

Donald R. Forsdyke

Evolutionary Bioinformatics

Third Edition

 Springer

Evolutionary Bioinformatics

Donald R. Forsdyke

Evolutionary Bioinformatics

Third Edition

 Springer

Donald R. Forsdyke
Department of Biomedical and Molecular Sciences
Queen's University
Kingston, ON, Canada

ISBN 978-3-319-28753-9 ISBN 978-3-319-28755-3 (eBook)
DOI 10.1007/978-3-319-28755-3

Library of Congress Control Number: 2016931596

Springer Cham Heidelberg New York Dordrecht London

© Springer International Publishing Switzerland 2006, 2011, 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

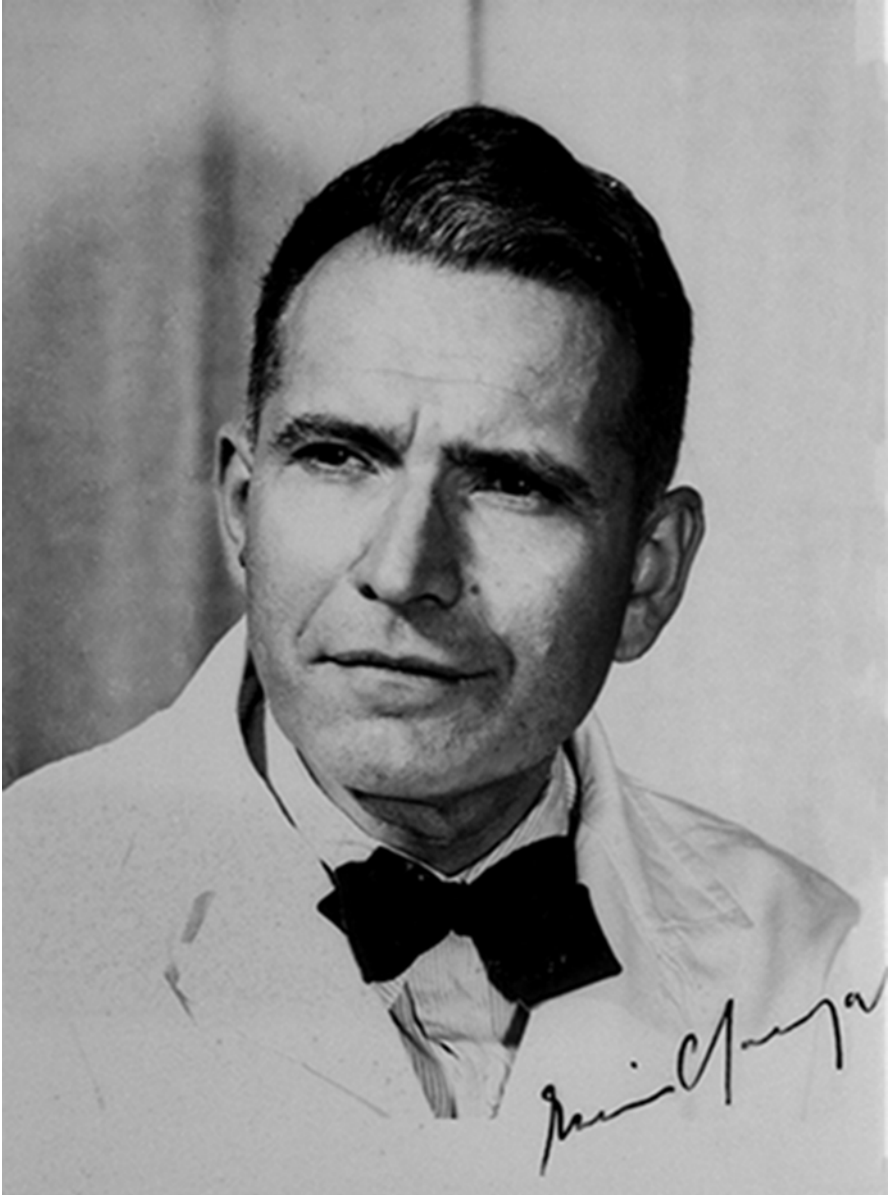
The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media
(www.springer.com)

*In memory of Erwin Chargaff whose studies
of the base composition 'accent'
of DNA underlie everything in this book,
and of Susumo Ohno who first understood
that DNA 'speaks' in palindromic verses;
with many thanks to Vinay Prabhu whose
neat little 1993 paper showed the way.*



Erwin Chargaff (1905-2002)

Contents

Note on the Third Edition	xvii
Note on the Second Edition	xix
Abbreviations	xxi
Prologue: To Select is Not To Preserve.....	xxiii

Part I Information and DNA

1 Memory: A Phenomenon of Arrangement.....	3
Priorities.....	3
Textual and Hereditary Information.....	5
Mental Information	6
Periodical Rhythms	9
Food	11
Informational Macromolecules.....	12
Translation	16
Variation.....	18
Latency.....	21
Summary	22
References.....	22
2 Chargaff’s First Parity Rule	25
Error-Detection	25
In-Parallel Redundancy.....	26
DNA Structure	26
Turnover and Channeling.....	32
Promiscuous DNA	33
Haploidy and Diploidy.....	34
In-Series Redundancy	35
Bits and Bats	36
Accidents and Non-Accidents.....	39
Summary	40
References.....	41

3 Information Levels and Barriers	43
Symbols and Code	43
Primary and Secondary Information	44
Primary Information for Protein	45
Information Barriers.....	46
Barriers Preserve Discontinuity	47
To Select is Not to Preserve	51
Another Order of Variation	52
Context as Barrier	54
Not a Palimpsest	57
Hybrid Disciplines and Hybrid Sterility	58
Summary	59
References.....	60

Part II Parity and Non-parity

4 Chargaff's Second Parity Rule	63
Information Conflict.....	63
Prose, Poetry and Palindromes	64
Intrastand Parity	65
Polarity and Complementarity	67
Duplets and Triplets	70
Species Specificity of K-mer Frequencies	74
Metagenomics	75
Introns Obey Better Than Exons.....	76
Level of Selection?	77
Inter-chromosomal Accounting.....	79
Base Frequency Interdependencies.....	80
Summary	81
References.....	81
5 Stems and Loops	83
Transfer RNA.....	83
The Bacteriophage Paradox	85
Genome-Level Selection.....	86
Calculation from Single Base Pairs	88
Role of Base Composition	90
Calculation from Dinucleotide Pairs.....	92
RNA Structure and Conflict	94
PR2 Revisited.....	96
Symmetrical Extrusions	97
DNA as Substrate	98
Summary	99
References.....	100

6	Chargaff's Cluster Rule	103
	Base Clusters.....	103
	Clusters of Clusters	104
	Polarity	106
	Origin of Replication	107
	Gene Movements	110
	Leading and Lagging Strands	111
	Purine loading	112
	Pyrimidine-Loading in Mitochondria	115
	Summary	117
	References.....	117

Part III Variation and Speciation

7	Mutation.....	121
	The Genetic Code	122
	One Base at a Time	124
	RNY Rule.....	125
	Negative and Positive Selection.....	127
	Neutral Mutations	129
	Genome Phenotype	130
	Compensatory Mutations.....	133
	Were First Mutations Synonymous?.....	134
	Phylogenetic Trees	136
	Limits of Natural Selection	137
	Summary	138
	References.....	139
8	Species Survival and Arrival.....	141
	Blending or Discrete Inheritance?	142
	Mendelism.....	143
	Multigenic Characters.....	145
	Reproductive Isolation	146
	Cycling Through the Generations	147
	Jurassic Park.....	149
	Branching Evolution	150
	Dobzhansky and Muller	152
	Pattern Change	153
	Summary	154
	References.....	155
9	The Weak Point	157
	Cryptic Variability of the Reproductive System	158
	Collective Variation of a Population Subset	159
	Sporadic and Regular Sterilities.....	160
	Chromosomes and the Residue or Irresoluble Base	163

Two Factors must Complement to Produce Sterility	164
Complementation like Sword and Scabbard.....	166
Modern Chromosomal Views	168
Darwin's Difficulties.....	169
Species Selection	170
Summary	172
References.....	173
10 Chargaff's GC rule	175
Uniformity of (G+C)%	176
The Holy Grail	176
Language.....	177
Crick's Unpairing Postulate	179
Structure-Mediated Homology Recognition.....	180
(G+C)% Controls Pairing	182
Mutational Meltdown.....	184
Coinfecting Viruses.....	185
Polyploidy Cures Hybrid Sterility	186
Protein Pressure Reinforces	187
Higher K-mer Differences as Ultimate Drivers	188
Saltum?	189
Summary	189
References.....	190
11 Homostability	193
Gene Duplication and Isochores	194
Isochores Early.....	198
The Gene as a Unit of Recombination.....	200
The Boundary Problem.....	202
Thermophiles	202
Summary	204
References.....	205
Part IV Conflict Within Genomes	
12 Conflict Resolution.....	209
Two Levels of Information	210
Codon Bias.....	210
GC-Pressure Versus Protein-Pressure	213
Species Barcode in Mitochondria	215
Species Versus Genes.....	217
Species Win at (G+C)% extremes.....	219
Intermediate (G+C)% Allows Compromise.....	222
Levels of Selection.....	223
Dog Wags Tail.....	224
AG-Pressure Versus Protein-Pressure.....	225

GC-Pressure Versus AG-Pressure	226
AG-Pressure and 'Placeholder' Amino Acids.....	228
Thermophiles	230
Summary	232
References.....	233
13 Exons and Introns	235
Introns Interrupt Information	237
Introns First.....	238
Protein Versus DNA.....	239
Achilles Heels	243
Mirror Repeats	244
RNA versus DNA.....	245
Overlapping Genes.....	245
Simple Sequences in Malaria.....	246
Multiple Pressures.....	248
Similarity Scores.....	249
Summary	250
References.....	250
14 Complexity	253
Scoring Information Potential.....	253
The New Bioinformatics.....	254
Protein or Nucleic Acid Level Function?.....	255
Base Pair Pressures	257
Epstein-Barr Virus.....	258
Malaria	263
Conflict with Fold Potential	266
Roles of Low Complexity Segments	268
Short Tandem Repeats	269
AC-Pressure and Asymmetric Fold Potential	271
Repetitive Sequences and Speciation.....	273
Meiotic Hotspots.....	274
Long-Range Periodicities.....	274
Summary	274
References.....	275
Part V Conflict Between Genomes	
15 Self/Not-Self?	279
Homology Search.....	280
Antibody Response	281
Antibody Variable Genes	283
Prototypic Immune Systems	284
Polymorphism Creates Host Unpredictability	287
Junk DNA.....	288

Repetitive Elements and “Selfish Genes”	289
All Hands on Deck!	291
The Double-Stranded RNA Alarm.....	292
Purine-Loading to Self-Discriminate	293
Entropy	295
The Hidden Transcriptome	297
CRISPR.....	299
Summary	301
References	301
16 The Crowded Cytosol	305
Piles of Coins	306
Homoaggregates	307
Collective Pressure.....	308
Concentration Fine-Tuning	309
Heteroaggregates.....	310
Protein ‘Immune Receptors’	311
Phenotypic Plasticity.....	312
Polymorphism Makes You Individual.....	313
Polymorphism Also Makes You Vulnerable	314
Death at Home or in Exile	315
Selfish Genes and the Menopause.....	316
Molecular Chaperones	317
Positive Repertoire Selection.....	320
Summary	321
References	321
Part VI Sex and Error-Correction	
17 Booting the Genome	327
Redundancy.....	328
Recombination Repair	329
Strand Guidance.....	331
DNA Damage.....	333
Piano Tuning and Calibrated Guessing.....	334
Spot Your Reprotype?	335
Haldane’s Rule	335
Sex Chromosomes	337
Sex and Speciation.....	339
Sex Chromosome Dosage Compensation.....	342
Gain in Aggregation Pressure	345
Thymus and Gonads as Auditioning Organs.....	345
Germ-Line Selection within Gonads	346
Summary	347
References.....	348

18 The Fifth Letter	351
Editing of DNA.....	351
Restriction Enzymes	353
CpG Islands.....	354
CpG Suppression	355
Methylation Differences in Twins.....	357
Imprinting	357
Transgenerational Inheritance.....	359
Editing of RNA.....	360
Summary	361
References.....	361

Part VII Information and Mind

19 Memory: What Is Arranged and Where?	367
The Organ <i>Par Excellence</i>	367
Not DNA	368
The Conventional Wisdom	369
Tomorrow.....	370
Savants and Information Measurement	371
Internal Signal Detection and Emission.....	372
Cloud Computing.....	373
The Physics of Space	374
Syntax	375
Argument from Incredulity	376
Summary	377
References.....	378
20 Certainty Now Uncertain	381
Allometry	381
Certainty.....	382
The Logic	383
Hydrocephalics	383
Lorber Confirmed.....	385
Savants and Microcephalics Don't Scale.....	385
Why Doesn't Size Matter?.....	386
The Search for the Soul.....	387
Theory of Mind.....	388
The Social Organism.....	389
The World as Eject.....	389
Consciousness	390
A Hilbertian List	391
Summary	392
References.....	392

Epilogue: To Preserve is Not To Select	395
Appendix A: What the Graph Says	409
Appendix B: Scoring Information Potential.....	415
Appendix C: Selection Level for PR2?	419
Appendix D: K-mer Probabilities and Genome Size	421
Appendix E: No Line?	425
Acknowledgements	435
Index.....	439

Note on the Third Edition

Since the second edition (2011), accesses of the text in ebook form have greatly increased. This parallels the growing trend towards online courses which supplement traditional educational approaches and facilitate self-directed learning. These courses can be much more than videos of formal lectures. The ‘Khan Academy’ is an interesting example. The pen-tablet of impresario Salman Khan captivates millions world-wide. Students see multicolored drawings, arrows, numbers and letters, moving across a black background to Khan’s melodious accompaniment. With pause and rewind options, they can proceed at their own pace.

While lacking Khan’s eloquence – but one has to start somewhere – I have experimented with his method, generating a series of 15 minute videos that explain evolutionary principles in everyday terms and, it is hoped, with a touch of humor. Rather than supplementing the present text, which has been extensively updated, the videos provide an elementary introduction – primarily directed at high school students and the general reader. For those who like this approach, I have also placed online a series of 1 hour talks on topics that link several chapters. The entire range may be accessed, either by going directly to YouTube, or through my webpages:

<https://www.youtube.com/user/forsdyke1/playlists>

<http://post.queensu.ca/~forsdyke/videolectures.htm>

Kingston, Canada
5th November 2015

Donald R. Forsdyke

Note on the Second Edition

Since the first edition in 2006 no text has risen to challenge its distinctive features: 1. Emphasis on non-genic bioinformatics ('the new bioinformatics'). 2. Belief that the key to a subject lies through its history. 3. Use of spoken and written language as a metaphor for DNA. 4. Intelligible to those not deeply versed in chemistry and mathematics. 5. 'Heads-on' rather than 'hands-on.' 6. 'Big picture' rather than 'small picture.' 7. Needs of both the 'bio' and 'informatic' communities met in one volume.

In the second edition, while some 'page creep' has been inevitable, I have tried to maintain the compact readability of the first. All chapters are revised and updated, and some lengthy chapters are split. To the six parts that composed the first edition a seventh is added. This reconnoiters that ultimate bioinformatic 'citadel' – the human brain. At present there is only one chapter in this section, but the rate of progress is such that more may be anticipated in future editions.

For too long there have been 'two cultures' – arts and sciences. There are numerous reasons that I will not dwell on here why this cannot be sustained. So, to the projected readership – primarily first year college science students – should be added the growing number of arts students for whom at least one science course is mandatory. Since the principles that govern language and our written texts also govern our genetic text, arts students may be surprised to find less difficulty than the title might suggest. Knowledge of the well known can enlighten the less well known. Naturally, there are sections, especially Parts II and IV, where those best versed in the humanities may find the going tough, but with plentiful java at hand, they can read selectively with some skimming and skipping, more for feel than for substance. Much should then percolate in, perhaps joyfully so!

Kingston, Canada
28th July 2010

Donald R. Forsdyke

Abbreviations

DNA	Deoxyribonucleic acid
dsRNA	Double-stranded RNA
EB	Evolutionary bioinformatics
EBV	Epstein-Barr virus
FONS	Folding of natural sequence
FORS-D	Folding of randomized sequence difference
FORS-M	Folding of randomized sequence mean
MHC	Major histocompatibility complex
ORF	Open reading frame
PR1	Chargaff's first parity rule
PR2	Chargaff's second parity rule
mRNA	Messenger RNA
R	Purine (A or G)
RNA	Ribonucleic acid
RNAi	RNA interference
rRNA	Ribosomal RNA
siRNA	Short interfering RNA
tRNA	Transfer RNA
Y	Pyrimidine (C or T)

Prologue: To Select is Not To Preserve

So long as uncertainty exists, there is nothing to be so much avoided as that sort of clearness which consists in concealing difficulties and overlooking ambiguities.

John Burdon Sanderson (1875) [1]

It is a truth universally acknowledged, that a single tract of text in possession of charms attractive to potential readers, must be in want of good binding and secure lodging. Universally acknowledged perhaps, but is charm alone sufficient to guarantee the latter? The loss of many works of antiquity with the burning of the great library of Alexandria suggests not. Texts that had been selected by generations of priests as most fitting to survive were not preserved. Selection did not guarantee preservation.

In the decade following James Watson and Francis Cricks' discovery of the structure of the genetic text in 1953, it was shown that certain DNA segments could be copied into RNA segments and that the information there encoded could then be translated into proteins. These were units of function that bestowed on an organism the discrete characters that Gregor Mendel had shown to be passed from generation to generation. Genes were defined as DNA segments encoding specific proteins, and were named accordingly. A protein in one organism might function better than the same protein in another organism, and on this basis the former organism might preferentially survive, as would the genes it contained. Thus, the natural selection of Charles Darwin would favor the survival of genetic texts because of the functions they encoded.

However, in 1966 George Williams proposed a more fundamental definition of the gene, not as a unit of function, but as a unit of preservation. A gene passed intact from generation to generation because it was preferentially preserved. A decade later Richard Dawkins popularized this idea as the "selfish gene." Yet, although there were already significant clues that preservation might require more than just successful function, neither Williams nor Dawkins would acknowledge agencies other than natural selection. Difficulties and ambiguities were shrugged off. Surely natural selection, by optimizing function, should keep organisms, and hence the

genes they contained, away from fires (enemies without). What threat could natural selection not evade? That to select *was* to preserve had been made abundantly clear in the full title of Darwin's great work – *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*? Williams thought that even the threat of dismemberment by the internal cutting-and-pasting of DNA segments, known as recombination (enemies within), would be evaded by natural selection.

The genic paradigm, with its faith in the supreme power of natural selection, reigned until the 1990s when, as part of various genome projects, computers began to spew forth DNA sequences and evolutionary bioinformatics (EB) emerged as a non-gene-orientated bioinformatics with roots dating back to the nineteenth century. Alas, when it came to evading recombination, natural selection did not guarantee preservation. The prevention of dismemberment was too important to be left to natural selection. Yet the paradigm shift implicit in this 'new bioinformatics' was not generally acknowledged. The genic juggernaut sped effortlessly from the twentieth century into the twenty-first.

That Sort of Clearness

Although now less costly, initially genome sequencing projects were very expensive. Several were completed with great expedition, not because of the scientific interest, not because a better understanding of genomes would make future gene manipulations safer, but because of the perception of those able to influence funding priorities that biology is about genes, and that the ability to identify and manipulate genes would be furthered by sequencing constellations of genes in the form of entire genomes. *Everyone* knew that genomes were about genes! So bioinformatics should be about genes and their products. Marketing the prospect of manipulating genes was easy when cures for cancer, schizophrenia and much more seemed within reach.

In marketing, simple messages work. "Concealing difficulties and overlooking ambiguities" becomes a way of life. The fanfare announcing that two research teams, one privately sponsored and one publicly sponsored, had sequenced the human genome led Sydney Brenner, one of the great pioneers in molecular biology, to declare in 2002 [2]:

Sequencing large genomes has nothing to do with intellectual endeavour. The creative work was done earlier by Fred Sanger and by others who improved the technology. The rest is about two things: money and management. As the various projects developed, their demand for money increased. The ... sequencing project quickly consumed most of the funds available for [other] genome work at a time when money was short.

On a book written by a leader of one of the sequencing teams, Brenner commented:

[He]...doesn't tell us anything about the genomes he sequenced. What did he find there that excited him? What did he learn about genes, about life, about evolution, about worlds to come? It is the play *Hamlet* without Hamlet.

While intending to disparage neither the great organizational and technical skills involved, nor the ultimate rewards that may accrue, my following comment in a medical journal reinforces Brenner's point [3]:

The overtaking of publicly funded research teams by Celera Genomics in its completion of the sequencing of the human genome smacks somewhat of the Sputnik episode. The USSR scaled up the payload of their rocket and put a man into space before the USA, but failed to make parallel progress in other technologies to capitalize on this advance. In a recent paper ... the CEO of Celera correctly states that “*understanding* the genome may help resolve previously unanswerable questions”. But Celera has merely made available the full text of the genome a little sooner than we might otherwise have had it. Those of us working to *understand* the genome no more need the full text of the genome than you would need the full text of a dictionary if you were trying to understand how a dictionary works.

A Buck or Two

So yes, let us applaud the progress in sequencing the human genome as a historic step. But let this not distract from the real issues. There is disquietude. At the beginning of the 21st century the AIDS pandemic continues its spread and funds continue to be directed/misdirected in search of the elusive vaccine. A strange element named “prion” lingers in latent form and slowly strangles its hosts, as it did the mad cows from which it seems to have arisen. Genes from genetically modified seed spread to normal crops. In Philadelphia, a genetically engineered virus (adenovirus) is used to treat a relatively minor genetic disease and the patient is dead three days later. In Paris, treatment of a life-threatening genetic disease with a genetically engineered virus (retrovirus) ends up with the ‘good’ gene (that was meant to replace a ‘bad’ gene) disrupting control of another gene that controls cell proliferation (insertional mutagenesis), so initiating cancer [4].

These are not acts of bioterrorism. They are acts of Nature and/or of well-intentioned people who think risk-benefit ratios are low enough to justify “concealing difficulties and overlooking ambiguities” – and perhaps making a buck or two as well.

The Dawn

Yet in 1991, at the dawn of the new era, Brenner set out the agenda quite clearly [5]:

Searching for an objective reconstruction of the vanished past must surely be the most challenging task in biology. I need to say this because, today, given the powerful tools of molecular biology, we can answer many questions simply by looking up the answer in Nature – and I do not mean the journal of the same name. ... In one sense, everything in biology has already been ‘published’ in the form of DNA sequences of genomes; but, of course, this is written in a language we do not yet understand. Indeed, I would assert that the prime task of biology is to learn and understand this language so that we could then compute organisms from their DNA sequences. ... We are at the dawn of proper theoretical biology.

In similar vein, biochemist Peter Little pointed out [6]: “It would be a great mistake to see the approach of enumeration of coding sequences as a substitute for the goal of the Human Genome Project. The goal is explicit ... to understand the information content of our DNA.” And, under the heading “Languages of the Chromosomes and Theoretical Biology”, molecular biologist Leroy Hood and his associates pointed to the numerous functions of DNA [7]:

Each of these functions places informational constraints upon our chromosomes. In a sense, these constraints represent distinct and sometimes overlapping languages. Deciphering some of these languages, and others that we do not have the imagination to envision, will require increasingly powerful computational and mathematical tools.

Homo bioinformaticus

Thus, from many quarters there were calls to cast aside genocentric blinkers and return to the fundamentals. This book explores the fundamental, sensed in England more than a century ago by George Romanes and William Bateson, that genomes are more than genes. Genomes are information channels, and genic information is but one of the forms of information which compete for genome space. Identifying these forms, revealing conflicts, and showing how the genomes of different species balance competing pressures on genome space, are the tasks of EB. These pages point the way and hint at rich rewards to follow.

You will find here that computation and mathematics are more than mere tools. In many respects, biological systems *are* computational systems [8], as indeed may be the universe itself [9]. Understanding computational systems (e.g. computers) may help us understand biological systems (e.g. you and me), and perhaps much more. In some respects, biology is more a branch – yes, a humble branch – of the discipline known as “information science,” than a component of a new hybrid discipline – bioinformatics.

Sadly, in the twentieth century many ‘wet’ laboratory-based biomedical scientists looked on ‘in silico’ laboratory-based computer scientists as capable only of tamely writing data-handling programs, rather than as capable of providing profound insights. Sometime in the twenty-first century a new breed of biomedical scientist, *Homo bioinformaticus*, will emerge. But current progress depends largely on partnership between the bio-people and the info-people. To facilitate this, biomedical scientists must eschew jargon and describe their perceived problems in simple, clearly defined, terms, and informatists must strive to make their art intelligible.

“Informatics” implies computing. To many, anything to do with computers sounds mathematical, and is therefore arcane and incomprehensible. Calls for “increasingly powerful computational and mathematical tools” tend to frighten people away. But there is very little actual mathematics in this book. Happily, I can retrieve sequences from on-line databases, and I can add, and subtract, and multiply, and divide. That is all *Homo bioinformaticus* will need to do to engage in the new

“informatics.” And a first year college biology background should suffice for the “bio.” This is not to say that the mathematically adept cannot take the subject to higher levels but, to paraphrase Wilhelm Johannsen, EB should be taken *with* mathematics, not *as* mathematics.

While mathematicians tend to seek broad generalizations by ‘mining’ large swaths of data that send statisticians purring with delight, biologists have found that conceptual clues often first arise from rare individual cases that do not conform. In the spirit of “treasure your exceptions” [10], we will often here be seeking out strange segments within a gene, strange genes within a biological species, and maverick species with extreme bioinformatic values.

Further Veils?

Will EB still be ‘new’ a decade (or even a century) hence? While so often one learns that to predict is to be wrong [11], there are substantial reasons for believing there will never be a *newer* bioinformatics with respect to the fundamentals. Unlike physics, where the conquest of each fundamental appears as the mere lifting of a veil to further fundamentals beneath, with the possible exception of the working of our own brains (see Part VII) it is likely there is nothing more fundamental in biology than the sequence of bases in a nucleic acid. To understand fully that sequence of four letters – **A**, **C**, **G** and **T** – is the goal of EB. In this context there can be nothing newer.

Of course, bioinformatics itself is bigger than this. It is literally the understanding of all biology in informational terms. When we focus on the bases of nucleic acids we call it genomics. When we focus on the proteins encoded by genes in genomes, we call it proteomics. Much of this book is about genomics, broadly interpreted to cover nucleic acids (DNA and RNA) as they operate both in cell nuclei (the usual site of DNA), and in the cytoplasm (the usual site of RNA).

Scope

“The historical mode of studying any question is the only one which will enable us to comprehend it effectually.” These words of the Victorian sheep-breeder and polymath, Samuel Butler, remind us that the ideas presented here once did not exist [12]. In the minds of humans seeking a path through a mist of incoherency, they emerged first, just as ideas. This was followed by an often much slower elaboration and definition of terms, which were eventually gathered into text. This was the creative process that happened in the past, and is ongoing. By understanding the past we prepare ourselves for participating in the ongoing present. To understand the past we look to biohistorians. However, for the all-too-human reason that workers at interfaces between disciplines need more than 24 hours in their days, they have not generally served us well. Palaeontologist Stephan Jay Gould pointed out [13]:

Many historians possess deeper knowledge and understanding of their immediate subject than I could ever hope to acquire, but none enjoy enough intimacy with the world of science ... to link this expertise to contemporary debates about the causes of evolution. Many more scientists hold superb credentials as participants in current debates, but do not know the historical background. ... A small and particular – but I think quite important – intellectual space exists, almost entirely unoccupied, for people who can use historical knowledge to enlighten ... current scientific debates, and who can then apply a professional's 'feel' for the doing of science to grasp the technical complexities of past debates in a useful manner inaccessible to historians (who have therefore misinterpreted, in significant ways, some important incidents and trends in their subject).

Could might have added that the biohistorian's dilemma is that many historical stones must be left unturned. Ideally, for complete objectivity, the turning should be random. This being difficult to achieve, then knowledge of contemporary debates can act as a powerful guide (heuristic). Of necessity, such debates contain speculative components and the lines between speculation and fact are never absolute. Cherished ideas that may have withstood centuries of debate are sometimes later seen to have been merely provisional.

Could might also have added that historical knowledge – at least that prior to the last century – is invariably of the 'big picture' variety, for the simple reason that no 'small picture' then existed. Our understanding of biological processes in biochemical terms is almost entirely a phenomenon of the twentieth century. We are very fortunate, but there is a flip side. Massive amounts of fine detail have drawn bioscientists into narrower and narrower fields of specialization and have bewildered biohistorians who increasingly tend to study sociological aspects rather than attempt to grapple with the science itself. Thus, bioscientists have lost touch with their Victorian antecedents some of whom, as will be shown here, had an uncanny 'feel' for the 'big picture' that is unequalled today.

Since its name was coined relatively recently, it might be thought that bioinformatics has a short history. It is shown here that the history of bioinformatics extends back to Darwin and beyond. This is made explicit in Chapter 1, *Memory: A Phenomenon of Arrangement*, which draws attention to the characteristics shared by different forms of stored information, be they mental, textual, or hereditary. Our understanding of forms we are familiar with (e.g. textual), may help our understanding of forms we are less familiar with (e.g. hereditary). At the outset the reader is introduced to Samuel Butler whose "Tottenham Court Road English" enlivens, and whose ideas enlighten, Darwin's ponderous prose.

The quantitative relationships between **A**, **C**, **G** and **T** were determined in the USA in the middle decades of the twentieth century by the biochemist Erwin Chargaff [14]. The regularities he discovered underlie much of the new bioinformatics. *Chargaff's First Parity Rule* (Chapter 2) considers features of the Watson-Crick structure of DNA that facilitate accurate transmission of hereditary information with the appropriate detection and correction of errors. This requires various levels of redundancy, the most obvious of which is that DNA molecules generally come in duplex form. At least two copies of the DNA message travel from one generation to the next. Yet, Crick questioned the potency of natural selection as an evolutionary force, and pointed to mechanisms involving recombination that might accelerate evolution.

Chapter 3, *Information Levels and Barriers*, focuses on informational aspect of heredity, noting that information often comes in quantal packets of restricted scope, like a book. The present book is about bioinformatics, not deep-sea diving, and it would be difficult to unite the two themes in a common text. Expanding the scope of information may require formation of new packets. This means that packets must be demarcated by appropriate informational barriers. Thus, the “primary” information a packet contains can be distinguished from the “secondary” information that demarcates and protects. There are different *levels* of information.

Chapter 4, *Chargaff's Second Parity Rule*, introduces conflict between different types and levels of information, as exemplified by prose, poetry and palindromes. Each of these imposes constraints on information transfer that enhance error-detecting power (e.g. if a poem does not rhyme you soon notice it). In Chapter 5, *Stems and Loops*, this theme is further explored in terms of the general ability of duplex DNA molecules, by virtue of the palindromes they contain, to extrude single-stranded segments as duplex-containing stem-loop structures with the potential to ‘kiss’ similar duplex structures. Paradoxically, the fittest organisms are those that can optimize the potential to form stem-loops, sometimes at the expense of other functions. The reader is left wondering why Nature has taken such pains to install stem-loops in all life forms.

Chapter 6, *Chargaff's Cluster Rule*, points to another informational constraint manifest as the preferential loading of loops in stem-loop structures with one type of base (‘purine-loading’). This should decrease the loop-loop ‘kissing’ interactions between single-stranded nucleic acids that may precede their pairing to form double-stranded nucleic acids (duplexes). Again, the reader is left wondering. Why should it be necessary to prevent the formation of such duplexes?

Chapter 7, *Mutation*, shows how each gene has its own rates of amino acid-changing and non-amino acid-changing mutations. These rates are correlated with each other, but not necessarily with those of neighboring genes. Thus each gene behaves as an individual mutational entity.

Chapter 8, *Species Survival and Arrival*, discusses the relative roles of natural selection and reproductive isolation in the generation of new species. Changes in DNA *primary* information, which affects an organism's structure or physiology (the *conventional phenotype*), provide genic targets for natural selection, so promoting linear evolution within a species. Changes in DNA *secondary* information, which affects the *genome phenotype*, have the potential to provide an initial demarcation (reproductive isolation), so sparking the origin of new species (branching evolution).

Surprising to many, the possibility that speciation might have a non-genic basis was worked out long ago by Darwin's research associate, George Romanes. In Chapter 9, what Thomas Huxley had referred to as *The Weak Point* in Darwin's theory – the need to explain hybrid sterility – is shown to have catalyzed Romanes' theoretical *tour de force* in the 1880s. These ideas were expanded by William Bateson in the years followed the discovery in 1900 of Mendel's now famous 1865 paper that launched the science of genetics.

The frequency of each base ‘letter’ in a sample of DNA can be counted. Chapter 10, *Chargaff's GC-Rule*, considers the frequencies of the bases **G** and **C** as the

‘accent’ of DNA that, like human accents, can serve to demarcate and even reproductively isolate. Differences in DNA accents work to impede the ‘kissing’ homology search between stem-loops extruded from potentially recombining DNA molecules during formation of gametes (e.g. spermatozoa and ova) in the gonads (e.g. testis and ovary). This can result in hybrid sterility, an interruption of the reproductive cycle within a species that permits certain species members to originate a new species. ‘Kissing’ turns out to be a powerful metaphor, since it implies an exploratory interaction that can have reproductive consequences.

Remarkably, the very same forces that serve to isolate genomes within a group of organisms (i.e. to originate a new species) also serve to isolate genes within a genome (i.e. to originate a new gene). Chapter 11, *Homostability*, describes Akiyoshi Wada’s observation that different genes have different ‘accents’ that preserve them as inherited units. While then not knowing that they could differ from each other in this way, the biologist George Williams defined genes as units that could resist dismemberment by recombination, rather than as units of function. This distinguished gene *preservation* from gene *function* and pointed to the involvement of an agency other than natural selection in the preservation (isolation) of genes and genomes.

The needs of ‘selfish genes’ and ‘selfish genomes’ can sometimes conflict. Chapter 12, *Conflict Resolution*, shows that the finite nature of genome space (‘channel capacity’) makes it likely that all demands on that space cannot be freely accommodated. It is unlikely that any enduring DNA can be dismissed as ‘junk’ or that any enduring mutation is truly ‘neutral.’ The multiple pressures acting at various levels on genomes require that there be trade-offs between different forms and levels of information. Chapter 13, *Exons and Introns*, shows how this can explain introns and low complexity elements. The latter are prominent in the genomes of herpesviruses and malaria parasites, which are considered in Chapter 14, *Complexity*. Here it is shown that some features of the amino acid sequence of proteins do not serve protein function; instead they reflect the needs of the encoding nucleic acid. The amino acids are mere ‘place-holders,’ slaves to the genomic pressures described in earlier chapters.

Organisms exist in an environment of predators, both external (enemies without, who remain without) and internal (enemies without, whose strategy is to get within). These coevolve with their prey or host, as they track them through evolutionary time. Thus, the full understanding of the genomes of any biological species requires an understanding of the genomes of the species with which it has coevolved. Chapter 15, *Self/Not-Self?*, turns from intracellular conflict *within* genomes to intracellular conflict *between* genomes. It is shown that, through purine-loading, ‘self’ RNA molecules could avoid ‘kissing’ and forming double-stranded RNA (dsRNA) with other ‘self’ RNAs. This would allow the RNA repertoire, derived from conventional and ‘hidden’ transcriptomes, to provide a selection of RNA ‘immunoreceptors’ that would ‘kiss’ viral (not-self) RNAs to generate dsRNAs, so activating intracellular alarms. The capacity of the genome information ‘channel’ being limited, it is advantageous for organisms selectively to display such anti-near-self immunoreceptors. Likewise, intracellular protein molecules are shown in Chapter 16, *The Crowded Cytosol*, to constitute a protein ‘immunoreceptor’ repertoire. Although reactions

between immunoreceptor proteins and viral proteins are likely to be weak, the crowded cytosol provides a conducive environment.

Chapter 17, *Rebooting the Genome*, reinforces the point that the need to detect and correct errors has been a major evolutionary force, driving DNA to ‘speak in palindromes’ and splitting members of species into two sexes. This, in turn, required sex chromosome differentiation, as a consequence of which the phenomena known as Haldane’s rule and dosage compensation emerged. The role of *The Fifth Letter* (Chapter 18) then assumed a new importance.

Returning to a theme of Chapter 1, in Chapter 19 – *Memory: What is Arranged and Where?* – there is exploration of aspects of brain function that make us uniquely human. Strangely, following a line of thought dating back at least to Avicenna in the 10th century, evidence on the location of long-term memory indicates that “the cupboard is bare”! Chapter 20 *On Certainty*, further examines this from the perspective of the philosophical quest for a frame-of-reference against which to assess phenomena we associate with the word ‘mind.’

Finally, the *Epilogue* turns to science politics and the opposition faced by those whose work is now our subject matter. There is all-too-often an all-too-human lack of objectivity that the reader should recognize. This has been with us in the past, is with us at present, and will doubtless persist. The flame of the torch of (what I believe is) enlightenment, has barely escaped extinguishment as it has passed from George Romanes to William Bateson, to the latter’s son Gregory, to Richard Goldschmidt, and to Stephen Jay Gould. The latter’s disagreements with Dawkins have received wide attention [15]. This book presents both theoretical and evidentiary bases for resolution of these disagreements. Thus, after a century of discontent, evolutionists may now hope to speak with one voice.

Goals

Who will read this book? Hopefully all who visit my bioinformatics web-pages daily from many parts of the world [16]. The site has been up-and-running since 1998 as a primary resource in its own right and to supplement my books. Links to my educational videos have recently been added. Those unfamiliar with the interpretation of graphical data and the laws of chance may find the *Appendices* of help. Each chapter has a summary of the main points so that readers who wish to skip should not miss much (e.g. Chapter 1, which outlines a history dealt with more fully elsewhere) [10, 17].

In my opinion much of the confusion in the evolution literature is semantic, so plain language is employed wherever possible, and usages are carefully defined (check the Index to locate definitions). Anticipating readers with a variety of backgrounds, for many of whom English may not be a first language, common usages (vernacular usages) have been preferred, while the corresponding more technical usages follow in round brackets (parentheses). There are abundant quotations (included in double quotation marks), which sometimes contain my explanatory

comments in square brackets. To imply a less literal meaning some words are in single quotation marks. Unless otherwise stated, all italicization is mine.

Literary allusions are scattered here and there. The foundations of EB were laid in Victorian London where literature and science intermingled and everyone either knew, or knew of, everyone else [1]. Our hero, Butler, contributed to, and battled with members of, both fields. Francis Galton, who coined the terms “nature and nurture” and “eugenics”, wrote a novel – *Kantsaywhere* – in the style of Butler’s most ‘successful’ work (*Erewhon*; the title is close to “nowhere” backwards). Authoress George Eliot contributed to the founding of the Physiological Society. A student of William Bateson met his future wife, the sister-in-law of authoress Virginia Wolfe, on the boat to Ceylon. Wolfe’s father advised Charles Darwin how to fend off Butler. Bateson’s sponsor, Christiana Herringham, inspired E. M. Forster’s *Passage to India* [10].

To make the text accessible to a wide audience I have adopted a ‘swimming pool’ approach so that moving between chapters should be somewhat like moving from the shallow to the deep end. More advanced readers who skip the early chapters should note that they are not shallow in substance. Rather they provide important images and metaphors that later chapters will flesh out in chemical terms. For advanced readers who happen to be still wedded to the classical Darwinian paradigm, the early chapters are mandatory! Less advanced readers eager to read the latest on junk DNA, introns, sex and speciation, are urged to be patient. In the style of a Norwegian folk tale, at each point I give the minimum necessary for comprehension, leaving the juicy details to be savored later [18].

Nevertheless, having splashed around in early chapters, there should be no hesitation in moving on. Though details may daunt, readers will not drown. Rather they will gain a general overview that I hope will satisfy. Those who cannot abide quantitation, however much simplified, should be able to skip Parts II and IV without losing the essence of the argument. Thus, there is a relatively easy route through the book – Parts I, III, V, VI and VII.

My ambitious hope is that, apart from being of interest to the general reader, the book will become a definitive resource on genome bioinformatics and evolution that will complement the numerous gene-orientated texts both currently available and to be written. As such it should serve the needs of all engaged in the biomedical sciences at college level and higher, from among whom *Homo bioinformaticus* will eventually emerge. It is not a ‘hands on’ text. The other texts do that very well. Rather, it is a ‘heads on’ text, designed to inspire ways of thinking that, for those so inclined, might make hands more productive. In short, readers are invited to escape the twentieth century box of gene reductionism and become part of the revolution that we may, for the time-being, refer to as “the new bioinformatics.”

So – I am at the end of my Prologue. Hopefully, you now have some feeling for the subject of the book, and are tempted to read on. But, is that all? What about ‘being,’ ‘soul,’ ‘ultimate meaning,’ and – a major concern of Francis Crick – ‘consciousness’? Are such matters, which touch the essence of our existence, beyond our present scope? It is difficult to improve on the words of Leslie Dunn when

anticipating that some might consider his text *A Short History of Genetics*, to be lacking in philosophical penetration [19]:

It is not that I, or another who may review the development of the science of heredity, may not be interested in what it [all] means, but rather that before we ask the ultimate question we should know what has caused it to be asked.

This book continues the quest to “know what has caused it to be asked.” But, as the reader will find in Part VII and Appendix E, the temptation to guess a little could not be resisted entirely.

References and Notes¹

1. Romano T (2002) *Making Medicine Scientific. John Burdon Sanderson and the Culture of Victorian Science*. John Hopkins University Press, Baltimore, p 105 [Sanderson was the great uncle of the twentieth century’s “JBS”, John Burdon Sanderson Haldane, and may have partly inspired Lydgate in George Eliot’s *Middlemarch*.]
2. Brenner S (2002) The tale of the human genome. *Nature* 416:793–794
3. Forsdyke DR (2001) Did Celera invent the Internet? *The Lancet* 357:1204
4. Wu X, Li Y, Crise B, Burgess SM (2003) Transcription start regions in the human genome are favored targets of MLV integration. *Science* 300:1749–1751
5. Brenner S (1991) Summary and concluding remarks. In: Osawa S, Honjo T (eds) *Evolution of Life: Fossils, Molecules and Culture*. Springer-Verlag, Berlin, pp 453–456
6. Little P (1995) Navigational progress. *Nature* 277:286–287
7. Hood L, Rowen L, Koop BF (1995) Human and mouse T-cell receptor loci. Genomics, evolution, diversity and serendipity. *Annals of the New York Academy of Sciences* 758: 390–412
8. Battail G (2014) *Information and Life*. Springer, Dordrecht
9. Wolfram S (2002) *A New Kind of Science*. Wolfram Media, Champaign, IL
10. Cock AG, Forsdyke DR (2008) “Treasure Your Exceptions.” *The Science and Life of William Bateson*. Springer, New York
11. Badash L (1972) The complacency of nineteenth century science. *Isis* 63:48–58
12. Butler S (1914) *The Humour of Homer and Other Essays*. Kennerley, New York, p 255
13. Gould SJ (2002) *The Structure of Evolutionary Theory*. Harvard University Press, Cambridge, MA, pp 36–37
14. Forsdyke DR, Mortimer JR (2000) Chargaff’s legacy. *Gene* 261:127–137
15. Morris R (2001) *The Evolutionists*. Norton, New York
16. Forsdyke DR (2015) Homepage of Donald Forsdyke (<http://post.queensu.ca/~forsdyke/homepage.htm>); Forsdyke Evolution Academy (<http://post.queensu.ca/~forsdyke/videolectures.htm>)

¹ Full text versions of some of the references in this book, together with much supplementary information, may be found in my web-pages, which are regularly updated. The pages may be accessed at: <http://post.queensu.ca/~forsdyke/homepage.htm>

Early versions of the latter web-pages are archived at:

<http://www.archive.org>

<https://qspace.library.queensu.ca/html/1974/421/homepage.htm>

The date of a written work sometimes does not coincide with that of its publication. When this arises I give the date of the written work after the author’s name, and the date of publication with the publisher’s name.

17. Forsdyke DR (2001) *The Origin of Species, Revisited. A Victorian who Anticipated Modern Developments in Darwin's Theory*. McGill-Queen's University Press, Montreal
18. Asbjørnsen PC (1963) *The Three Billy-Goats Gruff. A Norwegian Folk Tale*. Holt, Rinehart & Winston, New York [Early chapters serve to raise questions. Later chapters provide possible answers.]
19. Dunn LC (1965) *A Short History of Genetics*. McGraw-Hill, New York, p xxii

Part I

Information and DNA

Chapter 1

Memory: A Phenomenon of Arrangement

*A little learning is a dang'rous thing; drink deep, or taste not
the Pierian spring.*

Alexander Pope [1]

Because genes could explain so much in the biomedical sciences, in the twentieth century it was tempting to assume they could explain everything. So, for many, the bioinformatic analysis of genomes came to mean the bioinformatic analysis of genes and their products. However, bioinformatics has become less gene-centered. Evolutionary bioinformatics – the ‘new bioinformatics’ – views genomes as channels conveying multiple forms of information through the generations from the distant past to the present. As with information channels in general, genomes can become overloaded so that conflicts between different forms of information can arise, and ‘noise’ can delay information transmission and imperil its accuracy. The tasks of the new bioinformatics are to identify the forms of information that genomes transmit, and to show how conflict and noise are dealt with so that reliable transmission is sustained. The results should complete Charles Darwin’s work of understanding how, from simple organic molecules, “endless forms most beautiful and most wonderful have been, and are being, evolved” [2].

Apart from the satisfaction of this, there is an urgent practical purpose. Heeding not Pope’s admonition to “drink deep”, we have forged from fragile foundations new gene-based technologies that, while of immense potential benefit, are being applied to agricultural and medical problems with but a cosmetic nod towards possible hazards. We can only hope that the goals of the new bioinformatics will be attained before biotechnologists, with massive industrial support, open Pandora’s box.

Priorities

Evolutionary bioinformatics (EB) is only now new because the seductive simplicity of the genic paradigm (Fig. 1.1) and all that it promised, distorted perceptions, and hence, priorities. Thus a *gene-centered* bioinformatics was the first to emerge. The new, *non-gene-centered* bioinformatics emerged later. It is to the twenty-first

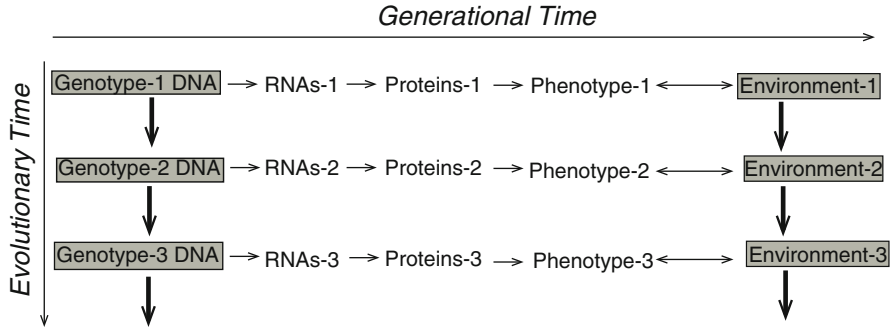


Fig. 1.1 The genic paradigm. Genomes and environments change over evolutionary time (heavy vertical arrows). Within a given generation the genetic program (contained in the DNA ‘archive’) is *transcribed* to provide instructions (messenger RNAs) that are *translated* into proteins. These are largely responsible for the anatomical and physiological characters (phenotype) with which organisms interact with the prevailing environment. Genotypes (i.e. constellations of genes) that most successfully track environmental changes (i.e. produce successful phenotypes) tend to persist

century, somewhat what genetics was to the twentieth. The words of William Bateson – the biologist who, struggling for recognition of genetics, brought Gregor Mendel to the attention of the English-speaking world – still seem apt [3]:

If the work which is now being put into these occupations were devoted to the careful carrying out and recording of experiments of the kind we are contemplating, the result ... would in some five-and-twenty years make a revolution in our ideas of species, inheritance, variation, and other phenomena.

Likewise Charles Darwin had declared in 1872 when praising a book on natural history [4]: “How incomparably more valuable are such researches than the mere description of a thousand species.” This might well be paraphrased today with respect to “a thousand genes” and their products.

Of course, descriptions of species and genes are important, and should be well supported. We have many complete genomic sequences and we need many more. The point is that humankind is likely to progress optimally when *all* aspects of science are supported – not just those that, being perceived as important by vested interests, provide a safe academic haven not only for serious researchers, but also for the intellectually non-curious and the politically ambitious. While some were ecstatically comparing the sequencing of the human genome to the discovery of the Rosetta Stone, molecular geneticist William Gelbart soberly suggested [5]:

It might be more appropriate to liken the human genome to the Phaistos Disk: an as yet undeciphered set of glyphs from a Minoan palace on the island of Crete. With regard to understanding the A’s, T’s, G’s and C’s of genomic sequences, by and large, we are functional illiterates.

Without apology we will begin our tasting of the “Pierian spring” by examining the historical roots of our subject. So often the understanding of a subject requires the understanding of its history. Yet, to really understand its history, one must understand

the subject. This paradox implies that the study of a subject and of its history must go hand in hand. Furthermore, by studying history we learn about *process*. By understanding the process of past discovery we prepare ourselves for understanding, and perhaps participating in, future discoveries.

History is about people, ideas and technologies. Our sources on these are usually written sources. However, often ideas arrive before there are suitable words to convey them. So the examination of the history of ideas can become an examination of the words people coined and used, and of what they meant by those words in particular circumstances. This is no light task but, as I hope to show, is well worth the effort.

Textual and Hereditary Information

At the heart of the new bioinformatics are the ideas of ‘information’ and ‘memory’ (i.e. stored information). To build a house you need bricks and mortar and something else – the ‘know-how,’ or ‘information,’ as to how to go about your business. You need a plan (i.e. again, stored information). Victorians like Charles Darwin understood this. In the context of the ‘building’ of animals and plants, the *idea* of ‘information,’ but seldom the word, recurs in their writings.

Classical scholars tell us Aristotle wrote that the "*eidos*," the form-giving essence that shapes the embryo, "contributes nothing to the material body of the embryo but only communicates its program of development" [6]. Darwin's mentor, the geologist Charles Lyell, may have had some intuition of this when in 1863, in his *Geological Evidences on the Antiquity of Man* [7], he wrote in a chapter entitled "Origin and Development of Languages and Species Compared:"

We may compare the persistence of languages ... to the force of inheritance in the organic world The inventive power which coins new words or modifies old ones ... answers to the variety making power in the animate creation.

Gregor Mendel may have thought similarly when wondering in 1865 why one pea plant looked so much like its parents, but another did not [8]. He considered that "development must necessarily proceed in accord with a law," and that "distinguishing traits ... can only be caused by differences in the *composition* and *grouping* of ... elements." One of the few familiar with Mendel's work, the botanist Carl von Nägeli in Munich [9], agreed:

To understand heredity, we do not need a special independent symbol for every difference conditioned by space, time or quality, but a substance which can represent every possible combination of differences by the fitting together of a limited number of elements, and which can be transformed by permutations into other combinations.

Darwin in 1868 considered the relative contributions of information in the male gamete (pollen), and in the female gamete (ovule), to the seed from which a new plant would emerge [10]. Thus, he wrote of "the peculiar *formative* matter" that was contained in male gametes and was needed for "the full development of the seed"

and for “the vigour of the plant produced from such seed.” He noted that “the ovules and the male element have *equal* power of transmitting every single character possessed by either parent to their offspring.” Furthermore, phenomena such as the regeneration of an entire plant from leaf cuttings (vegetative propagation) suggested that the “formative elements” were “not confined to the reproductive organs, but are present in the buds and cellular tissue of plants.” So “the child, strictly speaking, does not grow into the man, but includes *germs* which slowly and successively become developed and *form* the man.”

He equated “germs” with the “formative matter,” which “consists of minute particles or gemmules.” These were smaller than cells, were capable of independent multiplication by “self-division,” and were responsible for the generation of new organisms. Thus: “An organism is a microcosm – a little universe, formed of a host of self-propagating organisms, inconceivably minute and numerous as the stars in heaven.”

The major Victorian advocate of Darwin’s views, biologist Thomas Huxley, anticipated modern ideas on ‘selfish genes’ [11] and ‘molecular ecology’ [12] when writing in similar vein in 1869 [13, 14]:

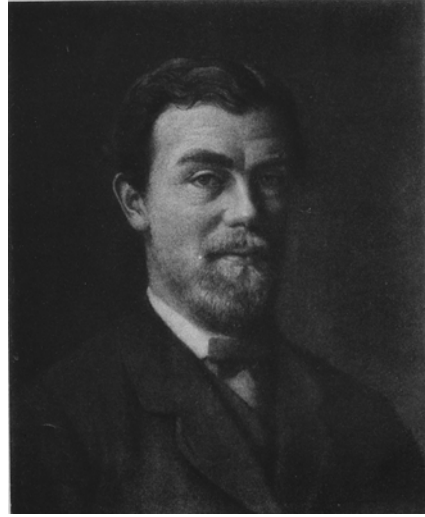
It is a probable hypothesis that, what the world is to organisms in general, each organism is to the molecules of which it is composed. Multitudes of these, having diverse tendencies, are competing with each other for opportunities to exist and multiply; and the organism as a whole, is as much a product of the molecules which are victorious as the Fauna or Flora of a country is the product of the victorious beings in it.

Mental Information

Darwin’s ideas were based on three fundamental principles – variation, heredity and natural selection. *Variation* caused an individual to differ in certain characters from others. *Heredity* caused characters to be transmitted from parents to children. *Natural selection* caused some individuals to survive and reproduce better than others. Of these three – variation, heredity and natural selection – only the basis of the latter was understood. For the Victorians, heredity and variation were unknowns to be entrusted to researchers of future generations. The words “heredity” and “variation” were handles with which to manipulate these principles, irrespective of their underlying mechanisms. One Victorian, however, thought otherwise (Fig. 1.2).

Initially heredity was seen as the process by which characters (which we now call the phenotype) were passed *forward* from parent to child. However, around 1870, this perspective was shifted by the physiologist Ewald Hering in Prague, and by Samuel Butler in London. They put the onus on the embryo, which *remembers* its parents. The newly formed embryo is a passive recipient of parental information (which we now call the genotype), and this information is used (recalled) by the embryo to construct itself (i.e. to regenerate phenotype from genotype). Heredity and memory are, in principle, the same. *Heredity is the transfer of stored information*. This powerful conceptual leap led to new territory. Evolutionary processes could thenceforth be thought of in the same way as mental processes. Hering [15] considered that:

Fig. 1.2 Self-portrait
(1878) of Samuel Butler
(1835-1902)



We must ascribe both to the brain and body of the new-born infant a far-reaching power of remembering or reproducing things which have already come to their development thousands of times in the persons of their ancestors.

More simply, Butler wrote [16]:

There is the reproduction of an idea which has been produced once already, and there is the reproduction of a living form which has been produced once already. The first reproduction is certainly an effort of memory. It should not therefore surprise us if the second reproduction should turn out to be an effort of memory also.

Perhaps sitting across from Karl Marx in the Reading Room of the British Museum in London, Butler devoured the evolution literature of his time and challenged the men of science. Bateson, who himself opposed a too simplistic interpretation of Darwin, came to regard Butler as "the most brilliant, and by far the most interesting of Darwin's opponents" [17]. Butler thought "a hen is only an egg's way of making another egg" [18]. In the twentieth century, zoologist Richard Dawkins chimed similarly that a body is only a gene's way of making another gene [11]. An effort at reading Butler's seemingly convoluted writings about plant seeds in *Erewhon* is well rewarded [19]:

The rose-seed did what it now does in the persons of its ancestors – to whom it has been so linked as to be able to remember what those ancestors did when they were placed as the rose-seed now is. Each stage of development brings back the recollection of the course taken in the preceding stage, and the development has been so often repeated, that all doubt – and with all doubt, all consciousness of action – is suspended. ... The action which each generation takes – [is] an action which repeats all the phenomena that we commonly associate with memory – which is explicable on the supposition that it has been guided by memory – and which has neither been explained, nor seems ever likely to be explained on any other theory than the supposition that there is an abiding memory between successive generations.

Hering wanted his strictly materialist position be understood [15]: “Both man and the lower animals are to the physiologist neither more nor less than the matter of which they consist.” He held that ideas that disappeared from consciousness would be stored in some material form in the unconscious mind, from which they could be recalled to consciousness. Thus, “the physiology of the unconscious is no ‘philosophy of the unconscious’.”

Actions that were repeated were more likely to be stored in memory, and hence more accurately recalled, than actions that were performed only once or twice. So practice-makes-perfect with respect both to psychomotor functions, such as reproducing music on the piano, and to hereditary functions, such as reproducing a cell from a parent cell, and reproducing an organism from a parent organism (asexual reproduction) or parent organisms (sexual reproduction) [15]:

But if the substance of the germ can reproduce characteristics acquired by the parent during its single life, how much more will it not be able to reproduce those that were congenital to the parent, and which have happened through countless generations to the organized matter of which the germ of today is a fragment? We cannot wonder that action already taken on innumerable past occasions by organized matter is more deeply impressed upon the recollection of the germ to which it gives rise than action taken once only during a single lifetime.

Among the recollections of the germ would be instinctual (innate) actions. A pianist, having practiced a piece so well that it can be played virtually unconsciously, does not, as far as we know, pass knowledge of that piece through the germ line to children. Yet, some psychomotor activities of comparable complexity are passed on [15]:

Not only is there reproduction of form, outward and inner conformation of body, organs, and cells, but the habitual actions of the parent are also reproduced. The chicken on emerging from the eggshell runs off as its mother ran off before it; yet what an extraordinary complication of emotions and sensations is necessary in order to preserve equilibrium in running. Surely the supposition of an inborn capacity for the reproduction of these intricate actions can alone explain the facts. As habitual practice becomes a second nature to the individual during his single lifetime, so the often-repeated action of each generation becomes a second nature to the race.

In 1884 in his *Mental Evolution in Animals*, Darwin’s ex-research associate, the physiologist George Romanes, was happy to consider instinctual activities, such as bird migration and nest-making, as requiring inherited (“ready-formed”) information [20]:

Many animals come into the world with their powers of perception already largely developed. This is shown ... by all the host of instincts displayed by newly-born or newly-hatched animals. ... The wealth of ready-formed *information*, and therefore of ready-made powers of perception, with which many newly-born or newly-hatched animals are provided, is so great and so precise, that it scarcely requires to be supplemented by the subsequent experience of the individual.

But, much to Butler’s frustration, he would not move beyond this [21]:

Mr. Romanes ... speaks of ‘heredity as playing an important part in *forming memory* of ancestral experiences;’ so that whereas I want him to say that the phenomena of heredity are due to memory, he will have it that the memory is due to heredity Over and over again Mr. Romanes insists that it is heredity which does this or that. Thus, it is ‘*heredity with natural selection which adapt* the anatomical plan of the ganglia;’ but he nowhere tells us what heredity is any more than Messrs. Herbert Spencer, Darwin and Lewes have done. This, however,

is exactly what Professor Hering, whom I have unwittingly followed, does. He resolves all phenomena of heredity, whether in respect of body or mind, into phenomena of memory. He says in effect, 'A man grows his body as he does, and a bird makes her nest as she does, because both man and bird remember having grown body and made nest as they now do ... on innumerable past occasions. He thus reduces life from an equation of say 100 unknown quantities to one of 99 only, by showing that heredity and memory, two of the original 100 unknown quantities, are in reality part of one and the same thing.' [Butler's italics]

Periodical Rhythms

In 1880 in his book *Unconscious Memory* Butler employed the word "information" in the context of memory [22]:

Does the offspring act as if it remembered? The answer to this question is not only that it does so act, but that it is not possible to account for either its development or its early instinctive actions upon any other hypothesis than that of its remembering, and remembering exceedingly well. The only alternative is to declare ... that a living being may display a vast and varied *information* concerning all manner of details, and be able to perform most intricate operations, independently of experience and practice."

In 1887 in *Luck or Cunning* [23], Butler showed a good understanding of the concept of information, and of its symbolic representation in a form conditioned by factors both internal and external:

As this idea [of a thing] is not like the thing itself, so neither is it like the motions in our brain on which it is attendant. It is no more like these than, say, a stone is like the individual characters, written or spoken, that form the word 'stone.' ... The shifting nature ... of our ideas and conceptions is enough to show that they must be symbolical, and conditioned by changes going on within ourselves as much as by those outside us.

Later, he was more poetic [24]:

Some ideas crawl, some run, some fly; and in this case words are the wings they fly with, but they are only wings of thought or of ideas, they are not the thought or ideas themselves.

Again emphasizing "the substantial identity between heredity and memory," Butler noted that "Variations [mutations to the modern reader] ... can only be perpetuated and accumulated because they can be inherited." Echoing Huxley, in 1880 he considered the unicellular amoeba [22]:

Let us suppose that this structureless morsel of protoplasm is, for all its structurelessness, composed of an infinite number of living molecules, each one of them with hopes and fears of its own, and all dwelling together like Tekke Turcomans, of whom we read that they live for plunder only, and that each man of them is entirely independent, acknowledging no constituted authority, but that some among them [DNA to the modern reader] exercise a tacit and undefined influence over the others.

Let us suppose these molecules capable of memory [of storing information], both in their capacity as individuals [individual molecules], and as societies [groups of molecules], and able to transmit their memories to their descendants, from the traditions of the dimmest past to the

experiences of their own lifetime. Some of these societies will remain simple, as having no history, but to the greater number unfamiliar, and therefore striking, incidents [mutations to the modern reader] will from time to time occur, which, when they do not disturb memory so greatly as to kill [lethal mutations], will leave their impression upon it [accepted mutations].

The body or society will remember these incidents, and be modified by them in its conduct, and therefore ... in its internal arrangements. ... This memory of the most striking events of varied lifetimes [inherited information] I maintain ... to be the differentiating cause, which, accumulated in countless generations, has led up from the amoeba to man ...

Thus, Butler identified a class of informational molecule that paradoxically would display *both* stability and variability:

We cannot believe that a system of self-reproducing associations should develop from the simplicity of the amoeba to the complexity of the human body without the presence of that memory which can alone account at once for the resemblances and differences between successive generations, for the arising and the accumulation of divergences – for the tendency to differ and the tendency not to differ.

Like most others, Butler discarded the “Russian doll” (“germs within germs” or “preformation”) model of inheritance [16]:

When we say that the germ within the hen’s egg remembers having made itself into a chicken on past occasions ... do we intend that each single one of these germs was a witness of, and a concurring agent in, the development of the parent forms from their respective germs, and that each of them therefore, was shut up with the parent germ, like a small box inside a big one? If so, then the parent germ with its millions of brothers and sisters was in like manner enclosed within a grandparental germ, and so on till we are driven to admit, after even a very few generations, that each ancestor has contained more germs than could be expressed by a number written in small numerals, beginning at St. Paul’s and ending at Charing Cross. ... Therefore it will save trouble ... to say that the germs that unite to form any given sexually produced individual were not present in the germs, or with the germs, from which the parents sprang, but that they *came into* the parents’ bodies at some later period.

So the germs that originally united to form each parent were no longer present in those parents. This raises the question as to how *new* germs “came into” the parents [16]:

We may perhaps find it convenient to account for their intimate acquaintance with the past history of the body into which they have been introduced by supposing that by virtue of assimilation [metabolism] they have acquired certain *periodical rhythms* already pre-existing in the parental bodies, and that the communication of the characteristics of these rhythms determines at once the physical and psychical development of the individual in a course as nearly like that of the parents [inheritance] as changed surroundings [environment] will allow.

So the material forms of the parental information are no longer present in their child, but during its life “periodical rhythms” are communicated to new material forms that will allow the generation of future children (the grandchildren of the original parents). Were the original parental material forms – the ‘papers’ upon which the ‘plan’ for their child was written – completely destroyed in the child so that fresh germs had to be reconstructed from scratch, like an architect taking measurements from an existing building in order to prepare a plan for a new, identical, building? Probably not. The term “periodical rhythms” suggests that Butler had some sort of copying in mind [16]:

For body is such as it is by reason of the characteristics of the vibrations that are going on in it, and memory is only due to the fact that the vibrations are of such characteristics as to catch on to, and be caught onto by, other vibrations that flow into them from without – no catch, no memory.

So, although the original parental material form of the plan (paper) would no longer be present, its information would have been copied to create identical plans on new papers. Thus, there would be a “recurrence” of the “rhythm” [16]:

I see no way of getting out of this difficulty so convenient as to say that a memory is the reproduction and *recurrence* of a rhythm communicated directly or indirectly from one substance to another, and that where a certain rhythm exists there is a certain stock of memories [stored information], whether the actual matter in which the rhythm now subsists was present with the matter in which it arose, or not.

In a later edition of *Unconscious Memory*, rhythm or vibration was deemed equivalent to “the molecular change in the particular nerves affected – for molecular change is only a change in the character of the vibrations going on within the molecules – it is nothing else than this.”

Food

In 1859 Butler had sailed for New Zealand with Justus von Liebig’s *Agricultural Chemistry* determined to make a fortune sheep ranching. This had set him, perhaps for the first time in his life, thinking deeply about food. He concluded that “matter ... introduced into the parents’ bodies during their life histories [i.e. food] ... goes to form the germs that afterwards become their offspring,” but he was troubled as to whether food should be considered as living or non-living, for “if living, it has its own *memories and life-histories* which must be *cancelled and undone* before the assimilation and the becoming *imbued with new rhythms* can be complete. That is to say it must become as near to non-living as anything can become” [16]. He toyed with the case of a grain of corn that if sown would make a corn plant. Yet, having eaten that grain, a bird remained a bird and produced bird offspring not corn offspring. What happened to the corn’s information?

Recall that he considered cells to consist of “living molecules, each one of them with hopes and fears of its own, and all dwelling together like Tekke Turcomans” [22]. If we think of polymeric DNA as “living,” and the subunits (letters) from which DNA is constructed (assimilated) as “non-living” (see Chapter 2), then Butler’s nineteenth century abstraction is far from wild. Food contains specific DNAs of other organisms, which encode their specific characteristics (“memories and life histories”). These polymeric DNAs are degraded (“cancelled and undone”) in digestive systems to subunits (“non-living”) that are absorbed into bodies where they are *repolymerized* in a new order (“imbued with new rhythms”) to produce copies of the bodies’ own DNAs (“living”).

He may have been thinking of the set of letters that a nineteenth century printer would assemble on a printer’s block. After a print run the letters would be removed from the block and then reutilized (i.e. assembled together in a different order) for the next day’s print run. Thus, when considering verbal communication Butler wrote [24]:

The spoken symbol ... perishes instantly without material trace, and if it lives at all does so only in the minds of those who heard it. The range of its action is no wider than that within which a voice can be heard; and every time a fresh impression is wanted the type must be set up anew.

Informational Macromolecules

In 1764 Ko  reuter in Germany deduced from plant breeding studies that maternal and paternal contributions to offspring were equal. That the “composition and grouping of ... elements” might require some material basis was recognized in the nineteenth century by the English philosopher Herbert Spencer, by N  geli, and by Darwin’s cousin, Francis Galton [25, 26]. It followed that, since “the ovules and male element have equal power of transmitting every single character,” then the quantity of that material could be no more than could be contained in the smallest gamete, which was usually the male gamete (pollen or spermatozoon; Fig. 1.3).

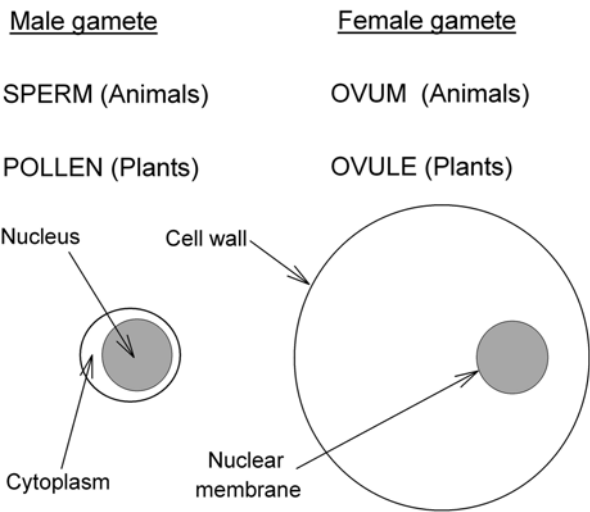


Fig. 1.3 If heredity has a non-compressible, non-stretchable, material basis, then, since male and female contributions to our genetic heritage are equal, the quantity of hereditary material cannot be more than is contained in the smallest of the two gametes, which is usually the male gamete (spermatozoon or pollen grain). Thus much of the material in the female gamete (ovum or ovule) cannot be directly concerned with the transmission of hereditary characters

In 1866 the English chemist William Crookes commented on the probable nature of the poisonous “particles” of the cattle plague that was ravaging Europe. The particles could not be seen with the microscopes then available. He deduced that they were not bland, but “organized” to a degree much greater than that of the chemical poisons with which people were then familiar [27]:

The specific disease-producing particles must moreover be organized, and possess vitality; they must partake of the nature of *virus* rather than of *poison*. No poison yet known to chemists can approach, even in a faint degree, the tremendous energy of the active agent of infectious diseases. A poison may be *organic*, but it is not *organized*. It may kill with far greater rapidity than the virus of infection, but, unlike this virus, it cannot multiply itself in the animal economy to such an extent as to endow within a few hours every portion of its juices with the power of producing similar results. A virus, on the contrary, renders the liquids of an infected animal as virulent as the original germ.

Viruses, as we now know them, were first identified and named decades later. Crookes used the word in the Latin sense understood by his contemporaries, meaning, “poison” or “toxin.” But his personal understanding appears modern in that he implied that a virus would use the resources available in “the animal economy” to multiply itself. The word “germ” meant a biological entity with the potential to sprout or “germinate” (like a plant seedling), and as such the word was used in the context of organisms we now refer to as viruses and bacteria. Darwin’s notion of the organism as “a little universe, formed of a host of self-propagating organisms” seems to have derived from this emerging understanding of the minuteness and great proliferative powers of the microorganisms responsible for the cattle plague. In his 1866 testimony to a government enquiry, John Simon, the Medical Officer of the Privy Council suggested [28]:

The several zymotic diseases are aetiologically quite distinct from one another. ... How their respective first contagia arose is ... quite unknown. This, in pathology, is just such a question as in physiology is ‘the origin of species.’ Indeed ... it is hardly to be assumed as certain that these apparently two questions may not be only two phases of one. Hourly observation tells us that the contagium of smallpox will breed smallpox, that the contagium of typhus will breed typhus, that the contagium of syphilis will breed syphilis, and so forth, – that the process is as regular as that by which dog breeds dog, and cat cat, as exclusive as that by which dog never breeds cat, nor cat dog.

In 1868 Darwin advanced his “provisional theory of pangenesis,” and named a minute fundamental unit of heredity a “gemmule” [10]. By this time, on the continent of Europe microscopic investigations had distinguished cell nuclei and the phenomena of cell division (29). Galton in 1872 [30] and August Weismann in 1885 noted that for many organisms there was a potential separation of the hereditary material in germ-line cells (that produced gametes) from that contained in other body cells (the cells of the mortal “soma”) [31]:

Splitting up of the substance of the ovum into a somatic half, which directs the development of the individual, and a propagative half, which reaches the germ-cells and there remains inactive, and later gives rise to the succeeding generation, constitutes *the theory of the continuity of the germ plasm*.

Bateson considered that Butler had arrived at “this underlying truth” in 1878 “by natural penetration” [17, 18]. The Dutch botanist Hugo de Vries in his *Intracellular Pangenesis* [32] boldly took the issue to the molecular level in 1889, noting that an observed character of an organism could be “*determined by a single hereditary character or a small group of them.*” These determining “hereditary characters” are “units, each of which is built up of numerous chemical molecules,” and are “arranged in rows on the chromatin threads of the nucleus.” The units which, rather than “gemmules,” he named “pangens,” could, for example, confer the “*power of assuming a red colour*” in a plant. This “*power of assuming*” was held to be different from the actual “*production of ... the red colouring matter,*” which was “more or less in the same manner as the action of enzymes or ferments.” Thus, de Vries distinguished three hierarchically arranged elements: hereditary character units (later known as the genotype), executive functions catalyzing chemical changes (associated with enzyme-like elements), and observed characters themselves (later known as the phenotype; Greek: *phainomai* = I appear).

The Swiss physiologist Johann Friedrich Miescher, discoverer of what we now call DNA, suggested in 1892 that, by virtue of structural differences, *one class* of “huge molecules” should suffice to convey the hereditary character units [33]:

In ... huge molecules ... the many asymmetric carbon atoms provide a colossal amount of stereo-isomerism. In them all the wealth and variety of heredity transmissions can find expression just as all the words and concepts of all languages can find expression in twenty-four to thirty alphabetic letters. It is therefore quite superfluous to make the egg and sperm cell a storehouse of countless chemical substances each of which carries a particular heredity quality.... My own research has convinced me that the protoplasm and the nucleus, far from consisting of countless chemical substances, contain quite a small number of chemical individuals which are likely to be of a most complicated chemical structure.

In 1893 in *An Examination of Weismannism* [34] Romanes, summarized what, for many at that time, were obscure abstractions:

But the theory of Pangenesis does not suppose the future organism to exist in the egg-cell *as a miniature* (Romanes' italics): it supposes merely that every part of the future organism *is represented* in the egg-cell by corresponding material particles. And this, as far as I can understand, is exactly what the theory of germ-plasm [of Weismann] supposes; only it calls the particles ‘molecules,’ and seemingly attaches more importance to the matter of variations in their arrangement or ‘constitution,’ whatever these vague expressions may be intended to signify.

Along similar lines in 1894 in *Materials for the Study of Variation* [35], Bateson stated:

The body of one individual has never *been* the body of its parent,... but the new body is made again new from the beginning, just as if the wax model had gone back to the melting pot before the new model was begun. [Bateson's italics]

Agreeing with de Vries, Bateson considered enzymes affecting color [36]:

We must not suppose for a moment that it is the ferment [enzyme], or the objective substance [color], which is transmitted [from generation to generation]. The thing transmitted can only be the *power or faculty to produce* the ferment or objective substance.

Rather than “hereditary characters,” Bateson spoke of “character units,” “genetic factors,” or “elements.” Grasping for what we now know as DNA, in 1913 in *Problems in Genetics* [37] he elaborated on the terms “genotype” and “phenotype” just introduced by the Danish botanist Wilhelm Johannsen:

Of the way in which variations in the ... composition of organisms are caused we have ... little real evidence, but we are beginning to know in what such variations must consist. These changes must occur either by the addition or loss of [genetic] *factors* [to/from the gametes and hence to/from the offspring formed by those gametes]. We must not lose sight of the fact that, though the [genetic] factors *operate by the production of*: (i) enzymes, (ii) bodies on which these enzymes can act [substrates], and (iii) intermediary substances necessary to complete the enzyme action [enzyme cofactors], yet these bodies themselves [all three of the above] *can scarcely themselves* be genetic factors, but *consequences* of their existence.

The genetic factors could be newly acquired and had a continuity from generation to generation, being apportioned symmetrically during cell division. Bateson continued:

What then are the [genetic] factors themselves? Whence do they come? How do they become integral parts of the organism? Whence, for example, came the *power* which is present in the White Leghorn of destroying ... the pigment in its feathers? That power is now a definite possession of the breed, present in all its germ-cells, male and female, taking part in their symmetrical divisions, and passed on equally to all, as much as [is passed on] the protoplasm or any other attribute of the breed.

It seems that Bateson never stated that a genetic factor conveyed “*information for*”, but only conveyed the “*power to cause*” enzymes, substrates and enzyme-cofactors to appear in the offspring. However, in addresses in 1914 and later he sensed that genetic factors (now interpreted as ‘genes’) might be some “phenomenon of arrangement” (now interpreted as ‘sequence’), and that variation (now interpreted as mutation; Latin, *mutatio* = I change) could occur when the arrangement changed. Indeed, he almost latched on to the idea of a code linking the observed characters of an organism (its phenotype) to “elements upon which they depend” (now interpreted as genes) [38]:

The allotment of characteristics among offspring is ... accomplished ... by a process of cell-division, in which numbers of these characters, or rather the elements upon which they depend, are sorted out among the resulting germ-cells in an orderly fashion. What these elements, or factors as we call them, are we do not know. That they are in some way directly transmitted by the material of the ovum and of the spermatozoon is obvious, but it seems to me unlikely that they are in any simple or literal sense material particles. I suspect rather that *their properties depend on some phenomenon of arrangement....* That which is conferred in variation must rather itself be a change, not of material, but of arrangement. ... By the re-arrangement of a very moderate number of things we soon reach a number of possibilities practically infinite.

Hering had spoken similarly, albeit more abstractly, in 1870 [15]. Thus, we can interpret “*power to cause*” as showing that Bateson (like Darwin, Galton, Hering, and Butler) understood the information concept, although he never used the word in this context (but he used the word in other contexts). He understood that different arrangements can convey different pieces of information. He was prepared to

believe his factors were molecules of the type we would today call “macromolecules,” but he was far from calling them “*informational* macromolecules”. There was not so much a conceptual gap as a semantic gap. The latter was bridged at the dawn of the “classical period” in molecular biology (1945-1966) by Austrian physicist Erwin Schrödinger [39–42]. His biological ideas grew from early studies of the works of German biologist Richard Semon, whose ideas, in turn, can be traced back to Hering and Butler [43].

Translation

Regarding the “difficulty of conceiving the transition of germinal substance into somatic substance,” Romanes considered “a transition of A into B ... where B is a substance which differs from A almost as much as a woven texture differs from the hands that weave it” [34]. But he could not see A and B as *languages*. He could not see the transition as *translation* from one form of information to another. The elements of language are, according to Butler [24]:

A sayer, a sayee, and a convention, no matter what, agreed upon between them as inseparably attached to the idea which it is intended to convey – these comprise all the essentials of language. Where these are present there is language; where any of them are wanting there is no language.

The need for a code or cipher (an agreed convention) to relate the information carried from generation to generation in the gametes (derived from germ cells in the gonads) to the actual characters displayed by an organism (e.g. parts of the body) was hinted at by Weismann in 1904 [44]:

The germ substance never arises anew, but is always derived from the preceding generation – that is ... the continuity of the germ plasm. ... We ... assume that the conditions of all the parts of the body [are] ... reflected in the corresponding primary constituents of the germ plasm and thus in the germ cells. But, as these primary constituents are *quite different* from the parts themselves, they would require to vary in quite a different way from that in which the finished parts had varied; which is very like supposing that an English telegram to China is there received in the Chinese language.

This implied that one form of information, namely the “primary constituents,” must be *translated* into another form, namely, “the parts.” We would now relate “primary constituents” to DNA (the genotype), and “the parts” (the phenotype) to certain classes of macromolecules, especially proteins.

The four main classes of macromolecules found in living systems, – nucleic acids, proteins, complex fats (lipids), complex sugars (carbohydrates) – all convey information. However, the former two are special in that the latter two are dependent upon them for their construction. Nucleic acids and proteins construct both themselves, and lipids and carbohydrates (Fig. 1.4).

Nucleic acids and proteins have sequences that are so varied that the ‘machines’ that construction them (enzymes which are themselves mainly proteins) have to be

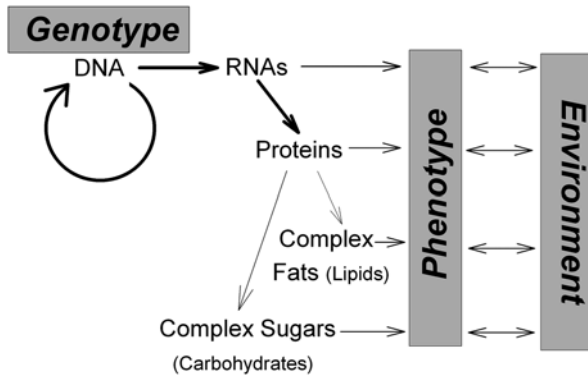


Fig. 1.4 Template-dependent reactions (heavy arrows), which require nucleic acids, result in the production from subunits (base-containing nucleotides, or amino acids) of ‘informational macromolecules’ (DNA, RNA and proteins). Proteins (enzymes) catalyze both these template-dependent reactions and the mechanical assembly from subunits (e.g. simple sugars, fatty acids) of other macromolecules – complex fats (lipids) and complex sugars (carbohydrates). When a nucleic acid acts as a linear template for the formation of another nucleic acid, information is copied (*replicated* or *transcribed*). When a nucleic acid acts as a linear template for the formation of a protein, one form of information is decoded (*translated*) to give rise to another form of information. RNAs, proteins, lipids and carbohydrates collectively compose the conventional Darwinian phenotype through which members of a species interact with (track) their changing environment through the generations. The curved arrow indicates the self-reproducing nature of the genotype

programmed with linear information templates (nucleic acids). On the other hand, lipids and carbohydrates have structural regularities that allow their mechanical assembly by enzymes in the absence of linear information templates. Thus, nucleic acids and proteins can be considered as ‘informational macromolecules,’ to denote their higher order of information.

Butler supposed “a memory to ‘run’ each gemmule” [18], but, of course, he knew the material basis of inherited information no more than he knew the material basis of mental information. We now know that DNA is the former, but have little understanding of the latter. Butler implied that both forms of storage were the same (Occam’s principle; see Chapter 3). Current evidence does not support this.

As DNA contains the information for the construction of, say, a hand, so DNA should contain the information needed for a particular pattern of cerebral ‘wiring’ (an “anatomical plan of the ganglia”). But it would be a big step to go from this wiring pattern to a particular instinct. And how such instinctual memories might relate to memories acquired during *individual* lifetimes (“cerebral alterations”) remains mysterious (see Chapters 19 and 20). Romanes pointed out in 1881 [21]:

We can understand, in some measure, how an alteration in brain structure when once made should be permanent, ... but we cannot understand how this alteration is transmitted to progeny through structures so unlike the brain as are the products of the generative glands [gametes]. And we merely stultify ourselves if we suppose that the problem is brought any nearer to solution by asserting that a future individual while still in the germ has already participated, say, in the cerebral alterations [changes in brain structure] of its parents.

Butler seems to have been the first to extrapolate the principle that ‘higher’ forms of life evolve from ‘lower’ forms, to suggest that eventually, beyond man, machines made by man would take over the planet [45]. Today we would equate this with computers. Butler’s genius was in seeing that inheritance and mental function, like today’s computers, are all “modes of memory” [18]:

Life is that property of matter whereby it can remember. Matter which can remember is living; matter which cannot remember is dead. *Life, then, is memory.* The life of a creature is the memory of a creature. We are all the same stuff to start with, but we remember different things, and if we did not remember different things we should be absolutely like each other. As for the stuff itself, we know nothing save only that it is ‘such as dreams are made of’. [Butler’s italics]

Variation

‘Memory’ refers either to the physiological process by which stored information is retrieved, or to the actual stored information itself. In the former sense, the storage of information is implicit. In the latter sense, memory *is* stored information. Whatever the sense, the concepts of memory and of stored information are very close. At some point in time information is selected and assigned to a store, and at a later point in time it may be retrieved. But storage of information is not enough. Information must be *safely* stored. To select and to store is not to preserve. On a hot summer’s day a package of ice-cream in your shopping bag will not be preserved. Stored information must also be *preserved* information. When preservation is not perfect, there is *variation*.

Butler declared that “the ‘Origin of Variation,’ whatever it is, is the only true ‘Origin of Species’” [18]. He knew the mechanisms by which Nature’s expressions vary from generation to generation, no more than he knew the mechanisms by which mental expressions vary from day to day. Nor did he know the scope of variation [35]. Variation can only occur in that which already exists, and the nature of that which already exists, limits the scope of its variation. Just as the scope of a dice is limited by its structure, so the scope of variation of a living form is limited. From time to time a human is born with six fingers but, as far as we know, no human has ever been born with a hundred fingers, or with feathers (see Chapter 3).

The Victorians delighted in games of chance. They appreciated that variation in landing position of a balanced dice is random. The variation in landing position of an unbalanced dice is partly non-random. So, from first principles, it was supposed that variation in living forms might be either a chance event, or biased. Whether unbiased or biased, a variation might be beneficial, deleterious, or somewhere in between. Within its scope, the dice of life might be multifaceted allowing fine gradations in variation. Alternatively variations might be discrete (e.g. the usual six-sided dice). And would the dice of life be thrown? If so, by what agency? If not thrown, then any variation that occurred would be considered ‘spontaneous’ and perhaps unbiased. If the dice of life were thrown, then there would be a thrower.

Would this agency be internal or external to the organism? In either case, could the agency direct the bias to the adaptive advantage of the organism? In other words, could the agency ‘design’ the organism?

Among these alternatives, current evidence shows (see Chapter 7) that variation is a spontaneous property of matter, but that it may be affected, both in type and amount, by external agents, such as X-radiation. Variation is usually unbiased. It is not directed. There is no design by means of variation, be it internally or externally induced. If an organism is closely adapted to its environment then a variation is unlikely to be beneficial. If the environment has recently changed dramatically, then a variation is more likely to be beneficial. But variations, *per se*, do not occur “for the good of the organism” (teleology).

However, the parallels he had drawn between memory and heredity suggested to Butler an alternative that had some plausibility at that time. He opted for an agency *internal* to the organism that would, in small steps, bring about variations that would accumulate to the advantage of the organism. This alternative had gained some scientific respectability in both England (e.g. Spencer), and continental Europe (e.g. Nägeli), and some came to call it “orthogenesis,” implying an innate tendency to progressive development [46].

Butler’s case rested on grounds that were both aesthetic and logical. Darwin’s argument that natural selection would cruelly send the weaker to the wall and select the fittest to survive, sufficed “to arouse instinctive loathing; ... such a nightmare of waste and death is as baseless as it is repulsive” [24]. To buttress this feeling, Butler turned to an argument that will be made here in a different context in Chapter 15. Microscopic studies of unicellular organisms, such as amoebae, showed them to possess organelles analogous to the organs of multicellular organisms. Thus, an amoeba, far from being an amorphous mass of protoplasm, was seen to extrude ‘arms’ (pseudopodia), fashion a ‘mouth,’ and digest its prey in a prototypic ‘stomach’ (digestive vacuole). If it could achieve this degree of sophistication, then perhaps it could, in an elementary way, also think? Butler was quite open with his premises [47]:

Given a small speck of jelly with ... some power of slightly varying its actions in accordance with slightly varying circumstances and desires – *given* such a jelly-speck with a power of assimilating other matter, and thus, of reproducing itself, *given* also that it should be possessed of a memory

When something was counter-intuitive Butler spurned common-sense, but when it served his purpose he often appealed to “people in ordinary life,” or “plain people” [24]:

The difference between Professor Weismann and, we will say, Heringians consists in the fact that the first maintains the new germ-plasm, when on the point of repeating its developmental process, to take practically no cognisance of anything that has happened to it since the last occasion on which it developed itself; while the latter maintain that offspring takes much the same kind of account of what has happened to it in the persons of its parents since the last occasion on which it developed itself, as people in ordinary life take things that happen to them.

In daily life people let fairly normal circumstances come and go without much heed as matters of course. If they have been lucky they make a note of it and try to repeat their success. If they have been unfortunate but have recovered rapidly they soon forget it; if they have suffered long and deeply they grizzle over it and are scared and scarred by it for a long time. The question is one of cognisance or non-cognisance on the part of the new germs, of the more profound impressions made on them while they were one with their parents, between the occasion of their last preceding development and the new course on which they are about to enter.

Thus, Butler downplayed chance (“luck”) and championed an *intrinsic* capacity for bias (“cunning”) as the means by which advantageous characters acquired by parents would be transmitted to their children. In short, he appealed to the doctrine of the inheritance of acquired characters that had been advocated in France in 1809 by Jean Baptiste Pierre de Lamarck. Butler even went so far as to assert, in Romanes’ words [21], “that a future individual while still in the germ has already participated ... in the cerebral alterations of its parents.” However, like Hering, Butler maintained a strictly materialist position. He used the words “intelligent” and “design,” often separately, and sometimes together, but never in a way as to suggest the involvement of an agency *external* to the organism [24]:

The two facts, evolution and design, are equally patent to plain people. There is no escaping from either. According to Messrs. Darwin and Wallace, we may have evolution, but are in no account to have it mainly due to intelligent effort, guided by ever higher and higher range of sensations, perceptions and ideas. We are to set it down to the shuffling of cards, or the throwing of dice without the play, and this will never stand. According to the older men [e.g. Lamarck], the cards [inheritance] did count for much, but play [environment] was much more. They denied the teleology of the time – that is to say, the teleology that saw all adaptation to surroundings as part of a plan devised long ages since by a quasi-anthropomorphic being who schemed everything out much as a man would do, but on an infinitely vaster scale. This conception they found repugnant alike to intelligence and conscience, but, though they do not seem to have perceived it, they left the door open for a design more true and more demonstrable than that which they excluded.

Butler soared beyond the comprehension of his contemporaries on the wings of his conceptual insight that heredity was the transfer of stored information. Although his dalliance with Lamarckism has not won support (for caveats on this see Chapter 18), the robustness of the underlying idea led him close to solutions to fundamental biological problems, such as the origin of sex, the sterility of hybrids (see Chapter 3), and aging (see Chapter 17). In his times, Butler’s *correct* equation of heredity with memory was beyond the pale, but his *incorrect* Lamarckism seemed plausible. Spencer was, at heart, a Lamarckian and, at Darwin’s behest, Romanes spent several years fruitlessly seeking to show that gemmules containing information for acquired characters could be transferred from normal parental tissues (soma) to the gonads (germ-line). And as late as 1909 the German evolutionist, Ernst Haeckel, was proclaiming that “natural selection does not of itself give the solution of all our evolutionary problems. It has to be taken in conjunction with the transformism of Lamarck, with which it is in complete harmony” [48].

Latency

One good reason why a character is not perceived in an organism is that the information for that character is not present. Yet, information can be present but remain undetected because it is not expressed. Such information is referred to as “latent.” Francis Galton suggested in 1872 that individuals consist of “two parts, one of which is *latent*, and only known to us by its effects on posterity, while the other is *patent*, and constitutes the person manifest to our senses” [30]. Darwin [10] believed that “every character which occasionally reappears is present in a latent form in each generation,” and in 1877 Butler, tongue in cheek, quoted him with approval in this respect [18]:

We should expect that reversion should be frequently capricious – that is to say, give us more trouble to account for than we are either able or willing to take. And assuredly we find it so in fact. Mr. Darwin – from whom it is impossible to quote too much or too fully, inasmuch as no one else can furnish such a store of facts, so well arranged, and so above all suspicion of either carelessness or want of candour – so that, however we may differ from him, it is he himself who shows us how to do so, and whose pupils we all are – Mr. Darwin writes: ‘In every living being we may rest assured that a host of long-lost characters lie ready to be evolved under proper conditions’ (does not one almost long to substitute the word ‘memories’ for the word ‘characters’?) ‘How can we make intelligible, and connect with other facts, this wonderful and common capacity of reversion – this power of calling back to life long-lost characters?’ ... Surely the answer may be hazarded, that we shall be able to do so when we can make intelligible the power of calling back to life long-lost memories. But I grant that this answer holds out no immediate prospect of a clear understanding.

Sometimes the information for an entire organism can be latent, as in the case of the AIDS virus genome that, under certain circumstances, becomes merged with human genomes but is not expressed. Darwin was thinking of characters, distributed among members of a species, some of which might remain latent for generations until suddenly they re-emerge among the members of a new generation. There are also characters that are usually present in all members of a species, but are only observed at certain times in the life cycle, or under particular environmental conditions. Thus, information for a butterfly is latent in its caterpillar. Information for human secondary sexual characters lies latent until adolescence.

Increased pigmentation of the skin occurs in white people exposed to excess sunlight; this inducible “tanning” is a latent ability. When white people in the northern hemisphere move south their skins darken. On the other hand, for black people, pigment formation is a permanent, “constitutive,” ability. When black people move north their skins do not whiten. Black pigment (melanin) protects skin cells against radiation and in equatorial climes there is still sufficient penetration of light to assist formation of the essential nutrient, vitamin D. When black people move north they risk vitamin D deficiency. Thus, there is a trade-off between the need to protect against radiation, implying more melanin, and the need to form vitamin D, implying less melanin. It is possible that hairless hominoids first evolved in hot equatorial climes, and were constitutively black. Through natural selection (see Chapter 8) those that migrated to and fro, from south to north and back again, acquired the ability to regulate melanin production [49].

All this implies that the expression of a piece of information can be regulated. It may, or may not, be expressed throughout a lifetime (constitutively), or its expression may be confined to certain periods of life or to certain environments. Information may be expressed fully or partially, sometimes in a precisely calibrated fashion, and sometimes more randomly. Just as a light from a forest dwelling may be perceived only when its path is not blocked by trees, so the penetrance of one piece of information (from within the organism to the phenotype perceived from outside the organism) can sometimes depend on the expression of other pieces of information that accompany it within the organism.

Summary

Heredity requires stored information (memory). Early ideas that this information storage might be “a phenomenon of arrangement” can now be interpreted in terms of the linear arrangement of bases in nucleic acid molecules. Genes differ from each other because they have different arrangements of these base building blocks. Typically, genes contain information for the construction of executive units (proteins) that facilitate chemical reactions whose products give rise to the observed features of an organism – its phenotype. Thus, the twentieth century’s obsession with genes is understandable. However, the gene-centered viewpoint spawned a corresponding bioinformatics very different from the emergent, non-gene-centered, bioinformatics. The tasks of EB – the new bioinformatics – are to identify *all* forms of hereditary information, both genic and non-genic, and show how conflicts between different forms are accommodated to permit the proper functioning of biological entities (genes, cells, individuals, species, genera). Because different types of information, be they mental, textual or hereditary, have common characteristics, our understanding of familiar forms (e.g. textual information) may assist our understanding of less familiar forms (e.g. hereditary information). Of high importance is that information, once selected and assigned to a store, must be *preserved*. If preservation is imperfect, stored information will change, and anything dependent on that information may also change. When hereditary information changes, parental characters may vary in offspring, and biological evolution can occur. Such ideas were elaborated in the nineteenth century by Hering and Butler, and passed by way of Semon and Schrödinger to illuminate the “classical period” in molecular biology (1945-1966).

References

1. Pope A (1711) *Essay on Criticism*. Macmillan, London (1896)
2. Darwin C (1859) *The Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. John Murray, London, p 490 [As later acknowledged

- by Darwin, aspects of his natural selection hypothesis had been anticipated by the Scottish horticulturalist Patrick Matthew in 1831; see Cock & Forsdyke (2008).]
3. Bateson W (1899) Hybridization and cross-breeding as a method of scientific investigation. In: Bateson B (ed) *William Bateson, F.R.S. Naturalist. His Essays and Addresses*. Cambridge University Press (1928) pp 161–171
 4. Darwin C (1872) Letter to Moggridge. In: Darwin F, Seward AC (eds) *More Letters of Charles Darwin*. John Murray, London (1903) Vol 1, p 337
 5. Gelbart WM (1998) Databases in genome research. *Science* 282:659–661
 6. Delbrück M (1971) Aristotle-totle-totle. In: Monod J, Borek F (eds) *Of Microbes and Life*. Columbia University Press, New York, pp 50–55
 7. Lyell C (1863) *Geological Evidences on the Antiquity of Man with Remarks on Theories of the Origin of Species by Variation*. Childs, Philadelphia, p 467
 8. Mendel G (1865) Versuche uber Pflanzen Hybriden. *Verhandlung des naturforschenden Vereines in Brunn* 4:3–47
 9. Nägeli C von (1884) *Mechanisch-physiologische Theorie der Abstammungslehre*. Munich, p 73 [Translated in Darwin F (1908) British Association for the Advancement of Science. Presidential Address.]
 10. Darwin C (1868) Provisional hypothesis of pangenesis. In: *The Variation of Animals and Plants under Domestication*. John Murray, London, Chapter 27
 11. Dawkins R (1976) *The Selfish Gene*. Oxford University Press, Oxford
 12. Holmquist G (1989) Evolution of chromosome bands: molecular ecology of noncoding DNA. *Journal of Molecular Evolution* 28:469–486
 13. Huxley TH (1869) The genealogy of animals. In: *Darwiniana. Collected Essays*. Macmillan, London (1893)
 14. Roux W (1881) *Der Kampf der Theile im Organismus*. Leipzig
 15. Hering E (1870) *Über das Gedachtniss als eine allgemeine Function der organisirten Materie*. Karl Gerold's Sohn, Vienna [On Memory as a Universal Function of Organized Matter. A lecture delivered at the anniversary meeting of the Imperial Academy of Sciences at Vienna, translated in Butler S (1880) *Unconscious Memory*. David Bogue, London, pp 97–133.]
 16. Butler S (1985) *The Notebooks of Samuel Butler*. Jones HF (ed) Hogarth Press, London, pp 58–59, 70–71
 17. Bateson W (1909) Heredity and variation in modern lights. In: Seward EC (ed) *Darwin and Modern Science*. Cambridge University Press, Cambridge, pp 137–151
 18. Butler S (1878) *Life and Habit*. Trübner, London, pp 134, 187, 196–197, 263, 299–300 [The book was actually released in December 1877.]
 19. Butler S (1872) *Erewhon or Over the Range*. Penguin Books, Harmondsworth (1935), pp 197–198
 20. Romanes GJ (1884) *Mental Evolution in Animals, with a Posthumous Essay on Instinct* by Charles Darwin. Appleton, New York, p 131
 21. Butler S (1884) *Selections from Previous Works, with Remarks on Mr. G. J. Romanes' "Mental Evolution in Animals," and a Psalm of Montreal*. Trübner, London, pp 228–254 [G. H. Lewes was a Victorian Philosopher/Biologist and member of the Physiological Society, who is best known as the partner of authoress, George Eliot.]
 22. Butler S (1880) *Unconscious Memory*. David Bogue, London, pp 252, 269–273 [Tekke Turcomans were members of a fierce nomadic tribe located to the north of what was then Persia. *Plus ça change, c'est la même chose!*]
 23. Butler S (1887) *Luck, or Cunning, as the Main Means of Organic Modification? An Attempt to Throw Additional Light upon the Late Mr. Charles Darwin's Theory of Natural Selection*. Trübner, London, p. 308
 24. Butler S (1914) *The Humour of Homer and Other Essays*. Kennerley, New York, pp 209–313
 25. Provine WB (1971) *The Origins of Theoretical Population Genetics*. University of Chicago Press, Chicago
 26. Olby R (1974) *The Path to the Double Helix*. University of Washington Press, Seattle

27. Crookes W (1866) On the application of disinfectants in arresting the spread of the cattle plague. In: *Appendix to Third Report of the Commissioners Appointed to Inquire into the Origin and Nature, etc. of the Cattle Plague*. Houses of Parliament, London, pp 187–201 [I have made this document available online: <http://post.queensu.ca/~forsdyke/rindpst3.htm>]
28. Romano TM (2002) *Making Medicine Scientific. John Burdon Sanderson and the Culture of Victorian Science*. John Hopkins University Press, Baltimore, p 63
29. Harris H (1999) *Birth of the Cell*. Yale University Press, New Haven
30. Galton F (1872) On blood-relationship. *Proceedings of the Royal Society* 20:394–402
31. Weismann A (1904) *The Evolution Theory*, Vol 1. Edward Arnold, London, p 411
32. Vries H de (1910) *Intracellular Pangenesis*. Gager CS (translator) Open Court, Chicago [Translated from Vries H de (1889) *Intracellulare Pangenesis*. Fischer, Jena.]
33. Olby R, Posner E (1967) An early reference on genetic coding. *Nature* 215: 556
34. Romanes GJ (1893) *An Examination of Weismannism*. Open Court, Chicago, pp 182–183 [Romanes pointed out that Weismann’s “germ-plasm” and Galton’s equivalent (called “stirp” from the Latin *stirpes* = root), had in common the properties of stability and continuity. However, unlike today’s DNA, “germ-plasm” as originally conceived by Weismann could only vary by the mixing with another germ-plasm through sexual reproduction. Galton held open the possibility that “stirp” might also vary by other means and, to “a very faint degree” (perhaps in deference to his cousin’s theory of pangenesis”) Lamarckian principles were admitted.]
35. Bateson W (1894) *Materials for the Study of Variation Treated with Especial Regard for Discontinuity in the Origin of Species*. Macmillan, London, p 33
36. Bateson W (1908) The methods and scope of genetics. In: Bateson B (ed) *William Bateson, F R S, Naturalist. His Essays and Addresses*. Cambridge University Press, Cambridge (1928), p 317–333
37. Bateson W (1913) *Problems of Genetics*. Yale University Press, New Haven, p 86
38. Bateson W (1914) Presidential address to the British Association, Australia. In: Bateson B (ed) *William Bateson, F R S, Naturalist. His Essays and Addresses*. Cambridge University Press, Cambridge (1928) pp 276–316
39. Schrödinger E (1944) *What is Life? The Physical Aspect of the Living Cell*. Cambridge University Press, Cambridge
40. Portugal FH, Cohen JS (1977) *A Century of DNA*. MIT Press, Cambridge, MA
41. Loewenstein WR (1999) *The Touchstone of Life. Molecular Information, Cell Communication, and the Foundations of Life*. Oxford University Press, Oxford
42. Kay LE (2000) *Who Wrote the Book of Life? A History of the Genetic Code*. Stanford University Press, Stanford
43. Forsdyke DR (2015) “A vehicle of symbols and nothing more.” George Romanes, theory of mind, information, and Samuel Butler. *History of Psychiatry* 26:270–287.
44. Weismann A (1904) *The Evolution Theory*, Vol 2. Edward Arnold, London, p 63
45. Celarius (alias Butler S) (1863) Darwin among the machines. *The Press*, Christchurch, New Zealand [Reprinted in Butler S (1914) *The First Year in Canterbury Settlement with Other Early Essays*. Fifeild, London.]
46. Kellogg VL (1907) *Darwinism Today*. Holt, New York, pp 274–290
47. Butler S (1879) *Evolution, Old and New; or the Theories of Buffon, Dr. Erasmus Darwin, and Lamarck, as Compared with that of Mr. Charles Darwin*. Hardwicke and Bogue, New York, p 40
48. Haeckel E (1909) Charles Darwin as an Anthropologist. In: Seward EC (ed) *Darwin and Modern Science*. Cambridge University Press, Cambridge, pp 137–151 [Decoys in the literature supported Lamarckism. Prominent among these were guinea-pig experiments by the eminent French physiologist Charles-Édouard Brown-Séquard.]
49. Jablonski NG, Chaplin G (2010) Human skin pigmentation as an adaptation to UV irradiation. *Proceedings of the National Academy of Sciences USA* 107:8962–8968

Chapter 2

Chargaff's First Parity Rule

The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously the same. ... Languages, like organic beings, can be classed in groups under groups. ... A language, like a species, when once extinct, never ... reappears. The same language never has two birthplaces The survival or preservation of certain favoured words in the struggle for existence is natural selection.

Charles Darwin (1871) [1]

Consider a creature, whose attributes you are probably familiar with, pausing in a posture similar to that which you are now perhaps adopting:

The cat sat on the mat.
The cat sat on the mat.

You might guess that, since the information is repeated and rhythms, it forms part of some artistic endeavor that we might refer to as poetry. Alternatively, the author may have had little faith in the typesetters and, to make assurance doubly sure, sent the message twice. It is with the latter explanation that we are most concerned.

Error-Detection

Since type-setting errors are usually random and rare, it is likely that, if an error were to occur, it would affect only one of the sentences. Instead of “mat” on the top line you might have seen “hat.” Coming across the two parallel sentences for the first time, and knowing that the repetition was deliberate, how would you decide which was the correct sentence? You might read the top sentence first as the ‘sense’ text, and then read the bottom sentence to check that each of the letters of the alphabet is faithfully matched by its ‘sense’ equivalent (i.e. “a” is always matched by “a;” “t” is always matched by “t,” etc.). You would check that there is *parity* between the two parallel lines.

Finding that “h” in “hat” in the top line is mismatched with “m” in “mat” in the bottom line, you would know that an error had occurred. But, in the absence of other information you would not be able to decide whether “h” in the top line was correct, or that it should be corrected to “m” based on the information in the bottom line. All you would know was that there had been an error. If for some reason you were forced to decide, you might toss a coin. But, in the absence of other information, if you accepted the coin's guidance there would only be a 50:50 chance of your being correct.

In-Parallel Redundancy

To increase the chance of accurate error-*detection*, and hence of accurate error-*correction*, the author might have repeated the sentence, in-parallel, three times. If two lines contained the word “mat” and only one line the word “hat,” then you would prudently choose the former. Your choice would be even more likely to be correct if the author repeated the sentence, in-parallel, four times, and only one line had the word “hat.”

All this requires much redundant information, which both takes up space in the medium conveying the message (in this case, the printed page), and imposes extra labor on the reader. For some purposes it might suffice merely to detect that an error had occurred. Having been alerted, you might then be able to consult other sources of information should the need to distinguish between “hat” and “mat” be critical. For example, there are 25 possible alternatives to “h” as the first letter in “hat.” Of the resulting words – aat, bat, cat, dat, eat, fat, gat, etc. – several can be excluded because there is no English dictionary equivalent, others can be excluded syntactically (e.g. “eat” is a verb not an object), and others can be excluded contextually (e.g. neighboring text might refer to mat, not to hat).

Thus, there is much to be gained by duplication, but with increasing levels of redundancy (triplication, quadruplication, etc.) the gains are less evident. At face value, this appears to be a strategy adopted by biological systems for accurately transferring information from generation to generation. Genetic messages are sent as duplexes, but with a ‘twist’ in more than one sense.

DNA Structure

Contrasting with the 26 letter English alphabet, the DNA alphabet has four letters – the bases **A** (adenine), **C** (cytosine), **G** (guanine), and **T** (thymine). Thus, a message in DNA might read:

TACGACGCCGATAGCGTCGTA (2.1)

With duplex in-parallel redundancy, the message could be sent from generation to generation as:

$$\begin{array}{l} \text{TACGACGCCGATAGCGTCGTA} \\ \text{TACGACGCCGATAGCGTCGTA} \end{array} \quad (2.2)$$

We can refer to this as “sense-sense” pairing since, like the cat sentences, both sentences read the same (i.e. **A** is matched with an **A**, and **T** is matched with a **T**, etc.). However, when arriving at their model for the duplex structure of DNA in 1953, James Watson and Francis Crick [2] took into account a ‘rule’ enunciated by Erwin Chargaff. He and his coworkers had found that bases did not match themselves. They matched other bases. In DNA, base **A** is quantitatively equivalent to base **T**, and base **G** is quantitatively equivalent to base **C**. Chargaff speculated in 1951 that this regularity might be important for DNA structure, noting [3]:

It is almost impossible to decide at present whether these regularities are entirely fortuitous or whether they reflect the existence in all DNA preparations of certain common structural principles, irrespective of far-reaching differences in their individual composition and the absence of an easily recognizable periodicity.

In 1952 Canadian biochemist Gerard Wyatt went further, suggesting a spiral structure [4]:

If you have a spiral structure ... [it is quite possible to have the bases] sticking out free so that they don't interfere with each other. Then you could have a regular spacing down the backbone of the chain, in spite of the differences in sequence.

Later he added [5]:

One is tempted to speculate that regular structural association of nucleotides of adenine with those of thymine and those of guanine with those of cytosine ... in the DNA molecule requires that they be equal in number.

If the top message were ‘sense,’ the bottom message could be considered as ‘antisense.’ The above ‘sense’ message could then be sent in duplex form as:

$$\begin{array}{l} \text{TACGACGCCGATAGCGTCGTA} \quad \text{‘sense’} \\ \text{ATGCTGCGGCTATCGCAGCAT} \quad \text{‘antisense’} \end{array} \quad (2.3)$$

Error-detection would still be possible. In this “sense-antisense” error-detection system, errors would be detected when an **A** was matched with **G**, **C** or another **A**, rather than with **T**. Similarly, if **G** was matched with **A**, **T** or another **G**, rather than with **C**, another error would have been detected.

That a base would not match itself was also right for chemical reasons. Just as the letters of the standard alphabet come as either vowels or consonants, so the bases of DNA are either purines (**A** and **G**) or pyrimidines (**C** and **T**; Table 2.1).

Vowels and consonants often match or ‘complement’ to the extent that vowels separate consonants giving words a structure, which facilitates their pronunciation. Purines are bigger than pyrimidines, and the chemical models that Watson and

Table 2.1 Symbols for groups of DNA bases. When picking symbols for collectivities of bases some logic is attempted. Since purines and pyrimidines both begin with the same letter, the second consonants **R** and **Y** are employed. Watson-Crick base-pairing involves interactions that are either 'weak' (**W**) in the case of **A-T** base-pairs, or 'strong' (**S**) in the case of **G-C** base-pairs

	R (Purines)	Y (Pyrimidines)
W (Weak)	A (Adenine)	T (Thymine)
S (Strong)	G (Guanine)	C (Cytosine)

Crick constructed required that a purine always match or 'complement' a pyrimidine. A molecular complex of two purines would be too big. A molecular complex of two pyrimidines would be too small. The solution is that the purine **A** pairs with the pyrimidine **T** and the purine **G** pairs with the pyrimidine **C**. By match we mean an actual structural (i.e. chemical) pairing. Although your eyes can detect that **A** on one line matches **T** on the other, inside our cells it is dark and there are no eyes to see. Matching is something molecules do for themselves by recognizing complementary shapes on their pairing partners, just as a key recognizes the lock with which it 'pairs.'

The key-lock analogy will serve us well here; however, pairing may also require subtleties such as similar molecular vibrations, or resonances [6]. To visualize this, in 1941 the geneticist Herman Muller likened molecular mixtures to imaginary mixtures of floating electromagnets each charged with an alternating current of a particular frequency. Since magnet polarity would depend on the direction of current flow, the polarity of each magnet would be constantly changing at a frequency determined by the frequency of the alternating current [7]:

If we had a heterogenous mixture of artificial electromagnets, floating freely about and having different frequencies of reversal of sign, those of the same frequency would be found eventually to orient towards and attract one another, specifically seeking each other out to the exclusion of others.

Of course, the final twist of Watson and Crick was, literally, a twist. The two sequences in DNA are two molecular strands that are wound round each other to form a spatially compact helix (Fig. 2.1). Perhaps the most famous throwaway line ever written came at the end of Watson and Cricks' first paper [2]. Here, as an apparent afterthought, they casually noted: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material." In other words, they were claiming not only to have discovered the fundamental structure of genetic information, but also to have discerned from that structure how the information would be faithfully replicated. When the underlying *chemistry* was understood, the *physiology* could be explained – a triumph for the 'reductionist' approach. They had shown how the chemical structure of DNA provided a basis for the continuity of inherited characteristics from organism

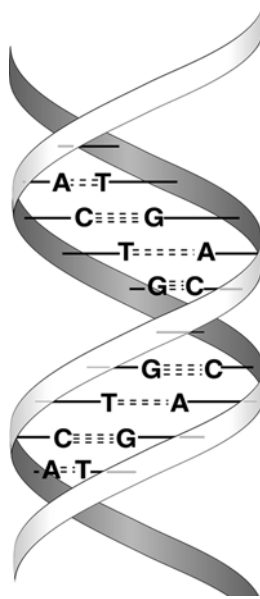


Fig. 2.1 Double helix model for DNA. The base ‘message’ is written on two helical strands, which are shown here as twisted ribbons – the ‘medium.’ Bases are arranged internally so that an **A** on one strand pairs with a **T** on the other strand (and vice versa), and a **G** on one strand pairs with a **C** on the other strand (and vice versa). The bases are attached to the strands by strong bonds, whereas the base-pairing interactions involve weak bonds (shown as dashed lines). Chemically, the bases are like flat discs that ‘stack’ on top of each other within the helical strands, like a pile of coins (rouleau). These stacking interactions stabilize the double-helical structure and, being largely entropy-driven (see Chapter 15), become greater as temperature increases. However, in solution at high temperatures (e.g. 80°C) this can be overcome, and the two strands separate (i.e. the duplex ‘melts’) to generate free single strands. This figure was kindly adapted by Richard Sinden from his book *DNA Structure and Function* [8]

to organism, and from cell to cell within an organism. In Bateson’s words, they had discovered how “the allotment of characteristics among offspring is ... accomplished.” This was made explicit in a second paper [9]:

Previous discussions of self-duplication [of genetic information] have usually involved the concept of a template or mould. Either the template was supposed to copy itself directly or it was to produce a ‘negative,’ which in its turn was to act as a template and produce the original ‘positive’ once again. ... Now our model for deoxyribonucleic acid is, in effect, a *pair* of templates [Watson and Crick’s italics], each of which is complementary to the other. We imagine that prior to duplication ... the two chains unwind and separate. Each chain then acts as a template for the formation on to itself of a new companion chain, so that eventually we shall have two pairs of chains, where we only had one before.

Armed with this powerful clue, within a decade biochemists such as Arthur Kornberg in the USA had shown Watson and Crick to be correct, and had identified key enzymes (e.g. DNA polymerase) that catalyze DNA replication [10]. The stunning novelty of the Watson-Crick model was not only that it was beautiful, but

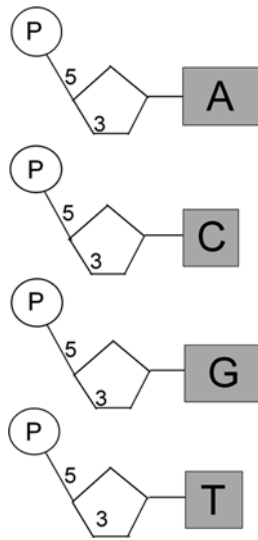


Fig. 2.2 The four nucleotide ‘building blocks’ of which DNA is composed. Each base is connected by a pentose sugar (pentagon) to a phosphate (circle). The purine bases (A and G) are shown as larger boxes than the pyrimidine bases (T and C). Nucleotides have in common a pentose sugar and a phosphate, and differ in their bases. The pentose sugar in DNA is deoxyribose (hence “deoxyribonucleic acid” = “DNA”). The sugar in RNA is ribose (hence “ribonucleic acid” = “RNA”). Pentose sugar carbon atoms are numbered to indicate the third and fifth. In RNA T is replaced by U (uracil) which, like T, pairs with A

that it also explained so much of the biology of heredity. One strand is the complement of the other, so that the text of one strand can be inferred from the text of the other. If there is an error in one strand, then there is the potential for its repair on the basis of the text of the opposite strand. When the cell divides the two strands separate. New ‘child’ strands are synthesized from nucleotide ‘building blocks’ corresponding to A, C, G and T. Each of these blocks, consisting of phosphate, ribose and a base (Fig. 2.2), replaces the former pairing partners of the separated strands, so that two new duplexes identical to the parental duplex are created. In each duplex one of the parental strands is conserved, being paired with a freshly synthesized child strand (Fig. 2.3).

All nucleotide ‘building blocks’ have in common phosphate and ribose, which continue the phosphate-ribose ‘medium,’ upon which the base ‘message’ or ‘pattern’ is ‘written.’ Thus, *any* nucleotide can serve to ensure continuity of the phosphate-ribose medium, and the message itself is determined only by which particular base-containing nucleotide is placed in a particular position. This, in turn, is determined by the complementary template provided by the parental DNA strands, which are recognized according to the specific base-pairing rules (Fig. 2.4).

The message you are now reading was imposed by the stepwise sequential addition of letters to a *pre-existing* medium (the paper). Each letter required a small local piece of the medium, but that medium was *already* in place when the letter

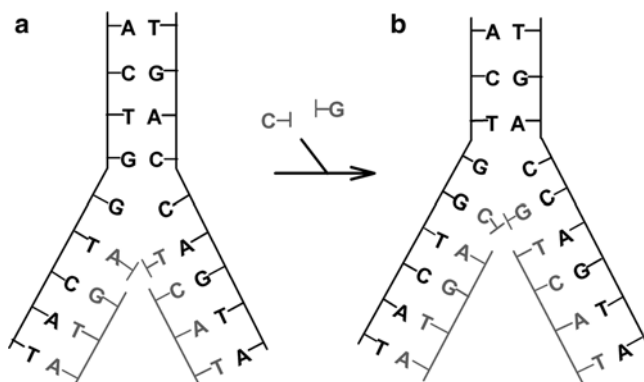


Fig. 2.3 DNA replication. Individual strands of a parental duplex partially separate, and fresh child strands are synthesized by sequential introduction and ‘zipping’ up (polymerization) of complementary base nucleotide ‘building blocks’ (shown in gray). Thus, DNA is a linear polymer (Greek: *poly* = many and *meros* = part) of nucleotide units (i.e. it is a polynucleotide). In (a), at the point of child strand growth in the left limb of the replication fork (inverted Y), an A (gray) is about to be joined to a G (gray). This join is complete in (b), where the two parental strands are further separated. The new duplexes each contain one parental strand (black), and one child strand (gray). Details of synthesis in the right limb are dealt with in Chapter 6 (Fig. 6.6)

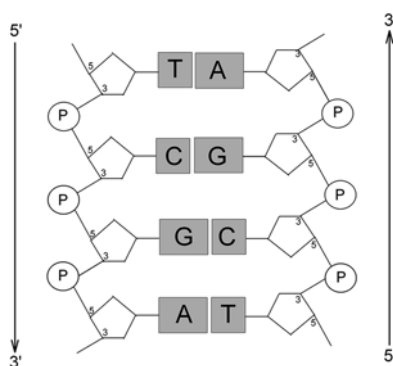


Fig. 2.4 Base-pairing between the two strands of a DNA duplex. The larger purines pair with the smaller pyrimidines, so the distance between the two strands remains relatively constant. Because of this size difference, the flat bases do not just ‘stack’ (form a ‘pile of coins’) above and below neighboring bases in the same strand (e.g. note that the two G’s on separate strands overlap each other, and thus partially stack together). Rather, base-pairs ‘stack’ with base-pairs, some better than others. Numbering associated with the pentose sugars indicates that strands have distinct directionality (polarity) that, by convention, is written from 5’ to 3’ (see vertical arrows). Thus, the left strand reads 5’TCGA3’ from top to bottom. The right strand reads 5’TCGA3’ from bottom to top. The two strands are described as “antiparallel.” Since this short duplex *as a whole* has symmetry (i.e. putting the purines first, the order is A-T base pair, G-C base pair, G-C base pair, A-T base pair), then it can be said to show palindromic properties (see Chapter 4)

arrived. The medium had already been generated. When DNA is synthesized each base 'letter' arrives in a pre-existing association with a small piece of the medium (phosphate-ribose) that it locally requires. Thus, the message and the medium are generated *at the same time*. The message and the medium increase in length simultaneously. Remarkably, all this had been sensed by Muller, who had mentored Watson in the 1940s. In 1936, while attempting to defend Russian genetics against Lysenko (see Epilogue), he distinguished less and more variable parts of gene structure, the latter comprising its "specific pattern" [11]:

The gene is, as it were, a modeller, and forms an image, a copy of itself, next to itself, and since all genes in the chain do likewise, a duplicate chain is produced next to each original chain, and no doubt lying in contact with a certain face of the latter. ... There are thousands of different levels of genes, i.e. of genes having different patterns, ... and ... each of these genes has to reproduce its own specific pattern out of the surrounding materials common to them all. When, through some microchemical accident, or chance quantum absorption, a sudden change in the composition ('pattern') of the gene takes place, known to biologists as a 'mutation,' then the gene of the new type, so produced, reproduces itself according to this new type, i.e. it now produces precisely the new pattern.

This shows that the copying property depends upon some more fundamental feature of gene structure [phosphate-ribose chain to the modern reader] than does the specific pattern which the gene has [base sequence to the modern reader], and that it is the effect of the former to cause a copying not only of itself but also of the latter, more variable, features. It is this fact which gives the possibility of biological evolution and which has allowed living matter ultimately to become so very much more highly organized than non-living. It is this which lies at the bottom of ... growth, reproduction, and heredity.

As we shall see (Chapter 7), the "possibility of biological evolution" occurs because, although mutations are often repaired, sometimes they are not. A change in "specific pattern" is then passed on, by copying, to the next generation. When considering pairs of bases, care should be taken to distinguish between: (i) a Watson-Crick base-pair (i.e. two bases on separate strands, or separate parts of a strand, which are involved in the classical **A-T** and **G-C** pairings), (ii) a dinucleotide consisting of two ordered contiguous bases on the same strand (e.g. **CG** often written as **CpG**; see Chapter 18), and (iii) the base composition of a nucleic acid segment (e.g. **(G+C)%**; see Chapter 10).

Turnover and Channeling

The "microchemical accident," to which Muller referred might have a definite cause (e.g. Muller himself had noted increased mutations following X-irradiation), or might loosely be described as 'spontaneous.' The accident might result in one regular letter being substituted for another (e.g. "hat" rather than "mat"), or a regular letter might change to something else (e.g. "pat" rather than "mat"), or simply be eliminated (e.g. "at" rather than "mat"). As will be discussed later (under the heading "entropy;" Chapter 15), it seems to be a general property of the universe that the

elements that compose it, whatever their size, tend to become disordered and evenly distributed. This is true at the chemical level where macromolecules tend to break down to their micromolecular building blocks, and the building blocks themselves, either separate or when they are part of macromolecules, live under a constant threat of structural change and dismemberment into their constituent atoms.

Photographers sometimes want to photograph a busy city scene but without the people, traffic and parked cars. The solution is to use time-lapse photography. A fixed camera takes a picture once a day with a very short exposure time. The film is not wound on, so daily pictures are superimposed. The first picture, if developed, would show nothing (because of short exposure). However, over weeks and months static objects begin to appear, whereas the transient objects are never present long enough to register. Since macromolecules tend to be transient, a magic time-lapse camera that could see individual molecules in bodies would tend to register nothing – except for molecules of DNA (and a few structural proteins like collagen). From this crude metaphor one should not deduce that DNA molecules are static. Even buildings vibrate and move within their confines. So do DNA molecules.

Two cell strategies for dealing with the constant breakdown of its parts are *recycling* (so that macromolecules are degraded and then resynthesized from their component parts), and *repair*. The former strategy (turnover) applies mainly to four of the five major classes of macromolecules (lipids, carbohydrates, proteins and RNA). The latter strategy applies mainly to the fifth class, DNA. Thus, whereas a damaged amino acid in a protein (a polymer of amino acid units) leads to the protein being degraded by specific enzymes (proteases) to its constituent amino acids, a damaged nucleotide in a DNA molecule (a polymer of nucleotide units) often invokes a ‘rapid response team’ of repair enzymes that will do its best to effect on-site repair without necessarily interrupting macromolecular continuity.

Synthesis of macromolecules from component parts is something cells do well because the assembly lines (biochemical pathways) for making the components are well established. DNA polymerase does not have to stand idle, waiting for a suitable nucleotide to turn up. However, the nucleotides that supply the needs of the cell for DNA synthesis can also supply the needs of a foreign invader – a virus. Ideally, nucleotides would be carefully dispensed to match the cell’s needs, but would be kept from the predators. Indeed, another cell strategy is to *channel* the nucleotides to the site of DNA synthesis [12]. Thus, the enzymes, both for the balanced synthesis of the four component parts, and for their incorporation into DNA, can exist as a large multi-enzyme aggregate close to the replication fork (Fig. 2.3).

Promiscuous DNA

Sometimes there is a break in a DNA duplex. The two ends may be reconnected by various enzymes (“ligases”). However, the tendency towards disorder sometimes means that a DNA segment is incorrectly reconnected. A random ‘cut’ followed by

a 'paste' may result in one segment of DNA recombining with a new segment of DNA so that the order of the information they contain is changed (transposed or inverted). To the extent that such changes are not critical for survival, genomes are vulnerable to an on-going kaleidoscopic diversification – a constant shuffling – of the sequences they contain.

More than this, DNA molecules are promiscuous – meaning, literally, that DNA molecules are “pro-mixing.” Place two duplex DNA molecules within a common cell wall and they will seek each other and attempt to recombine. We shall see that biological evolution became possible when DNA ‘learned,’ by adjusting sequence and structure, how to constrain and channel this tendency. Often the order of information in DNA is critical. Specific segments of DNA have specific ‘addresses’ in the chromosomes that contain them. The ability to accurately recombine specific segments of duplex DNA, while maintaining segment order and the integrity of functional units, is a fundamental property of living organisms. Indeed, US biologist George Williams, one of those responsible for our modern ‘selfish gene’ concept, thought it better to define genes in terms of their abilities to resist dismemberment by recombination, than in terms of their functions (see Chapter 11). The great evolutionary significance of recombination was pointed out by Crick in 1970 [13]:

There is also a major problem to which I believe biologists have given insufficient attention. All biologists essentially believe that evolution is driven by natural selection, but ... it has yet to be adequately established that the rate of evolution can be ... explained by the processes which are familiar to us. It would not surprise me if nature has evolved rather special and ingenious mechanisms so that evolution can proceed at an extremely rapid rate – recombination is an obvious example.

A year later Crick presented his “unpairing postulate” to explain how the inward-looking bases in a DNA double-helix might look *outward* to recognize complementary bases in another helix (see Chapter 10).

Haploidy and Diploidy

So breath-taking was Watson and Crick's model that some potentially major criticisms were overlooked. If every line of the present book were repeated, after the fashion of the cat sentences at the beginning of this chapter, then the book would be twice as long as it now is. Not only does it make sense to minimize the duplication of information in books, but there are also circumstances where it would appear advantageous not to duplicate information in biological systems. Despite this, *duplication is the rule*. For example, one of the two forms of gamete, usually the male spermatozoon, has to be highly mobile and hence has a streamlined shape and, tadpole-like, is often equipped with a flagellum. There appears to have been a selection pressure to keep the quantity of contained information (i.e. DNA) to a minimum (Fig. 1.3). Virus genomes, which have to be packaged for transfer from organism to organism, are also very compact. Yet, the DNA of spermatozoa and viruses is always in duplex form (with a few special exceptions).

Many organisms alternate during their life cycle between haploidy (one copy of each chromosome, containing one DNA duplex, per cell) and diploidy (two copies of each chromosome per cell). Gametes are haploid and so contain only one copy of each DNA duplex. When male and female gametes unite, the product (zygote) is diploid with two copies of each DNA duplex-containing chromosome, one from the father and one from the mother. Some organisms, such as the malaria parasite *Plasmodium falciparum*, quickly switch back to the haploid state, so its adult form is haploid. But for many organisms, diploidy is the adult norm. Only when new gametes are formed is there a brief flirtation with haploidy.

Thus, there is redundancy of information not only because DNA molecules come as duplexes, but also because many organisms ‘choose’ for most of their life cycles to have two copies of each duplex. Since each duplex has *at least* two-fold redundancy, diploid organisms have *at least* four-fold redundancy in their content of DNA. Why “*at least*”? There is only at least four-fold redundancy because we have so far considered only *in-parallel* redundancy. The phenomenon of *in-series* redundancy was discovered when measurements were made of the rates at which duplexes would reform from single strands when in solution in test tubes.

In-Series Redundancy

In the 1950s it became possible to synthesize artificial single-stranded RNA sequences such as UUUUUUUUUUUU referred to as poly(rU), and AAAAAAAAAA referred to as poly(rA). The single strands when mixed together (e.g. poly(rU) + poly(rA)) formed a double-stranded hybrid, which had a helical structure similar to that of double-stranded DNA. Omitting the helix, this can be represented as:



At the time it appeared amazing that hybridization could occur in a simple salt solution at room temperature in the absence of enzymes. The biologist Julian Huxley (grandson of Thomas Huxley) announced the discovery of “molecular sex” [14]. Whether said in jest, or from profound insight, the description fits perfectly (see Chapter 10).

What was going on in the privacy of the test-tube when millions of flexible, snake-like, poly(rU) molecules were mixed with millions of flexible, snake-like, poly(rA) molecules? Following the Watson-Crick base pairing rules, molecules of poly(rU) react only weakly with each other (since U pairs weakly with U). Furthermore, there is little inclination for the molecules to fold back on themselves, permitting internal pairing of U with U. The same applies for poly(rA). So there was nothing left but for As to pair with Us (analogous to A-T pairing in DNA). Since the molecules had little internal secondary structure (no folding back on themselves), it

was easy for a writhing chain of **Us** to find a writhing chain of **As**. Millions of relatively rigid, duplex molecules resulted. Their formation could be monitored by changes either in light absorption or in the viscosity of the solution.

Hybridization was studied similarly with natural nucleic acid duplexes that had been randomly fragmented into smaller duplexes (each about 1200 base pairs in length). From knowledge of the length of an original unfragmented DNA duplex and the number of such duplexes in a solution it was possible to calculate how rapidly the sheered duplex fragments should reform after their two strands had been separated from each other by heating. Like separated partners on a dance floor, to reform, each single-strand would have to find its complement. If there were just one original DNA duplex present, then each single strand would have no option but to find its *original* complementary partner. If two identical duplexes were present it would not matter if a strand found a partner from the other duplex (i.e. it would switch dancing partners). However, in this case there would be twice the chance of finding a partner in a given space and time, compared with when only one duplex was present. Thus, the more identical DNA duplexes present, the more rapidly should the strands reform duplexes (anneal) after heating.

When the experiment was carried out, it was found that for many DNA samples the rate of duplex reformation was far *greater* than anticipated [15]. This was particularly apparent in the case of species with very long DNA molecules. Further studies showed that within DNA there is a redundancy due to the presence of repetitive elements. There are many more copies of certain segments of DNA than the four expected from in-parallel considerations. Molecular ‘dancing partners’ may be found *in-series* as well as in-parallel. Why is there so much sequence redundancy? Could it all be beneficial (see Chapter 15 for discussion of “junk DNA”)? We note below, that ‘high flyers’ long ago found they could manage quite nicely, thank you, without some of their repetitive elements.

Bits and Bats

There is a link with information theory. Since there are two main types of bases, purines (**R**) and pyrimidines (**Y**), then, disregarding the phosphate-ribose medium upon which the base message is written, a nucleic acid can be represented as a binary string such as:

$$\mathbf{YRYRRYRYRRYRRYRYRYR} \quad (2.5)$$

Electronic computers work with information in this form – represented as strings of 0s and 1s. If a **Y** and an **R** are equally likely alternatives in a sequence position, then each can be measured as one “bit” (binary digit) of information, corresponding to a simple yes/no answer.

Confronted with a generic base (often expressed as **N**) you could first ask if it was a purine (**R**). A negative reply would allow you to infer that **N** was a pyrimidine (**Y**).

You could then ask if it was cytosine (C). A positive reply would allow you to infer that N was not thymine (T). By this criterion, each position in a DNA sequence corresponds to two potential yes/no answers, or two “bits” of information. So the entire single-strand information content in the human haploid genome (3×10^9 bases) is 750 megabytes (since 8 bits make a byte). This is of the order of the amount of information in an audio compact disk.

This way of evaluating DNA information has been explored [16], but so far has not been particularly illuminating with respect to DNA function. One reason for this may be that DNA is not just a binary string. In the natural duplex form of DNA, a base in one string pairs with its complementary base in another string. Each base is ‘worth’ 2 bits, so that a base pair would correspond to 4 bits. However, even if not paired the two bases would still collectively correspond to 4 bits. Thus, the chemical pairing of bases increases their collective information content to some value greater than 4 bits. But does this come at a price?

At another level (literally and otherwise) consider flying organisms – bats, birds, insects. In every case we find duplex DNA in cells. Every cell of all multicellular organisms has duplex DNA, and flying organisms are no exceptions. Bats have 5.4 picograms of DNA per cell [17], whereas equivalent mammals (mice) have 7 picograms per cell. Bats appear to have shed some of the ‘excess’ DNA, but that which remains is still in *duplex* form. Birds have approximately 2.5 picograms of DNA/cell. A bird that could shed half its DNA and exist with single stranded DNA would seem to have a weight advantage compare with a bird that had duplex DNA. It should be able to fly faster and further than those with duplex DNA, a feature of particular importance for migratory birds. But again, the DNA is always in duplex form.

Nevertheless, relative to humans, birds seem to have shed (or to have not acquired) ‘excess’ DNA. Where did this excess originate? Humans and chickens have similar numbers of genes, but the average chicken chromosome is more crowded – 1 gene/40 kilobases compared with 1 gene/83 kilobases [18]. This suggests that birds have shed the DNA between genes. But the average chicken gene is half the size of the average human gene – 27 kilobases compared with 57 kilobases. So birds have also shed some DNA that is deemed “genic.” Yet chicken proteins are the *same* size as human proteins. This suggests that birds have not shed the protein-encoding parts of genes (exons; see Chapter 13). Instead, they have shed non-protein-encoding parts of genes (introns).

In Figure 2.5 the lengths of exons and introns in some corresponding genes of chickens and humans are compared. In the case of exon lengths the slope is 1.0; so, on average, each chicken exon is the *same* size as the corresponding human exon. This is in keeping with bird proteins being the same size as human proteins. But in the case of intron lengths the slope is 0.4. The dashed lines indicate that, on average, a large 8 unit human intron would correspond to a large 6.5 unit chicken intron, and a small 4.5 unit human intron would correspond to a small 5.2 unit chicken intron. There is a big difference in the case of large introns, and a much smaller difference in the case of small introns (the scale is logarithmic). Large human introns are much bigger than the corresponding large chicken introns.

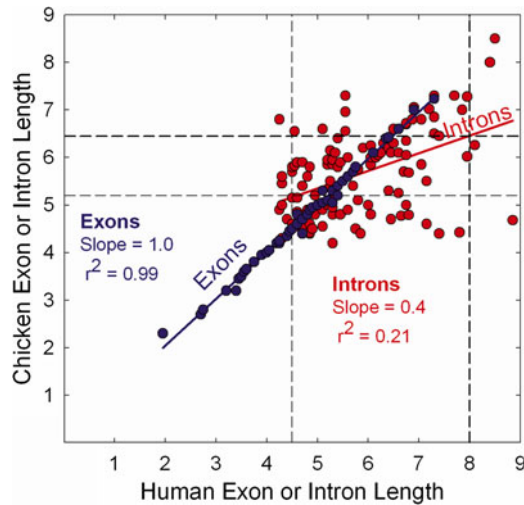


Fig. 2.5 Comparison of the lengths of exons and introns in corresponding genes of humans and chickens. The lines are drawn to best fit the points (see Appendix A). Both slopes are significantly different from zero. Note that, in general, exons (blue) are smaller than introns (red; see Chapter 13), and the points for exons fit more closely to their line (SEE = 0.11; $r^2 = 0.99$) than the points for introns fit to their line (SEE = 0.80; $r^2 = 0.21$). The lesser scattering of exon length values from their line, indicates that the exons in the set of genes studied have been under strong negative selection (i.e. individuals with exon mutations have tended not to survive). Thus, since the time when humans and chickens diverged from a common ancestor, exon sequences have been less successful at varying than intron sequences (i.e. individuals with intron mutations have survived more often than individuals with exon mutations; see Chapters 7 and 8). Dashed lines facilitate comparison between small and large introns (see text). This figure is adapted from reference [19]

When we examine the sequences we find that humans have an excess of repetitive elements (see Chapter 15). These lie both *between* genes (in intergenic DNA) and *within* genes (in introns).

It seems more likely that birds have *discarded* some repetitive elements, than that humans have acquired repetitive elements [19, 20]. When there is less DNA, the nucleus – the ‘brain’ of the cell – is smaller. So bird nuclei are smaller than human nuclei. And there are similar trade-offs at the level of the whole organism. We should not be surprised to find that migratory birds have smaller brains than non-migratory. Indeed, the further the distance migrated, the smaller the brain [21, 22]. Would the demands on cognition and memory be less in migratory birds? Or perhaps “bigger is not always better” where brains are concerned (see Part VII).

From all this it seems that there are compelling reasons for keeping DNA in duplex form at all stages of life. In biological systems there are conflicts and there have to be trade-offs. But abandoning the duplication of DNA information is seldom one of them.

Accidents and Non-Accidents

In our lives we encounter two classes of adverse events – random and non-random. The non-paranoid designate random adverse events as “accidents.” There is conflict between the forces of disorder and the forces of order, the former usually being not deliberately hostile, but merely reflecting the tendency of things, if left alone, to become untidy rather than tidy. This tendency gets greater when things move faster, which usually means they get hotter, as will be considered at the molecular level in Chapter 15.

In this chapter we have considered error-generation as driven by random processes (“microchemical accidents”), and sequence redundancy as having arisen to permit error-detection, and so, possibly, error-correction. Redundancy means that the *qualitative* characteristics (e.g. sequence) of an organism’s *own* DNA molecules can be compared, so allowing *quality control*. By mechanisms to be touched on in Chapter 6, the total quantity of DNA in a cell is maintained relatively constant. This is *quantity control*. Since the quantity of DNA determines the ‘dose’ of gene-products (e.g. proteins) that a cell contains (see Chapter 17), this implies that the quantity of cellular macromolecules can be regulated, directly or indirectly, by DNA quantity-control mechanisms.

Sometimes the forces of disorder have an appreciable non-random component, as when a virus (i.e. foreign DNA; V) deliberately enters a cell. The repertoire of ‘self’ macromolecules (M) then is supplemented by sets of ‘not-self’ macromolecules (VM). So the total quantity of cellular macromolecules (TM) can be written:

$$M + VM = TM \quad (2.6)$$

Under normal circumstances, the quantity of macromolecules of virus origin (VM) would be zero. As will be seen in Chapters 13, there are sophisticated host strategies for distinguishing ‘self’ from ‘not-self,’ and to be successful (i.e. to increase VM) a virus must outsmart them. The closer the virus, in its *qualitative* characteristics, can approach to self (i.e. become ‘near-self’ with respect to its host) the more likely it is to succeed. Host quality-control mechanisms, are then likely to be less effective. There is, however, the theoretical possibility of using *quantitative* characteristics of viruses (i.e. VM itself) as a basis for distinction by the host. The available strategies for organisms to respond *internally* to non-random adversities, in the forms of viruses, are somewhat similar to the available strategies for countries to respond *internally* to non-random adversities, in the forms of forgers of their currencies. The metaphor may be helpful.

The aim of a forger is to fool you with counterfeit currency. If successful the forger prospers, but if too successful there is the possibility that the entire monetary system would collapse. This would not serve the forger well and, so far as we know, no forger has gone to this extreme. Nevertheless, the counterfeit notes must be as like the real thing as the forger can contrive. At the qualitative level, your visual and tactile sensory mechanisms for distinguishing real notes from counterfeit notes

must be evaded. Accordingly, manufacturers of a country's true currency are engaged in an 'arms race' with the illegitimate manufacturers of false currency. As forgers get progressively better at counterfeiting currency that approaches progressively closer to the real thing, so the manufacturers of true currency must add embellishments that are difficult for forgers to imitate. This allows you to continue to make *qualitative* distinctions on a note-by-note basis.

At the level of the entire currency system, however, it should *in theory* be possible to detect that forged notes (FN) are present without looking at individual notes. Designating the quantity of real notes as N, we can write:

$$N + FN = TN \quad (2.7)$$

TN represents the total quantity of notes. Here is how it would work. Given knowledge of the initial quantity of real notes, and their rates of manufacture and of destruction when worn-out, then it should be possible to know how many real notes (N) exist. If there were a way of directly monitoring how many notes actually existed at a particular time-point (e.g. knowing the average 'concentration' of notes and the area over which they were distributed), then the actual number (TN) could be compared with the calculated number (N). If the actual number exceeded the calculated number, then the presence of forged notes (FN) would be inferred, alarm bells would ring, and appropriate corrective measures implemented. In principle, if the system were sufficiently sensitive, a small initial increase in forged notes would be immediately responded to. A forger would have difficulty opposing this form of monitoring. But, of course, in practice such monitoring is difficult for countries to implement.

Biological organisms are not so constrained. In general, 'self' molecules are manufactured at rates that have been fine-tuned over millions of years of evolution. Similarly, rates of destruction have been fine-tuned. Accordingly, the concentrations of many molecules, *especially proteins*, fluctuate between *relatively narrow limits*. Intrusive foreign 'not-self' macromolecules would tend to increase total macromolecule concentrations in ways that, in principle, should be detectable. This theme will be explored in Parts V and VI.

Summary

Most cell components undergo cycles of degradation and resynthesis ("turnover"), yet their concentrations fluctuate between only very narrow limits. The DNA of a cell provides information that specifies both the quality and quantity of these components. Accurate transmission of this information requires that errors in DNA be detected and corrected. If there is more than one copy of the information (redundancy) then one copy can be compared with another. For hereditary transmission of information, a 'message' is 'written' as a sequence of four base 'letters' – **A, C, G, T** – on a strand of phosphate and ribose (the 'medium'). In duplex DNA there is

two-fold redundancy – the ‘top’ strand is the complement of the ‘bottom’ strand. **A** on one strand matches **T** on the other, and **G** on one strand matches **C** on the other. A check for non-complementarity permits error-detection. Thus, Chargaff’s first parity rule is that, for samples of duplex DNA, the quantity of **A** (adenine) equals the quantity of **T** (thymine), and the quantity of **G** (guanine) equals the quantity of **C** (cytosine). In diploid organisms there is four-fold *in parallel* sequence redundancy, due to the presence of a DNA duplex (chromosome) of maternal origin, and a DNA duplex (chromosome) of paternal origin. There is also some *in-series*, within-strand, redundancy. Trade-offs to optimize utilization of sequence space do not include abandonment of duplex DNA or diploidy. Birds lighten their DNA load by decreasing the number of repetitive sequences that would be located either between or within genes. DNA is promiscuous in readily acquiescing to a ‘cutting-and-pasting’ (recombination between and within strands) that shuffles the information it contains. Indeed, George Williams thought it better to define genes in terms of their abilities to resist dismemberment by recombination, than in terms of their functions. Furthermore, a codiscoverer of DNA structure, Francis Crick, questioned the potency of the natural selection of functional differences as an evolutionary force, and pointed to possible “ingenious mechanisms” involving recombination that might accelerate evolutionary processes.

References

1. Darwin C (1871) *The Descent of Man, and Selection in Relation to Sex*. Appleton, New York, pp 57–59
2. Watson JD, Crick FHC (1953) Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature* 171:737–738
3. Chargaff E (1951) Structure and function of nucleic acids as cell constituents. *Federation Proceedings* 10:654–659
4. Wyatt GR (1952) Specificity in the composition of nucleic acids. *Experimental Cell Research*, Supplement 2:201–217
5. Wyatt GR, Cohen SS (1953) The bases of the nucleic acids of some bacterial and animal viruses. *Biochemical Journal* 55:774–782
6. Israelachvili J, Wennerstrom H (1996) Role of hydration and water structure in biological and colloidal interactions. *Nature* 379:219–224
7. Muller HJ (1941) Resumé and perspectives of the symposium on genes and chromosomes. *Cold Spring Harbor Laboratory Symposium on Quantitative Biology* 9: 290–308
8. Sinden RR (1994) *DNA Structure and Function*. Academic Press, San Diego
9. Watson JD, Crick FHC (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature* 171:964–967
10. Kornberg A (1989) *For the Love of Enzymes: the Odyssey of a Biochemist*. Harvard University Press, Cambridge, MA
11. Muller HJ (1936) The needs of physics in the attack on the fundamental problems of genetics. *Scientific Monthly* 44:210–214
12. Forsdyke DR, Scott FW (1979) Exogenous purine deoxyribonucleosides do not prevent inhibition of DNA synthesis by hydroxyurea: evidence for nonconvergence of *de novo* and salvage pathways. In: Lynen F, Mothes K, Nover L (eds) *Cell Compartmentation and Metabolic Channelling*. Elsevier, Amsterdam, pp 177–184 [Viruses can sometimes escape dependence

- on host *de novo* biosynthetic pathways by encoding their own enzymes. In turn, when alarmed, some hosts can activate enzymes that degrade the free nucleotides needed by viral polymerases; Yan N, Lieberman J (2012) SAMHD1 does it again, now in resting T cells. *Nature Medicine* 18:1611–1612]
13. Crick F (1970) Molecular biology in the year 2000. *Nature* 228:613–615
 14. Forsdyke DR (2007) Molecular sex: the importance of base composition rather than homology when nucleic acids hybridize. *Journal of Theoretical Biology* 249:325–330
 15. Waring M, Britten RJ (1966) Nucleotide sequence repetition: a rapidly reassociating fraction of mouse DNA. *Science* 154:791–794
 16. Gatlin LL (1972) *Information Theory and Living Systems*. Columbia University Press, New York [In ‘the classical period of molecular biology’ the often misleading input of cyberneticists and mathematicians tended to discourage a fuller exploration of informational aspects of biology.]
 17. Burton DW, Bickham JW, Genoways HH (1989) Flow-cytometric analyses of nuclear DNA contents in four families of neotropical bats. *Evolution* 43:756–765
 18. Bellott DW et al. (2010) Convergent evolution of chicken Z and human X chromosomes by expansion and gene acquisition. *Nature* 466:612–616
 19. Hughes AL, Hughes MK (1995) Small genomes for better flyers. *Nature* 377:391 [In 2013 the Animal Genome Size Database maintained by Ryan Gregory had 475 avian records; the mean genome size was 1.39 gigabases (Gb) and 90% of the genomes fell between 1.0 and 1.6 Gb. The corresponding values for 657 mammalian species were 3.36 Gb with 90% between 2.2 and 4.9 Gb.]
 20. Organ CL, Shedlock AM, Meade A, Pagel M, Edwards SV (2007) Origin of avian genome size and structure in non-avian dinosaurs. *Nature* 446: 180–184 [Mammalian red blood cells (Fig. 16.1) have no nuclei. It would seem advantageous, regarding body weight, that birds should be similar. Yet, avian red blood cells are nucleated.]
 21. Safi K, Seid MA, Dechmann DKN (2005) Bigger is not always better: when brains get smaller. *Biology Letters* 1:283–286
 22. Sol D, Garcia N, Iwaniuk A, Davis K, Meade A, Boyle WA, Szekely T (2010) Evolutionary divergence in brain size between migratory and resident birds. *PLOS One* 5:e9617

Chapter 3

Information Levels and Barriers

All messages and parts of messages are like phrases or segments of equations which a mathematician puts in brackets. Outside the brackets there may always be a qualifier or multiplier which will alter the whole tenor of the phrase.

Gregory Bateson (1960) [1]

Aristotle wrote that the form-giving essence, the *eidos*, “does not become part of the embryo, just as no part of the carpenter enters the wood he works... but the *form is imparted* by him to the material” [2, 3]. A carpenter may hold the information for imparting form to wood in his head. If he wishes to communicate this information to other carpenters separated from himself in time and space, he has to choose symbols and a code. The symbols, D, O, and W, if repeated and arranged in the sequence W, O, O, and D, would be meaningless unless the receiver knew the transmitting carpenter’s code. Thus, implicit to the information concept is the idea that a grouping of symbols can have a meaning, and that there can be a linkage (‘mapping’) between the grouping and the meaning, which we call a code (an agreed convention).

Symbols and Code

If symbols are arranged as a string in one dimension, then the corresponding code is also likely to be one-dimensional. The code can be three-dimensional, as in the way a lock, by its complementary shape, ‘decodes’ the contours (i.e. the symbols) of a key. Indeed, it is in this three-dimensional manner that biomolecules communicate with other biomolecules within and between cells, to create the phenomenon we know as life. If the time of the intermolecular communication were also encoded, then the code would be four-dimensional, with time being the extra dimension. This would mean that the decoding process might also involve the molecular resonances or vibrations referred to in Chapter 2.

Primary and Secondary Information

Returning to our human written form of information, the sentence “Mary had a little lamb its fleece was white as snow” contains the *information* that a person called Mary was in possession of an immature sheep. The sentence is a *message*, whereas the information – the meaning of the message – is itself an abstract entity that is conveyed by the message. The message can be in the form of an English text as above, or a French text, or a German text, or in some (at the time of this writing) unknown form in our brains (see Part VII). The determination that the information was the same in these sources would be made when it was extracted from each message form by applying a decoding process and shown to be convertible to a common message form (e.g. the English text).

The information about Mary and the lamb can be conveyed in different languages. So the message itself can contain not only its *primary* information, but also *secondary* information about its source – e.g. it is likely that an author is more familiar with one language than others. Some believe that English is on the way to displacing other languages, so that eventually it (or the form it evolves to) could constitute the only language used by humans on this planet. Similarly, in the course of early evolution it is likely that a prototypic nucleic acid language, perhaps initially in a two-letter form (**R** and **Y**), displaced contenders.

Because languages diverged from primitive root languages many thousands of years ago, it would be difficult to discern directly a relationship between the English, French and German versions of “Mary had a little lamb its fleece was white as snow.” However, in England, a person with a Cockney accent might say “Miree ader liawl laimb sfloyce wors woyt ers snaa.” Cockney English and ‘regular’ English diverged more recently (they are “allied languages”) and it is easy to discern similarities. Take the following:

yewas htbts llem ws arifea ac Mitte alidsnoe lahw
irsnwwis aee ar lal larfoMyce b sos woilmyt erdea

One line of text is the regular English version with the letters shuffled. The other line is the Cockney version with the letters shuffled. Can you tell which is which? If the shuffling was thorough, the primary information has been destroyed. Yet, there is still *some* information left. With the knowledge that Cockneys tend to ‘drop’ their Hs, it can be deduced that the upper text is more likely to be from someone who spoke regular English. With a longer text, this could be more precisely measured.

Thus, languages have characteristic letter frequencies. You can take a segment (‘window’) and count the various letters in that segment. In this way you can identify a text as English, Cockney, French or German. We can call this information “secondary information.” There may be various other *levels* of information in a sequence of symbols. To evaluate the secondary information in DNA (with only four base ‘letters’), you can select a window (say 1000 bases) and count each base in that window. You can apply the same window to another section of the DNA, or to another DNA molecule from a different biological species, and repeat the count. Then you can compare DNA ‘accents.’

Primary Information for Protein

The best understood type of primary information in DNA is the information for proteins. The DNA sequence of bases (one type of ‘letter’) encodes another type of ‘letter,’ the “amino acids.” There are 20 members of the amino acid alphabet, with names such as aspartate, isoleucine, serine and valine. These are abbreviated as **Asp**, **Ile**, **Ser** and **Val** (and sometimes as single letters; e.g. D, I, S, V). Under instructions received from DNA, the decoding and assembling machinery of cells joins amino acids together in the same order as they are encoded in DNA, to form proteins. The latter, amino acid chains that fold in complicated ways, play a major role in determining how we interact with our environment. Proteins constitute a major part of, for example, the enzymes, which interact with other molecules (substrates) to catalyze chemical changes. The proteins largely determine our phenotype (Figs. 1.1, 1.4).

In an organism of a particular species (“A”), consider the twenty one base DNA sequence:

$$\text{TACGACGCCGATAGCGTCGTA} \quad (3.1)$$

This, read in sets of three bases (“codons”), can convey primary information for a seven amino acid protein fragment (**TyrAspAlaAspSerValVal**). All members of the species will tend to have the same DNA sequence, and differences between members of the species will be rare and of minor degree. If the protein is fundamental to cell function (like insulin or growth hormone) it is probable that organisms of *another* species (“B”) will have DNA that encodes the *same* protein fragment. However, when we examine their DNA we might find major differences compared with the DNA of the first species (the similarities are underlined):

$$\text{TATGATGCTGACAGTGTGTT} \quad (3.2)$$

This sequence *also* encodes the above protein fragment, showing that the DNA contains the *same* primary information as DNA sequence 3.1 above, but it is ‘spoken’ with a different ‘accent.’ Such *secondary* information could have a biological role. It is theoretically possible (but unlikely) that all the genes in an organism of species *B* would have the ‘accent,’ yet otherwise encode the *same* proteins as species *A*. In this case, organisms of species *A* and *B* would be both anatomically and functionally (physiologically) *identical*, while differing dramatically with respect to secondary information. Biologists may refer to species that show this tendency as “sibling species” (*espèces jumelles*, *Geschwisterarten*).

On the other hand, consider a single change in the sequence of species *A* to:

$$\text{TACGACGCCGTTAGCGTCGTA} \quad (3.3)$$

Here the difference (underlined) would change one of the seven amino acids. It is likely that such *minor* changes in a *very small* number of genes affecting development

would be sufficient to cause anatomical differentiation *within* species *A* (e.g. compare a bulldog and a poodle, as “varieties” of dogs). Yet, in this case the secondary information would be hardly changed.

A view to be explored in this book is that, like the Cockney’s dropped H’s, the role of secondary information is to *initiate* and, for a while, *maintain*, the reproductive isolation (recombinational isolation) that is essential for the process by which biological species are formed. Secondary information is information that *preserves species*. Indeed, such is the promiscuity of DNA that, even *within* the genome of a species, secondary information seems to play a similar role in maintaining the recombinational isolation of individual genes. Secondary information *preserves genes* (see Chapter 11).

Secondary information can exist because the genetic code is a ‘redundant’ or ‘degenerate’ code (see Chapter 7); for example, the amino acid serine is not encoded by just one codon – there are six possible codons (**TCT**, **TCC**, **TCA**, **TCG**, **AGT**, **AGC**). In species *A* (DNA sequence 3.1) serine (**Ser**) is encoded by **AGC**, whereas **AGT** is used in species *B* (DNA sequence 3.2). On the other hand, the change from **GAT** (sequence 3.1) to **GTT** (sequence 3.3) changes the encoded amino acid from aspartic acid (**Asp**) to valine (**Val**). This might suffice to change the properties of the corresponding protein, and hence to change one or more of the anatomical or physiological characters of an organism (its phenotype), which depend on that protein.

As a consequence of the degeneracy of the genetic code, given a nucleic acid sequence one can deduce the corresponding protein, but not the converse. Given a protein sequence one can hazard a guess at, but cannot accurately deduce, the nucleic acid that encodes it. Thus, in this context, information is doomed to flow one way from nucleic acid to protein. In the early days of molecular biology, Crick referred to this as the “central dogma.”

Information Barriers

A language has both content and accent. You cannot have intelligible content without an accent, but you can have an accent without intelligible content. Remove intelligible content and all you have left is the accent. For example, you could ask a Frenchman, an Englishman and an American to read a gobbledygook sentence (zero intelligible content) and most listeners would be able to tell which accent was which. To this extent, accent is likely to be more fundamental than content.

That it might be appropriate to compare the evolution of languages and of species has long been recognized [4]. A biological interest of linguistic barriers is that they also tend to be reproductive barriers. Even if a French person and an English person are living in the same territory (biologists call this living “sympatrically”), if they do not speak the same language they are unlikely to marry. Of course, being of the same species, if there were not this early barrier to reproduction, then the union would probably be fruitful in terms of the number and health of children. Thus, if this early barrier were overcome, there would probably be no later barriers.

However, the French tend to marry French and produce more French. The English tend to marry English and produce more English. Even in England, because of the class barriers so colorfully portrayed by G. B. Shaw in *Pygmalion*, Cockneys tend to marry Cockneys, and the essence of the barrier from people speaking ‘regular’ English is the difference in accent [5]. Because of other (‘blending’) factors at work in our society, it is unlikely that this linguistic ‘speciation’ will continue to the extent that Cockney will eventually become an independent language.

The point is that when there is incipient linguistic ‘speciation,’ it may be the *secondary* information (dropped H’s), rather than the primary information, which constitutes the barrier. Similarly, secondary information in DNA may provide a species barrier that keeps organisms reproductively separate as distinct quantal ‘packages.’ Organisms are reproductively isolated (i.e. only able to reproduce sexually with members of their own species). For purposes of reproduction, members of a species assort with their own type. An organism that breeds with a member of an allied species may sometimes produce healthy offspring (hybrids), but these offspring are usually sterile (“hybrid sterility”).

Secondary information in DNA can be considered as part of the “genome phenotype” (see Chapter 7) which, in the context of reproduction, we may refer to as the “reprotype.” In contrast, primary information in DNA is concerned with conventional phenotypic attributes – often dependent on proteins – such as the color of a flower, or the speed at which a horse can run. Such attributes (characters) are the grist for the mill of classical Darwinian natural selection.

Barriers Preserve Discontinuity

If you fully understand a concept then you should be able to deduce its implications. While it can be argued that the Victorians understood the information concept (Chapter 1), with the exception of Butler they did not explore what was implicit in the language they spoke and in the texts they read and wrote. For example, information can be inaccurate; it may be accurate at one moment at one place, but at a moment later, at the same or a different place, it may be inaccurate. A concept of information that does not include the concept that information is subject to error is an incomplete concept of information.

Galton in 1876 compared gametes to “mailbags” containing “heaps of letters,” and regretted that, with current technology biologists were like persons looking through the windows of a post office; they “might draw various valuable conclusions as to the postal communications generally, but they cannot read a single word of what the letters contain.” Noting that the “germs” from which bodily units were “derived” should have independently varied in each parent, and hence might have become “deficient” in *different* ways, Galton saw an advantage of sexual reproduction [6]:

When there are two parents, and therefore a double supply of material, the chance deficiency in the contribution of either of them, of any particular species of germ [gene], tends to be supplied by the other. No doubt, cases must still occur, though much more rarely than

before, in which the same species of germ is absent from the contribution of both, and a very small proportion of families will thereby perish.

Butler said this more colorfully in 1877 when pondering what sex was for [7]:

We should expect to find a predominance of sexual over asexual generation, in the arrangements of nature for continuing her various species, inasmuch as two heads are better than one, and a *locus poenitentiae* is thus given to the embryo – an opportunity of correcting the experience of one parent by that of the other. And this is what the more intelligent embryos may be supposed to do; for there would seem little reason to doubt that there are clever embryos and stupid embryos, with better or worse memories, as the case may be, of how they dealt with their protoplasm before, and better or worse able to see how they can do better now.

The possibility of error did not escape Miescher (Fig. 3.1) to whom we will be returning in Chapter 17. In 1892 he suggested that sex might have arisen to correct structural defects in molecules [8]:

To me the key to the problem of sexual reproduction is to be found in the field of stereochemistry. The ‘gemmules’ of Darwin’s pangenesis are no other than the numerous asymmetrical carbon atoms of organic substances. As a result of minute causes and external conditions these carbon atoms suffer positional changes and thus gradually produce structural defects. Sexuality is an arrangement for the correction of these unavoidable stereometric architectural defects in the structure of organized substances. Left handed coils are corrected by right-handed coils, and the equilibrium restored.

As noted in Chapter 2 when considering Muller’s “microchemical accidents,” error detection usually requires some form of redundancy, so that one copy can be compared with another. To correct a play by Shakespeare you need another copy of the same play. You cannot correct a copy of *Macbeth* with a copy of *The Tempest*. A blended form, – a “*Macpest*” or a “*Tembeth*” – also would not suffice. Thus, like

Fig. 3.1 Johann Friedrich Miescher (1844-1895)



biological species, the plays must be discrete, or “discontinuous.” They must come as packets, but not clonal packets. If you do not believe something in your newspaper you do not go out and buy another copy of the same newspaper which, having been produced as part of the same print run, contains identical information with identical errors.

The property of discreteness implies that there are limits. Shakespeare was able to explore the theme of personal ambition in *Macbeth*. A separate play was required to explore the theme of shipwreck on an enchanted island. The diversity of Shakespeare’s plays can be compared with the diversity of biological species. Within each species there is room for change, but within limits. A species can attain a certain type and level of complexity, but for further change, barriers must be breached and new barriers established. There must be an origin of species. Butler related this to information [7]:

Many plants and animals do appear to have reached a phase of being from which they are hard to move – that is to say, they will die sooner than be at the pains of altering their habits – true martyrs to their convictions. Such races refuse to see changes in their surroundings as long as they can, but when compelled to recognize them, they throw up the game because they cannot and will not, or will not and cannot, invent. This is perfectly intelligible, for a race is nothing but a long-lived individual, and like any individual, or tribe of men whom we have yet observed, will have its special capacities and its special limitations, though, as in the case of the individual, so also with the race, it is exceedingly hard to say what those limitations are, and why, having been able to go so far, it should go no further. Every man and every race is capable of education up to a point, but not to the extent of being made from a sow’s ear into a silk purse.

Sometimes one adaptation will prevent the full development of another. Adaptation for flight invokes counter-adaptations in body weight (Chapter 2). Since better cognition seems to require an increase in brain weight, we should be thankful that flight imposed a limit on brain size. Smart birds, as portrayed by Daphne du Maurier in *The Birds*, have yet to destroy mankind [9]. Each adaptation invokes counter-adaptations, thus achieving an organism best balanced to pursue a particular evolutionary strategy (e.g. flight).

In Scotland in 1867 a professor of engineering, Fleeming Jenkin, saw members of a biological species as enclosed within a sphere (Fig. 3.2). For a particular character – say potential running speed – most members would be at, or near, the center of the sphere, since most members are capable of average speed. Jenkin noted that, under the then prevailing view that parental characters were blended in offspring, there would always be a tendency for individuals to vary towards, rather than away from, the center [10]:

A given animal ... appears to be contained ... within a sphere of variation; one individual lies near one portion of the surface; another individual, of the same species, near another part of the surface; the average animal at the centre. Any individual may produce descendents varying in any direction, but is more likely to produce descendents varying towards the centre of the sphere, and the variations in that direction will be greater in amount than the variations towards the surface. Thus a set of racers of equal merit indiscriminately breeding will produce more colts and foals of inferior than of superior speed, and the falling off of the degenerate will be greater than the improvement of the select.

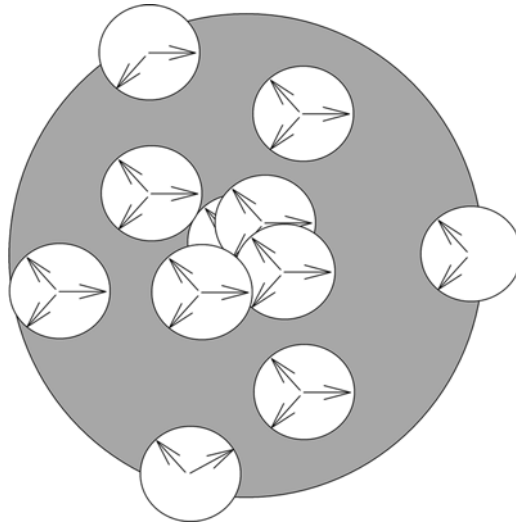


Fig. 3.2 Jenkin's sphere of species variation (gray). Members of a species and the ranges of variation seen in their children are represented as white circles, with arrows representing the directions of variations. Most individuals are clustered at the center (average phenotypes) where they are free to vary in any direction. Those at the precise center can only vary away from the center. The few individuals at the periphery (rare phenotypes with exceptionally positive or negative qualities), tend to vary towards the center. Thus their children cannot escape the confines of the species (i.e. their exceptional qualities cannot develop further). Note that the variations Jenkin refers to reflect differences in nature, not nurture. He is not referring to variations acquired through interaction with the environment within an individual lifetime (e.g. increased muscular strength in an individual that exercises)

This “regression to the mean,” given much weight by Galton, was invariably observed when individuals of *mixed pedigree* were crossed. He investigated the observation that tall people are often the offspring of tall parents and small people are often the offspring of small parents. Thus, in theory, average child height (when adult) should equal average parent height (Fig. 3.3). In fact, the children of tall parents, although still tall, are usually smaller than the parents. Likewise, the children of small parents, although small, are usually taller than the parents. Thus, to maintain the range of variant forms (e.g. to maintain racers of superior speed), it appeared that on-going variation (the nature of which was then unknown) would have to occur in order to replace what seemed to have been lost through regression.

In addition to studying height in humans, Galton also compared the size of plant seeds with those of parental plants. Johannsen later followed up on this approach to show that if an individual of *unmixed pedigree* (“pure line”) were crossed with one of the same pedigree, then a position in Jenkin's sphere could be sustained. There was no regression to the population mean, but there was still no escape beyond the surface (see Chapter 8).

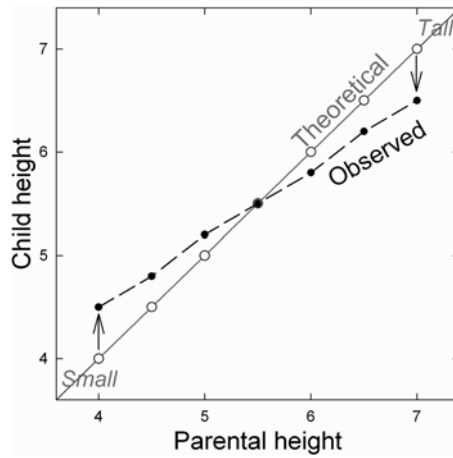


Fig. 3.3 Galton’s “regression to the mean.” The gray line indicates the theoretical expectation when the adult height of an individual closely follows that of its parents. However, on average, the children of tall parents, though still tall, are smaller than their parents, and the children of small parents, though still small, are taller than their parents. Thus, the children tend to “regress” towards the mean height in the parent population (in this case 5.5 units)

To Select is Not to Preserve

Johannsen in 1909 suggested the word “gene” as a *Rechnungseinheiten*, or accounting unit, that symbolized one of Bateson’s “character units.” In most definitions of the “gene” there is a loose or explicit reference to function. Biologists talk of a gene encoding information for tallness in peas. Biochemists talk of a gene encoding information for growth hormone (a protein), and relate this to a segment of DNA. One popular biochemical text states that “each region of the DNA helix that produces a functional RNA molecule constitutes a gene” [11]. Another states that “a gene is defined biochemically as that segment of DNA (or in a few cases RNA) that encodes the information required to produce a functional biological product” [12]. However, before it can function, information must be preserved. Classical Darwinian theory proposes that function, through the agency of natural selection, is itself the preserving agent (see Chapter 8). Thus, function and preservation (conservation) go hand-in-hand, but function is held to be more fundamental than preservation.

There is nothing strange in this. Shakespeare’s plays entertained far better than plays by other authors, and by virtue of this function they are with us today. But the physical preservation of the plays, the maintenance of their physical discontinuity from their surroundings, required much more – durable ink and paper, and the bindings and covers that make a folio or book. This was not all. The folios and books had to be carefully housed. The selection and reproduction of ancient texts by generations of priests and scribes came to nothing when the library of Alexandria was

burned to the ground. It is only by references from other sources that we now know the magnitude of our loss. The epic cycle poems of Homer's contemporaries ("The Little Iliad," "The Destruction of Troy"), some of the plays of Sophocles, and the works of ancient historians such as Hecataeus, Hellanicus, and Ephorus, are preserved only in fragments by later authors.

Bringing the metaphor closer to home, Mendel's ideas lay fallow for 35 years until their discovery in 1900. Initially, the ideas were in Mendel's head and in a certain issue of the obscure journal *Verhandlungen des naturforschenden Vereines in Brunn*, copies of which were distributed to many academic centers. Some copies were lost. Mendel died in 1884. But some of the copies were bound into journals. For 35 years Mendel's ideas were held between pages 3 to 47 in volume four of the journal, not because their value was recognized, but because the volumes were securely bound and stably housed. The preservation of the journal, although a *pre-condition* for the positive selection of its contents, did not guarantee that this selection would eventually occur (see Epilogue).

Eventually Mendel's ideas gained a life of their own. The texts that contained them were sought out, reproduced, and translated into many languages, and the ideas ('memes') began to spread from head to head. An interesting side-product was that the *Verhandlungen des naturforschenden Vereines in Brunn*, and in particular the issue containing Mendel's paper, gained increased attention. Consequently, other papers in that issue may have got a second reading, but not by virtue of their intrinsic merits. They, like the journal, were able to 'hitch-hike' on the success of Mendel's paper. Geneticists refer to this as a 'selective sweep' (see Chapter 7).

So, in biological systems which is more important, continuing demand for function, or actual physical preservation? This book argues that preservation is often more fundamental than function. Furthermore, secondary information (context) is an agency, distinct from natural selection, which brings this about. Once biological information has been preserved then its meaning can be displayed and its accuracy checked.

Another Order of Variation

William Bateson was concerned with the discreteness of biological species as distinct entities and the relationship of this discreteness to that of the environments they occupied. Thus fish species live in water and non-fish species live on land. The two species groups are discrete, as are their respective environments. So the discontinuity of the species matches the discontinuity of their environments. But in the case of allied species – likely to have diverged relatively recently from a common ancestral species – often differences in environments do not match the observed differences in species. Allied species are often found in the same territory (i.e. they are sympatric). Conversely, members of a single species sometime thrive in a wide range of environments. In 1894 the 33-year-old Bateson noted in his book *Materials for the Study of Variation Treated with Especial Regard to Discontinuities in the Origin of Species* [13]:

The differences between Species on the whole are Specific, and are differences of kind, forming a discontinuous Series, while the diversities of environment to which they are subject are on the whole differences of degree, and form a continuous Series; it is therefore hard to see how the environmental differences can thus be in any sense the directing cause of Specific differences, which by the Theory of Natural Selection they should be. [Bateson's capitalizations]

The preservation of a species did not seem readily explicable on the basis of conventional natural selection, often a relatively continuous environmental force capable of bringing about profound changes, but usually *within the confines of a species*. For the discontinuity of species, Bateson believed there should be a discontinuity in the way in which members of a species vary from one another.

His use of the term “discontinuity” can perhaps be best understood if we first consider the term “continuity.” If a road is continuous then *one process* – walking, running or riding – should suffice to reach a destination. If a road is discontinuous, perhaps because there is an unbridged river or ravine, then *another process* – wading, swimming or climbing – must intervene. If variation is continuous then, in principle, every possible intermediate between two forms can now exist, and one process that has removed intermediates that do not match their environment suffices to explain the distinctness of the diverse types we see around us. In short, natural selection reigns. If variation is discontinuous then each of the two forms is limited in the extent to which it can now vary. Every possible intermediate between the two forms cannot now exist (although intermediates may have existed in the past when the forms had just begun to diverge from a common ancestor). So, in addition to natural selection, another process that has something to do with the gap between forms, is needed. However, at that time Bateson could describe this only in abstract terms:

Upon the received hypothesis it is supposed that Variation is continuous and that the Discontinuity of Species results from the operation of Selection. ... There is an almost fatal objection in the way of this belief, and it cannot be supposed both that all variation is continuous and also that the Discontinuity of Species is the result of Selection. With evidence of the Discontinuity of Variation this difficulty would be removed. [Bateson's capitalizations]

The need for some other order of variation had been perceived earlier by Jenkin [10]:

The theory advanced [by Darwin] appears ... to be that, if owing to some other qualities a race is maintained for a very long time different from the average or original race (near the surface of our sphere), then it will spontaneously lose the tendency to relapse, and acquire a tendency to vary outside the sphere. What is to produce this change? Time simply, apparently. ... Not only do we require for Darwin's theory that time shall first permanently fix the variety near the outside of the assumed sphere of variation, we require that it shall give the power of varying beyond that sphere.

In his 1869 text *Heredity Genius*, Galton likened this power of varying beyond the sphere to the effort needed to roll a rough stone from one position of stability to another [14]. Initially, Bateson examined the observable characters of a wide range of plants and animals to catalogue the range of variations that could occur. With the emergence of Mendel's laws on quantitative aspects of variation, which demonstrated the

discontinuity of genes (i.e. the discreteness of the characters they encoded; see Chapter 8), it was tempting to think that this *genic* discontinuity might be the evidence he sought. However, although the discontinuity of genes clarified many issues, Bateson soon recognized that a *different type of discontinuity* was needed (see Chapter 9).

From the discontinuity of genes it followed that if a *rare* new mutant organism crossed with one of the *abundant* non-mutant forms in its species (the most likely occurrence), then the emergent mutant character would not immediately be diluted (seen as blending of parental characters). The character's increase or decrease in the population would be statistically determined, depending on (i) its degree of dominance (see Chapter 8), (ii) whether it was seen by natural selection as beneficial or detrimental, and (iii) whether at some point, perhaps after preservation for many generations, there was a cross with another organism with the same mutation. This latter point was emphasized by Jenkin [10]:

Any favourable deviation must ... give its fortunate possessor a better chance of life; but this conclusion differs widely from the supposed consequence that a *whole* species may or will gradually acquire some one new quality, or *wholly* change in *one* direction and in the *same* manner.

Easy access to a mate with the same mutation had been a strength of the Lamarckist doctrine of the inheritance of acquired characters. According to Lamarck, organisms subjected to the same environmental provocation would *simultaneously* adapt by mutation. Being colocated in the same environment there would then be little difficulty in finding a mate with the same adaptation. Mendel's discovery of genic discontinuity raised the possibility of preservation for many generations until a potential mate with the same mutation was found – there was no need for immediate colocalization.

Context as Barrier

Some mutations bring about dramatic, sometimes large scale, changes in a single generation – for example, dwarfism, extra fingers, and a leg where a fly's antenna should be (see Chapter 7). Although initially misled into believing that these “monstrosities” or “sports” might be examples of the discontinuity he sought, William Bateson came to realize that the process of variation had to generate a *non-genic* discontinuity as a barrier that would preserve species (see Chapter 9). Yet he did not see this discontinuity as a barrier that might relate to error detection and correction. His thoughts on evolution greatly influenced his son Gregory (named after Mendel; Fig. 3.4), who became an anthropologist/psychologist much interested in the hierarchy of information levels in human communication.

In *Steps to an Ecology of Mind*, Gregory Bateson pointed to the all-important role of context [15]. He noted that schizophrenic patients often appear unable to recognize the context in which primary information is conveyed. More simply stated, there is no smile or wink acting as a “classifying metamessage.” This work of the younger Bateson and his colleagues is well summarized by his biographer [16]:

Fig. 3.4 Gregory Bateson
(1904-1980)



Assuming the view that in communication there are primary messages which are verbal and classifying messages which express the on-going relationship of the actors, they proposed that the discourse of schizophrenics especially distorts or omits the signals about relationship. ... The ability to discriminate the various contextual cues which specify both the semantic and personal meaning of a message seemed to be impaired. Human communication was construed in terms of multiple levels of reference, ... The problem ... posed for the schizophrenic was how does a simple message *mean* if it is not pinned down by normal metamessages.

Gregory Bateson extrapolated this to biological information transfer [15]:

Both grammar and biological structure are products of communicational and organizational processes. The anatomy of the plant is a complex transform of genotypic instructions, and the 'language' of the genes, like any other language, must of necessity have contextual structure. Moreover, in all communication there must be a relevance between the contextual structure of the message and some structuring of the recipient. The tissues of the plant could not 'read' the genotypic instructions carried in the chromosomes of every cell unless cell and tissue exist, at that given moment, in a contextual structure.

Gregory Bateson also equated a change in contextual structure to a special type of change in the genomic information required for species divergence that his father had postulated [17]:

If evolution proceeded in accordance with conventional theory, its processes would be blocked. The finite nature of somatic changes indicates that no ongoing process of evolution can result only from successive externally adaptive genotypic changes, since these must, in combination, become lethal, demanding combinations of internal somatic adjustments of which the soma is incapable. We turn therefore to a consideration of *other classes of genotypic change*. What is required to give a balanced theory of evolution is the occurrence of genotypic changes, which shall *increase* [Bateson's italics] the available range of somatic flexibility. When the internal organization of the organisms of a species has been limited by environmental or mutational pressure to some narrow subset of the total range of living states, further evolutionary progress will require some sort of genotypic change which will compensate for this limitation.

In Germany, Richard Goldschmidt (Fig. 3.5) had termed this “macroevolution.” In Chapter 7 we will consider Goldschmidt’s proposal that, within the confines of a species, the plasticity of the phenotype is limited since changes in genotypic ‘pattern’ (microevolution) are limited. A chromosomal ‘repatterning’ (macroevolution) is required if species are to diverge into new species, so expanding the scope of possible phenotypes. The relationships of some of the dichotomies considered here to Goldschmidt’s terminology are indicated in Table 3.1.

Fig. 3.5 Richard Goldschmidt (1878-1958)



Table 3.1 Some of the dichotomies considered in this chapter. “Physiological selection” is a term introduced by Romanes to contrast with “Natural selection.” When a particular *genome phenotype* (“reprotype”) confers reproductive success or failure, then physiological selection has operated. Similarly, when a particular *conventional phenotype* confers reproductive success or failure, then natural selection has operated (see Part III)

	Dichotomy	
	Primary	Secondary
Information Type	Genic	Non-Genic
DNA involved	Natural	Physiological
Selection Type	Conventional	Genome
Phenotype	Function	Preservation
Adaptation for:	Continuous	Discontinuous
Scope of Variation	Non-branching	Branching
Evolution Type	Microevolution	Macroevolution
Goldschmidt Terminology		

Not a Palimpsest

Today's metaphors may be tomorrow's truths. As we saw in Chapter 1, Romanes could tolerate Butler's relating heredity and memory *as a metaphor*, but the thought that within cells there might be something akin to a written text was going too far. Metaphors can help us understand our genomes. But they have to be the right metaphors. Some metaphors are just not helpful. Others can actively mislead.

The idea of levels of information invites comparison with palimpsests (Greek: *palimpsestos* = scraped again). In ancient times messages on clay tablets could be scraped off to allow reuse. The scrapings may have been incomplete, but sufficient for reuse. Likewise, the memory discs of today's computers can have their contents erased so allowing fresh messages to be entered. Again, the erasures may be incomplete. Computers can 'crash' and the messages on their discs can be accidentally erased. There are processes by which erased messages can sometimes be restored.

In 1906 in Constantinople a faded tenth century copy of Archimedes' long lost *Method of Mechanical Theorems* was discovered on a parchment beneath the lettering of a twelfth century Greek prayer book [18]. Such a document is also referred to as a palimpsest. With modern technologies some palimpsests are found to have multiple layers of text. Clearly, the writers of the top layers had no use for the preceding layers. For their purposes they were garbage and the only lament, if any, would be that they had been insufficiently erased. Translating this mode of thought to the layers of information that have been discovered in genomes, it was easy to assume that, while these may be of interest as possible remnants of ancient messages laid down over billions of years of evolution, they are not of relevance to genomes as they now exist. Thus, the term "junk DNA" did not seem inappropriate.

This view is challenged in Chapter 15. Because we are unable to conjure up a functional explanation for something, it does not follow that it has no function. Following the admonition attributed to William of Occam, we should accept a simple explanation if it fits the facts as well as, or better than, a more complex explanation. But we should also heed the admonition attributed to Albert Einstein: "Make everything as simple as possible, but not simpler." Certainly, from modern DNA sequences we can infer some of the characteristics of ancient sequences, but it does not therefore follow that modern sequences are not strictly functional, serving the needs of modern organisms. Nature can only adapt that which already exists, but when conflicts arise she sternly dismisses that which no longer serves her purposes.

A popular form of 'molecular archaeology' is to compare the sequences of two species that are considered to have derived from a common ancestral species (phylogenetic analysis; see Chapter 7). What they have *in common* may have been in the ancestral sequence (i.e. has been conserved). What they do not have in common may have been acquired since their divergence into separate species. For example, modern passenger automobiles and airplanes both serve the practical and aesthetic needs of modern humans. They have certain features in common, such as wheels, windows and at least one forward-looking seat at the front. It is not unreasonable to infer that many 'ancient' horse-drawn carriages had these features.

Hybrid Disciplines and Hybrid Sterility

Barriers protect and preserve. People join clubs and societies to associate with groups of like-minded types. This is also true for academic disciplines. To become a member of a club, society or discipline, various entry criteria must be satisfied (i.e. barriers must be breached). Here again, if not used with care, metaphors can mislead. When emphasizing the importance of chemistry for understanding genetic function, Crick noted [19]:

Classical genetics is ... a black-box subject. The important thing was to combine it with biochemistry. In nature hybrid species are usually sterile, but in science the reverse is often true. Hybrid subjects are often astonishingly fertile, whereas if a scientific discipline remains too pure it usually wilts. In studying a complicated system, it is true that one cannot even see what the problems are unless one studies the higher levels of the system, but the *proof* of any theory about higher levels usually needs detailed data from lower levels. [Crick's italics]

Hybrid subjects are indeed astonishingly fertile, but not quite in the way Crick implied. Horses and donkeys (asses) have been artificially crossed for millennia because the mule that results, although sterile, is endowed with great energy and vigor. Thus, hybrid animals and hybrid disciplines have *in common* the feature that they tend to be *more vigorous* than the corresponding parental animals or disciplines. They are both highly productive, but not highly reproductive. This is because they also have *in common*, a tendency towards sterility, and for much the same reasons.

Within the barrier afforded by an academic discipline, members are likely to thrive, provided their activities remain within the confines of the group. Here they are likely to be best understood, so their ideas become subject to constructive criticism and error-correction. However, those who stray beyond the confines of one discipline (in Mendel's case, Biology) and engage with another discipline (in Mendel's case, Physics), may be misunderstood by those in both parental disciplines. Barrier breachers may be vigorously productive, but the communication gap may imperil both the novel ideas they wish to convey and their personal fortunes. They run the risk of an alienation that can disrupt their lives. At least in the short term, they are academically sterile (i.e. their ideas are not reproduced; see Epilogue).

Again, Butler, unique among the Victorians in his presentation of evolution in informational terms, had some sense of this [7]. First he considered the case of a cross between two types so disparate that the embryo would not develop correctly ("hybrid inviability"):

We should expect to find that all species, whether of plants or animals, are occasionally benefited by a cross; but we should also expect that a cross should have a tendency to introduce a disturbing element, if it be too wide, inasmuch as the offspring would be pulled hither and thither by two conflicting memories or advices, much as though a number of people speaking at once were without previous warning to advise an unhappy performer to vary his ordinary performance – one set of people telling him he has always hitherto done thus, and the other saying no less loudly that he did it thus; – and he were suddenly to become convinced that they each spoke the truth. In such a case he will ... completely break down, if the advice be too conflicting, ... through his inability to fuse the experiences into a harmonious whole.

Thus, the ‘advice’ of the paternally-contributed gamete and the ‘advice’ of the maternally-contributed gamete would be so much in conflict that the puzzled embryo would be unable to undertake its ‘ordinary performance.’

Next Butler considered, in turn, first the reproductive events leading to a parental male horse and to a parental female donkey, and then their cross to produce a hybrid embryo that was not so puzzled as to be unable to undertake its ‘ordinary performance’ of development. The result would be a mule, vigorously healthy (“hybrid vigor”), but sterile (“hybrid sterility”):

[In] the case of hybrids which are born well developed and healthy, but nevertheless perfectly sterile, it is less obvious why, having succeeded in understanding the conflicting memories of their parents, they should fail to produce offspring. ... The impregnate ovum from which the mule’s father was developed remembered nothing but its horse memories; but it felt its faith in these supported by the recollection of *a vast number* of previous generations, in which it was, to all intents and purposes, what it now is.

In like manner, the impregnate ovum from which the mule’s mother was developed would be backed by the assurance that it had done what it is going to do now a hundred thousand times already. All would thus be plain sailing. A horse and a donkey would result.

These two [parents] are brought together; an impregnate ovum is produced which finds an unusual conflict of memory between the two lines of its ancestors, nevertheless, being accustomed to *some* conflict, it manages to get over the difficulty, *as on either side it finds itself backed by a very long series of sufficiently steady memory*. A mule results – a creature so distinctly different from either horse or donkey, that reproduction is baffled, owing to the creature’s having nothing but its own knowledge of itself to fall back upon.” [Butler’s italics]

Sadly, Butler himself turned out to be an academic mule. With a few late-in-the-day exceptions (e.g. William Bateson and Francis Darwin), his ideas were scorned by the scientific establishment (see Epilogue) [20].

Summary

Information is an abstract entity that we recognize in the form of a message conveyed in a medium. Messages are read (decoded) to extract meanings (information content) and are compared with copies of the same messages to detect and correct errors. These functions require that a message contain both its primary information and a barrier (secondary information) that preserves the context within which the primary information is read for meaning and for errors. There must be discontinuity between different messages. Just as the themes that can be developed within the context of a single play (e.g. *Macbeth*) are limited, so that new themes require new plays (e.g. *The Tempest*), evolutionary themes that can be developed within the context of a biological species are limited, so that further evolutionary progress requires a barrier-breaching change that establishes a new species. Just as error-correction of the text of *Macbeth* requires another text of *Macbeth* from an independent source, so error-correction of a genome text is best carried out with a genome

text from an unrelated member of the same species. Sex is primarily a within-species error-correcting device requiring that species be discontinuous and reproductively isolated from each other. Members of allied species that fail to recognize their context and breach the reproductive barrier may produce offspring, but these are sterile (hybrid sterility). Similarly, humans who fail to recognize the context of cognitive discourse may be schizophrenic.

References

1. Bateson G (1967) Cybernetic explanations. *American Behavioural Scientist* 10:29–32
2. Romanes GJ (1891) Aristotle as a naturalist. *Contemporary Review* 59:275–289
3. Delbrück M (1971) Aristotle-totle-totle. In: Monod J, Borek F (eds) *Of Microbes and Life*. Columbia University Press, New York, pp 50–55
4. Lyell C (1863) *The Geological Evidences of the Antiquity of Man with Remarks on Theories of the Origin of Species by Variation*. Childs, Philadelphia, p 467
5. Shaw GB (1913) Pygmalion. In: *Bernard Shaw. Complete Plays with Prefaces*. Volume I. Dodd, Mead, New York (1963)
6. Galton F (1876) A theory of heredity. *Journal of the Anthropological Institute* 5:329–348
7. Butler S (1878) *Life and Habit*. Trübner, London, pp 172–180, 200–201 [*Locus poenitentiae* is Latin for a “place of repentance.” Legally the term signifies a, perhaps brief, temporal safe-haven where there is an opportunity to change one’s mind. Thus, between the last bid and the fall of the auctioneer’s hammer, there is a *locus poenitentiae* for a decision, which cannot be stretched out for ever. Technically, a cross between a male horse and a female donkey results in a “hinny,” whereas a cross between a female horse and a male donkey results in a “mule.” Butler did not bother with such niceties, claiming that “I neither know, nor want to know, more detail than is necessary to enable me to give a fairly broad and comprehensive view of my subject.”]
8. Olby R, Posner E (1967) An early reference on genetic coding. *Nature* 215:556
9. Du Maurier D (1952) *The Birds*. Penguin Books, London.
10. Jenkin F (1867) The origin of species. *The North British Review* 46:277–318
11. Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson JD (1994) *Molecular Biology of the Cell*. 3rd Edition. Garland Publishing, New York, p 340
12. Lehninger AL, Nelson DL, Cox MM (1993) *Principles of Biochemistry*. 2nd Edition. Worth Publishers, New York, p 789
13. Bateson W (1894) *Materials for the Study of Variation Treated with Especial Regard to Discontinuities in the Origin of Species*. Macmillan, London, pp 16, 69 [Bateson and his St. John’s College colleague, the physicist Joseph Larmor, were friends. Some of the ideas associated with the new physics may have influenced Bateson’s thinking.]
14. Galton F (1869) *Hereditary Genius*. Macmillan, London
15. Bateson G (1973) *Steps to an Ecology of Mind*. Paladin, St. Albans, p 127
16. Lipset D (1980) *Gregory Bateson: The Legacy of a Scientist*. Prentice-Hall, Englewood Cliffs, p 208
17. Bateson G (1963) The role of somatic change in evolution. *Evolution* 17:529–539
18. Wilson NG (1999) Archimedes: the palimpsest and the tradition. *Byzantinische Zeitschrift* 92:89–101
19. Crick F (1988) *What Mad Pursuit. A Personal View of Scientific Discovery*. Basic Books, New York, p 150
20. Forsdyke DR (2006) Heredity as transmission of information. Butlerian intelligent design. *Centaureus* 48: 133–148

Part II

Parity and Non-parity

Chapter 4

Chargaff's Second Parity Rule

Poetry's unnat'ral; no man ever talked poetry 'cept a beadle on boxin'day, or Warren's blackin' or Rowland's oil, or some o' them low fellows.

Mr. Weller. Pickwick Papers [1]

Information has many forms. If you turn down the corner of a page to remind you where you stopped reading, then you have left a message on the page ('bookmark'). In future you 'decode' (extract information from) the bookmark with the knowledge that it means "continue here." A future historian might be interested in where you paused in your reading. Going through the book, he/she would notice creases suggesting that a flap had been turned down. Making assumptions about the code you were using, a feasible map of the book could then be made with your pause sites. It might be discovered that you paused at particular sites – say at the ends of chapters. In this case pauses would be correlated with the distribution of the book's 'primary information.' Or perhaps there was a random element to your pausing – perhaps when your partner wanted the light out. In this case pausing would be influenced by your pairing relationship. While humans not infrequently get in a flap about their pairing relationships, we shall see that flaps are useful metaphors when we observe DNA molecules in a flap about theirs.

Information Conflict

The primary information in this book is in the form of the linear message you are now decoding. While not exactly bed-time reading, we can assume that there will be pauses. If you turn down a large flap it might cover up part of the message. Thus, one form of information can interfere with the transmission of another form of information. To read the text you have to correct (fold back) the 'secondary structure' of the page (the flap) so that it no longer overlaps the text. Thus, there is a conflict. You can either retain the flap and not read the text, or remove the flap and read the text.

In the case of a book page, the text is imposed on a flat two-dimensional base – the paper. The message (text) and the medium (paper) are different. Similarly, in the case of our genetic material, DNA, the ‘medium’ is a chain of two units (phosphate and ribose), and the ‘message’ is provided by a sequence of ‘letters’ (bases) attached at intervals along the chain (Fig. 2.4). It will be shown in Chapters 5 and 13 that, as in the case of a written text on paper, ‘flaps’ in DNA (secondary structure) can conflict with the base sequence (primary structure). Thus, the pressures to convey information encoded in a particular sequence, *and* to convey information encoded in a ‘flap,’ may be in conflict. If it is possible to recognize that one form of information (e.g. flap) is in some way higher than another form (e.g. text), then hierarchical levels can be assigned, and it can then be said that there is a conflict not only between different *forms* of information, but also between different *levels* of information. In biological systems where there is competition for genome space, the ‘hand of evolution’ has to resolve these *intrinsic* conflicts while dealing with other pressures (*extrinsic*) from the environment.

Prose, Poetry and Palindromes

Primary information can be conveyed in Pidgin English. “Cat sat mat” can convey essentially the same primary information as the more formal “the cat sat on the mat.” It is known that the act of sitting is something cats frequently do, and that mats usually occupy lowly places where they are vulnerable to being sat upon. An important function of the syntax of formal prose – the form the primary information would take in normal human discourse – is that it can suggest the possibility that there has been an error in the primary information. The sentence “The cat sat on the mat” is syntactically correct, and on this basis we might judge that the author had really written what he/she had meant to write. However, this comes at a price. “Cat sat mat” takes less time to write or speak, and occupies less space. Definite articles and prepositions can be considered as redundant information. Provided we know that it is cats that do the sitting, the less formal prose of Pidgin English has certain efficiencies.

To those whose first language is English this may not be obvious. Like Monsieur Jourdain in Moliere’s *Le Bourgeois Gentilhomme*, we discover that we have been speaking and writing a high level formal prose quite effortlessly all our lives [2]. But, in fact, it has not been all our lives. As infants we began by stringing single words together in forms not too different from “cat sat mat,” and had to acquire, slowly and laboriously, the modifications that led to a higher level of discourse.

An even higher level of discourse would be poetry. The error-detecting function would then be considerably increased. In this case the ear and eye would, from childhood, have been so accustomed to meter and rhyme that, if these elements were missing, the listener or viewer would be alerted to the possibility of error in the primary information. Yet, as Mr. Weller notes, we do not naturally speak poetry. The advantage of this extra layer of error-detecting ability does not appear to have outweighed its disadvantages.

The same can be said of the writings of Christian Bök [3]. You should be able to spot the potential error-detecting device:

Relentless, the rebel peddles these theses, even when vexed peers deem the new precepts mere dreck. The plebes resent newer verse; nevertheless, the rebel perseveres, never deterred, never dejected, heedless, even when hecklers heckle the vehement speeches.

Bök constrains himself to use only the vowel “e”. Georges Perec also tried this in 1972 [4], and in an earlier novel he employed all the vowels with the exception of “e” [5]. That an error had occurred in Bök’s text, and in Perec’s 1972 text, would become detectable if a vowel other than “e” appeared. That an error had occurred in Perec’s earlier text would become detectable if “e” appeared.

The constraints become greater when discourse is palindromic (Greek: *palindromos* = running back again) as when, to prevent any misunderstanding, there was a coy introduction in the Garden of Eden:

Madam I’m Adam.

The sentence reads the same forwards and backwards. If we count letters we get: A = 4, D = 2, I = 1, M = 4. The palindrome has a lone central “I.” Otherwise, all the letters correspond to even numbers. Palindromes obey a parity rule.

With sufficient ingenuity you can come up with quite long palindromes. Indeed, they have been known since ancient times. Inscribed in Greek on medieval fountains one may find the admonition “Wash sins, not only your face:”

NIYON ANOMHMATA MH MONAN OYIN

Here, A = 4, H = 2, I = 2, M = 4, N = 2, O = 4, Ψ = 2, and T = 1. The following, by geneticist Susumo Ohno, has no lone central letter:

Doom! Sad named was dog DNA, devil’s deeds lived, and God saw demand as mood.

In this case the parity rule is absolute (A = 8, D = 10, E = 6, G = 2, I = 2, L = 2, M = 4, N = 4, O = 6, S = 4, V = 2). Thus, a check for parity, which disregarded lone central letters, should be capable of detecting errors quite powerfully, but at a very high cost. Normal human discourse would be severely constrained if our communications were so limited.

Yet we shall see that this form of human language serves as a good metaphor for the language of our genomes. In Ohno’s words: “All DNA base sequences, regardless of their origins or functions (coding versus non-coding) are messages written in palindromic verses” [6]. This indicates that the error-detecting role of the genome language (involving various forms of information that can be referred to as secondary) may be of *more* importance than the immediate efficiencies of communicating primary genetic messages (primary information).

Intrastand Parity

Chargaff’s second parity rule (PR2) is that his first parity rule (PR1), which applies to duplex double-stranded DNA, also applies, to a close approximation, to single-stranded DNA. If the individual strands of a DNA duplex are isolated and their base

compositions determined, then $A\% \cong T\%$, and $G\% \cong C\%$ [7]. This also means that the 'top' and 'bottom' strands of duplex DNA have approximately the same base composition. For example, if we arrange for base parity in a single top strand:



Then, following PR1, the complementary bottom strand must have the same base composition:



The validity of PR2 became clearer when complete genomic sequences became available. For example, the cowpox virus (*Vaccinia virus*; Latin, *vacca* = a cow) that is used to vaccinate humans against smallpox, has in its top strand, 63,921 As, 63,776 Ts, 32,030 Gs, and 32,010 Cs, for a total of 191,737 bases. Cut the strand into two equal halves and the rule still holds – $A\% \cong T\%$ and $G\% \cong C\%$. Cut each of these halves into halves, and again the rule holds. Indeed, it holds, albeit less precisely, even when the divisions continue down to segments of a few hundred bases [8].

To simplify, the relationship can be expressed as $(A+G)\% = (C+T)\%$, or $R\% = Y\%$. Since the proportions of purines (**R**) and pyrimidines (**Y**) are reciprocally related (i.e. if a base is not an **R** it must be a **Y**, and vice-versa), then the degree of parity can be expressed as $R\%$ or $(A+G)\%$, with a value of 50% representing precise parity (i.e. $R\% = Y\%$). This will be considered more in Chapters 6 and 12.

It is easy to dismiss PR2 as a product of random forces. For example, in many species there are approximately equal numbers of males and females. This sexual parity appears to follow from the fact that in one of the sexes there are usually two types of gamete (e.g. in human males there are X-chromosome bearing spermatozoa and Y-chromosome bearing spermatozoa; see Chapter 17). These have approximately equal opportunities of encountering a gamete of the opposite sex (i.e. the female ovum bearing a solitary X-chromosome; see Fig. 1.3). Thus, on a chance basis, there should be equal numbers of males (XY) and females (XX). However, chance can be supplemented by powerful evolutionary forces [9, 10].

Imagine that, say, after a war, there were more females than males. Then there would be more potential partners for the rare sex (male). Each male would, on average, come to leave more offspring than each female. Biologists call this "frequency-dependent selection." Now, whereas most families have a genetic tendency to produce equal numbers of males and females, some families have a genetic tendency to produce more males and some have a genetic tendency to produce more females (for an example, see Chapter 17). When males are the rare sex, families with a genetic tendency to produce more males would leave more offspring. Hence the proportion of males in the population would increase. When the male/female ratio was restored, then families tending to produce more males would no longer have this advantage, and their genetic characteristics, *including the tendency to produce more males*, would no longer gain a greater representation in future populations.

Table 4.1 Given that any base (**N**) can accompany any of the four bases (**A, C, G, T**) to form a dinucleotide (i.e. **NA, NC, NG, NT**), then there are 16 possible dinucleotides (a), which include 6 pairs of complementary dinucleotides (b). The four dinucleotides **TA, GC, CG** and **AT** (in grey boxes) have themselves as complement (e.g. the reverse complement of **TA** is **TA**, and so **TAs** on one strand have **TAs** opposite them on the other strand). Thus, there are 10 complementary dinucleotide pairs, of which four are ‘self-complementary’

a					b		
		A	C	G		A	T
A		AA	AC	AG		AA	TT
						AC	GT
C		CA	CC	CG		AG	CT
						CA	TG
G		GA	GC	GG		CC	GG
						GA	TC
T		TA	TC	TG			

A similar converse argument would apply if there were a population fluctuation causing excess females. We shall see that, as in the case of sexual parity, PR2 is the result of randomness supplemented by powerful evolutionary forces.

If PR2 holds for the 2 pairs of complementary bases (**A** and **T, G** and **C**), what about the 10 pairs of complementary base duplets, or the 32 pairs of complementary base triplets, or the 128 pairs of complementary base tetruplets, or ... ? But we are getting a little ahead of ourselves. Let’s first sort out what we mean by complementary base duplets (dinucleotides), triplets (trinucleotides), tetruplets (tetranucleotides), quintruplets (pentanucleotides), etc. (Table 4.1).

Polarity and Complementarity

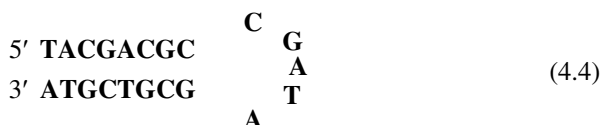
Sentences in English show polarity (i.e. they have a beginning and an end), and are read sequentially from left to right. Following this, by convention we write the ‘top’ strand of a DNA duplex from left to right. The strand has polarity beginning on the left at what, for chemical reasons, is referred to as the 5’ (“five prime”) end, and terminating at the right at what, for chemical reasons, is referred to as the 3’ (“three prime”) end (see Fig. 2.4). The top strand is the strand that is recorded in databases,

such as the GenBank database maintained at the National Center for Biotechnological Information (NCBI) in Washington, DC. It should be noted that although only one strand is recorded in Washington there is global redundancy, since the same sequence is stored at equivalent sites in Europe and Japan.

Watson and Crick found that the chemical constraints on their double helix model were best met if the bottom strand was in opposite orientation to the top strand. So a duplex is like a two-lane highway. The duplex sequence shown in Chapter 2 can be labeled to show “antiparallel” strand polarities:



Under certain circumstances, this double helix can form ‘flaps.’ Indeed, the sequences of the two strands have been pre-arranged so that they each can form stem-loop secondary structures due to pairing with complementary bases in the *same* strand. For example, if we unpeel the top strand away from the duplex, it can fold into a form with two major elements, a stem and a loop:



For this ‘flap’ in DNA to occur there have to be matching (complementary) bases. The stem consists of paired bases. Only the bases in the loop (**CGATA**) are unpaired. In the loop there happen to be equal numbers of Cs andGs, but there is an extra **A**. It can be said that, in this case Chargaff’s parity rule applies, to a close approximation, to a *single strand* of DNA (i.e. **A**=5, **C**=6, **G**=6 and **T**=4).

PR2 usually applies to the single strands of duplex DNA from biological sources. It turns out that, just as I arranged with my word-processor for the above sequence to have regions where the base sequences complement, so throughout evolutionary time the ‘hand of Nature’ has arranged for biological sequences to have similar base arrangements. Take another look:



Here the trinucleotide **GTA** in the top strand (underlined) has as its complement **TAC** in the bottom strand (underlined; Table 4.2). Both are, by convention written, in the sentence preceding this one, with 5’ to 3’ polarity. The strands of duplex DNA being antiparallel (Fig. 2.4), whenever there may be any doubt, the complement is referred to as the “*reverse complement*.” Note that **CAT** in the bottom strand (underlined), written *in this particular instance* (i.e. in the sentence you are now reading) with 3’ to 5’ polarity, would be the “*forward complement*” of **GTA** in the top strand, if it were also in the top strand and were hence written with 5’ to 3’ polarity.

Table 4.2 Any of the 16 dinucleotides can join any of the four bases to form a trinucleotide, so there are $16 \times 4 = 64$ trinucleotides ($k = 3$). Each can pair with one of the others (e.g. **GTA** and **TAC**, in boxes) and there are 32 pairs of reverse complementary trinucleotides. Similar tables can show the $64 \times 4 = 256$ tetranucleotides ($k = 4$), and the $256 \times 4 = 1024$ pentanucleotides ($k = 5$), etc.. When k is even (see Table 4.1) there are self-complementary tuples (k -mers). When k is odd there are no self-complementary tuples (k -mers)

	A	C	G	T
AA	AAA	AAC	AAG	AAT
AC	ACA	ACC	ACG	ACT
AG	AGA	AGC	AGG	AGT
AT	ATA	ATC	ATG	ATT
GA	GAA	GAC	GAG	GAT
GC	GCA	GCC	GCG	GCT
GG	GGA	GGC	GGG	GGT
GT	GTA	GTC	GTG	GTG
CA	CAA	CAC	CAG	CAT
CC	CCA	CCC	CCG	CCT
CG	CGA	CGC	CGG	CGT
CT	CTA	CTC	CTG	CTT
TA	TAA	TAC	TAG	TAT
TC	TCA	TCC	TCG	TCT
TG	TGA	TGC	TGG	TGT
TT	TTA	TTC	TTG	TTT

Considering the top strand alone, **TAC** at the beginning of the sequence is the *reverse* complement of **GTA** at the end of the sequence, meaning that its three bases could potentially base pair with the complementary three bases in **GTA** to form the stem in a stem-loop secondary structure (see sequence 4.4 above). Given the tendency of DNA sequences to form palindromes (see Chapter 5), single-stranded DNAs from most biological sources tend to have equifrequencies of inverse complements (e.g. **GTA** and **TAC**), but equifrequencies of forward complements (e.g. **GTA** and **CAT**) are much less apparent [11].

This was described as a “symmetry principle” by Indian physicist Vinayakumar Prabhu in 1993 [12]. Dinucleotides, trinucleotides, tetranucleotides, etc., were referred to as “2-tuples,” “3-tuples,” “4-tuples,” etc., and the frequency of each was scored in overlapping segments (e.g. the 3-tuple **GTA** contains two 2-tuples – **GT** and **TA**):

A study of all sequences longer than 50000 nucleotides currently in GenBank reveals a simple symmetry principle. The number of occurrences of each n -tuple of nucleotides in a given strand approaches that of its complementary n -tuple in the same strand. This symmetry is true for all long sequences at small n (e.g. $n = 1, 2, 3, 4, 5$). It extends to sets of n -tuples of higher order n with increase in length of the sequence.

The demonstration of symmetry depended on the length of the sequence under study. That this should be so is evident when one thinks about our own written language. Members of the word-pairs “rat”–“tar,” and “dog”–“god,” can, for present purposes, be considered as complementary (i.e. if there were sense-antisense pairing, one would match the other, letter by letter). If one took a text of only a few paragraphs one might, on a chance basis, come across a disproportionate number of instances of one of these (say 95% rat and 5% tar). In a longer text one would arrive at a more proportionate number (say 60% rat and 40% tar), which might hold for similar texts of similar lengths. For “dog” and “god” it is likely that the proportionate number in the same texts would differ (say 75% dog and 25% god). In a veterinary text the proportions might be 90% dog and 10% god, and the reverse in a religious text.

The amazing thing about the DNA language is that the proportions are often close to 50% and 50% (i.e. PR2), and this holds in *all* ‘texts’ examined and for *all* pairs of complementary ‘words.’ Tuple pairs of higher order than “rat” and “tar,” are not easy to find in our language – e.g. “lager” and “regal.” One would have to search extremely long texts for these words in order to arrive at a proportionate number that might hold for other texts of similar length. Similarly, one has to search very long DNA texts to approximate 50%:50% symmetry for high order n -tuples. But if a long enough sequence is available, this is found.

Duplets and Triplets

Prabhu's n -tuples are usually referred to as ‘ k -mers.’ He showed that, if the frequency of each k -mer in a given sequence is plotted against the frequency of its complementary k -mer in the same sequence, then the data points fall close to the diagonal. Figure 4.1a shows this for single bases (1-mers) in Vaccinia virus (the values were given earlier in this chapter). The four bases form two complementary base pairs, so there are only two points on the graph. A line through the two points extends back to zero. More points could be produced for the graph by taking base composition values from other organisms. However, since different organisms, or segments from different organisms, often have different absolute numbers of bases, it is better to present values as percentages of the total number of bases (Fig. 4.1b).

Even better, the values can be presented as percentages of the frequency of the most abundant base. This is shown for the frequencies of six complementary dinucleotide pairs of Vaccinia virus in Figure 4.1c. Again, a line through the data points extends (extrapolates) back to zero. Thus, complementary dinucleotides (2-mers) are present in approximately equal quantities. This also applies to complementary trinucleotides (3-mers; Fig. 4.2a).

Fig. 4.1 Equiprequencies of complementary bases (a, b), and dinucleotides (c) in the ‘top’ strand of the *Vaccinia virus* genome. Since there are only 2 pairs of complementary bases, there are only two points in (a) and (b). Complementary dinucleotide pairs (see Table 4.1b) generate six points in (c). Location on diagonals signifies equipfrequency. In (c) dinucleotide frequencies are expressed as a percentage of the frequency of the most abundant dinucleotide (AA). Note that in (a) and (b) pyrimidines are assigned to the X-axis, and purines are assigned to the Y-axis. This is not possible with dinucleotides since some dinucleotides contain both a pyrimidine and a purine. In this case assignment to one or other axis is arbitrary

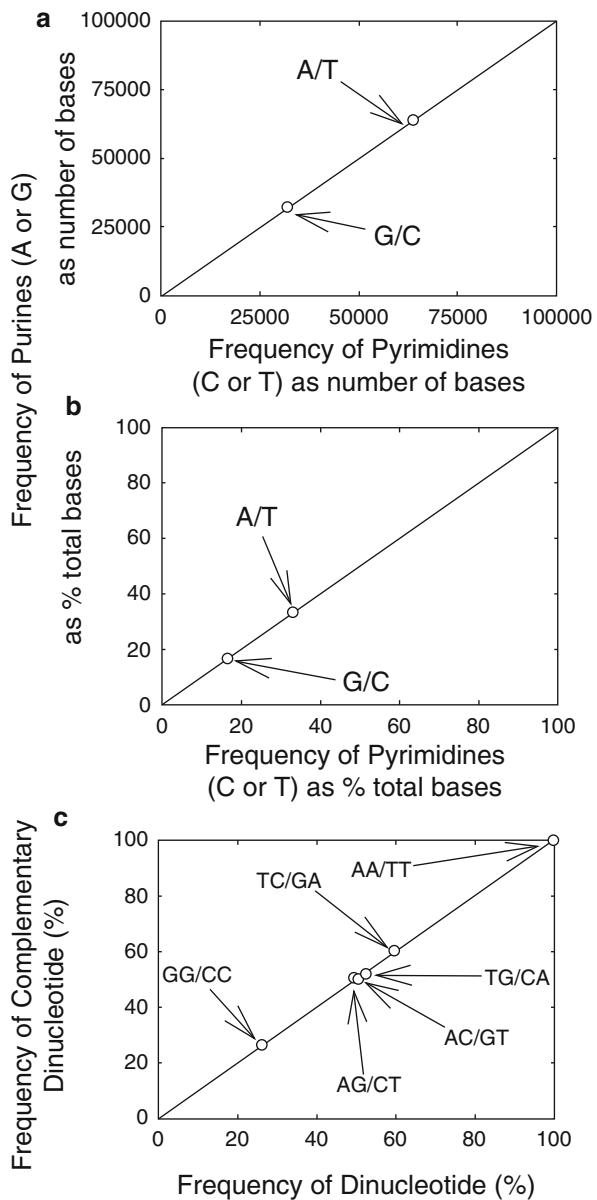


Figure 4.2 shows that the relationship holds for 3-mers in a low (G+C)% genome (*Vaccinia virus*), in an intermediate (G+C)% genome (a segment of human chromosome 19), and in a high (G+C)% genome (*Herpes simplex virus*). In the low (G+C)% genome there are many AT-rich 3-mers (rich in the W-bases) and few GC-rich 3-mers (rich in the S-bases; Fig. 4.2a). So, 3-mers of general formula WWW (W₃) are abundant (e.g. ATA with its complement TAT), and 3-mers of

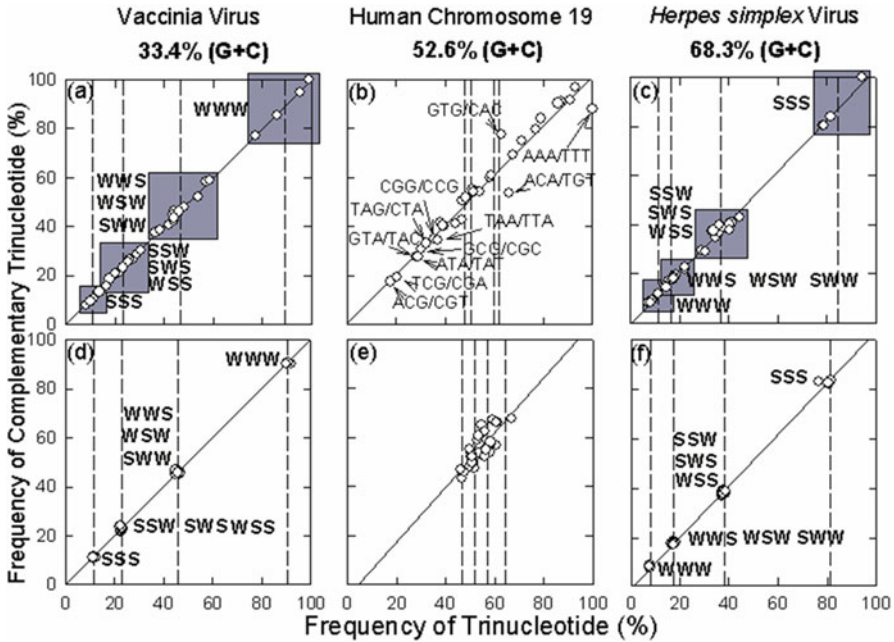


Fig. 4.2 Approximate equifrequencies of complementary trinucleotides in the natural (a, b, c) and randomized versions (d, e, f) of the 'top' strand of the *Vaccinia* virus genome (a, d), a human chromosome 19 segment (b, e), and the *Herpes simplex* virus genome (c, f). Frequencies are expressed as percentages of the most abundant trinucleotide in the natural sequence. Each member of the **32** sets of trinucleotide pairs is assigned to either the X-axis or the Y-axis in a standardized way (e.g. in the case of the point marked "GTG/CAC" the frequency of GTG refers to the Y-axis, and the frequency of CAC refers to the X-axis). "W" represents either A or T, and "S" represents either G or C. In (a) and (c) gray boxes surround points corresponding to the **8** members of the W_3 group of trinucleotides (4 complementary pairs), the **24** members of the W_2S group, the **24** members of the S_2W group, and the **8** members of the S_3 group. Vertical dashed lines indicate the average frequency among members of each group. The best line fitting the 32 points (regression line) forms the diagonal

general formula SSS (S_3) are not abundant (e.g. GCG with its complement CGC). The converse applies to the high $(G+C)\%$ genome (Fig. 4.2c). In the case of the intermediate $(G+C)\%$ genome, 3-mers corresponding to the extremes (W_3 and S_3) are more generally distributed (Fig. 4.2b).

2-mers (dinucleotides) and 3-mers (trinucleotides), are examples of oligonucleotides (Greek: *oligo* = few). The presence of a given oligonucleotide (k-mer) in a sequence is partly a reflection of the overall $(G+C)\%$ composition of the sequence, and partly a reflection of the *order* of the bases. Thus, GCG would satisfy a pressure for a high $(G+C)\%$ genome, but the pressure would be equally satisfied by other combinations of G and C (e.g. by CGG , GGC , CGC ; see Appendix B).

To distinguish pressures on base *composition* (i.e. pressures determining mononucleotide frequency) from pressures on base *order* (i.e. pressures determining oli-

gonucleotide frequencies), sequences can be shuffled (randomized). This maintains the base composition (mononucleotide frequency), but disrupts the natural order of the bases, hence randomizing oligonucleotide frequencies, (which can then be directly predicted from mononucleotide frequency; see later).

When this is done the data points become much less scattered (Figure 4.2d, e, f). Furthermore, as DNA segments depart from 50% **G+C**, there is an *increasing* influence of base composition (1-mer frequency) on n-mer frequency (i.e. the oligonucleotide frequency becomes more, but not entirely, predictable from mononucleotide frequency; see later). 3-mers are distributed among four distinct groups (centered at X-axis values marked by vertical dashed lines in Figure 4.2). These are the **W₃**, **W₂S**, **S₂W**, and **S₃** groups [13].

A visual measure of the influence of base order is provided by plotting the 64 frequencies of trinucleotides of the shuffled sequences against the corresponding 64 frequencies of trinucleotides of the natural sequences (Fig. 4.3). Here the degree of horizontal dispersion provides a measure of the role of base order. One ‘principle component’ of the plots can be distinguished from another. There is much more scope for base order in the DNA segment with an intermediate (**G+C**)% value (52.5%; Fig. 4.3b). In other words, departures from base equipfrequencies *constrain* base order. Thus, at extreme (**G+C**)% values a biological factor dependent on base composition could more likely overrule a biological factor dependent on base order (see Chapter 12).

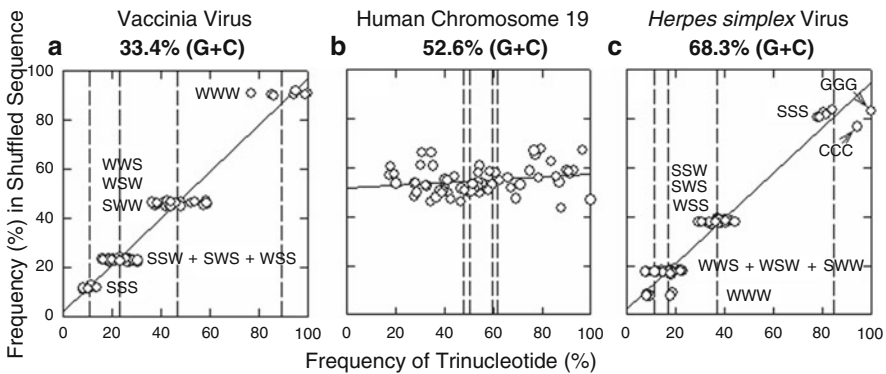


Fig. 4.3 Comparison of frequencies of the 64 trinucleotides in randomized (‘shuffled’) sequences with their frequencies in the corresponding natural, unshuffled, sequences. Trinucleotide frequencies, expressed as a percentage of the most abundant trinucleotide in the natural sequence, are plotted against each other. For other details see Figure 4.2. This procedure allows us to dissect out one ‘principle component’ of the plot, base order, from another ‘principle component,’ base composition (see Appendix A)

Species Specificity of K-mer Frequencies

Base compositions (1-mer frequencies) often differ between species (see Chapter 10). But even when species have similar base compositions, they usually differ at higher k-mer levels [14–16]. To display this, frequencies of k-mers in DNA from one source can be plotted against the frequencies of the same k-mers from another source [17]. For trinucleotides (3-mers) there are 64 data points. Figure 4.4a shows that when one segment of human DNA (from chromosome 4) is plotted against another segment of human DNA (from chromosome 19), which has about the same (G+C)%, the points fall close to the diagonal. Yet, a plot of human DNA (chromosome 19) against a segment of DNA from the gut bacterium *Escherichia coli*, which has about the same (G+C)%, reveals no correlation (Fig. 4.4b). The *E. coli* sequence behaves much as the shuffled human sequence shown in Figure 4.3b.

Thus, with respect to human DNA, *E. coli* DNA would appear as randomized sequence. But two segments of DNA from different parts of the human genome, which encode *entirely different* functions, have *very similar* base order-dependent contributions to k-mer frequencies (Fig. 4.4a). Given its particular base-composition, and other pressures being equal, the genome of a given species will opt for particular sets of k-mers, rather than others. Genome k-mer frequency values tend to be *species specific*, and *relatively uniform* throughout a genome, independently of the underlying functions [18].

One cause of this is a pressure in organisms considered higher on the evolutionary scale (e.g. humans, fruit fly) to loose the self-complementary dinucleotides **CG** and **TA** (see Chapter 18). This is not so strange. K-mer frequencies also show specificity

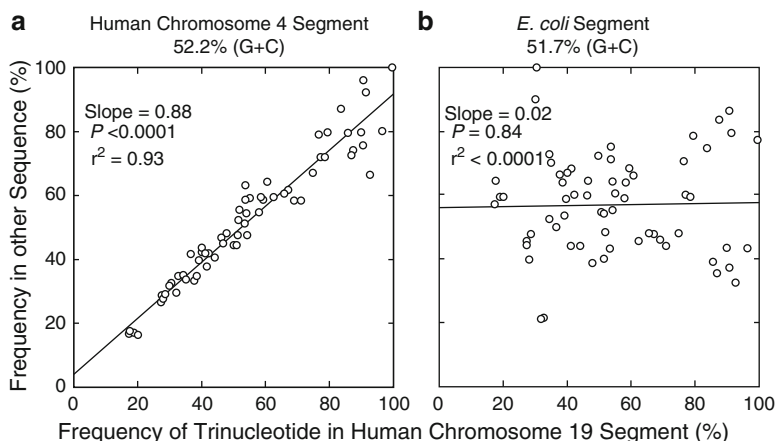


Fig. 4.4 Comparison of frequencies of 64 trinucleotides in different sequences each with G+C percentages around 50%. Trinucleotide frequencies in either a segment from human chromosome 4 (a), or a segment from the bacterium *E. coli* (b), were plotted against the corresponding trinucleotide frequencies in a segment from human chromosome 19 [13]

in spoken languages. For example, the 2-mer “sz” is found in Polish names (e.g. Szybalski), but infrequently in English names. Because of this pressure, acting primarily at the dinucleotide level, there may be a deficiency in associated trinucleotides (CGA, CGC, CGG, CGT) and their reverse complementary set (ACG, CCG, GCG, TCG). Thus, the question arises if, and under what conditions, *low order* oligonucleotide frequency drives *intermediate order* (say 4–6-mers), and *high order*, oligonucleotide frequencies?

In Figure 4.4a there is a high slope (0.88) and the probability that the difference from zero slope had occurred by chance is very low ($P < 0.0001$). In contrast, in Figure 4.4b the slope is low (0.02) and the probability that the difference from zero slope had occurred by chance is very high ($P = 0.84$). Thus, the former slope is highly significant, but the latter is not (see Appendix A). By plotting DNA segments from different species against each other in this manner, we can obtain a measure of species similarity, and this generally fits in with evolutionary (phylogenetic) expectations [13, 18].

Another measure is the square of the correlation coefficient. In Figure 4.4a 93% of the variation between the two human segments can be explained by the correlation ($r^2 = 0.93$). Only 7% remains unaccounted for (‘missing heritability’). In contrast, in Figure 4.4b there is no correlation ($r^2 < 0.0001$). 100% of the variation remains to be accounted for. Again, by comparing r^2 values (or just r values) for DNA segments from different species, the degree of similarity can provide a measure of phylogenetic relationship.

Metagenomics

For DNA to be sequenced, all or part of its host organism is first isolated from other organisms in its environment. Then DNA is extracted and purified. Eventually a set of sequenced DNA fragments is obtained. These can be matched end-to-end by virtue of sequence overlaps, to generate an entire genomic sequence. However, because each species has a distinctive hierarchy of oligonucleotide frequencies it is possible to identify species-specific fragments from the total DNA obtained directly from a particular environment, *without a prior isolation of organisms*. For example, there are about 160 different species of microorganisms in one milliliter of sea water. DNA can be prepared directly from a sample of sea water and the resulting fragments then sequenced. Assignment of a particular fragment to a particular species can be made *retrospectively* based on differences in oligonucleotide hierarchies.

Thus, in principle, the genomes of multiple species occupying a particular habitat can be simultaneously sequenced *without* identifying each species in advance. This process, sometimes referred to as “metagenomic” analysis, can lead to the discovery of new species. *It signifies the existence of a level of information in DNA sequences that may be used to distinguish species*. Whether such species-identifying information appears *after* two species have diverged from a common ancestral

species, or can be an agency *initiating* the process of divergence (speciation), is considered in Part III.

For some purposes we can confine our study to species with similar (G+C)% values as in Figure 4.4. This avoids the effects on oligonucleotide frequencies of differences in base composition (i.e. differences in mononucleotide frequencies). Alternatively, various software packages provide ‘principle component analysis,’ which allows us to dissect out the role of (G+C)% differences, so that the contributions of other factors (i.e. base order) to the variation can be assessed (Appendix A).

Introns Obey Better Than Exons

In Figure 4.4a, two human DNA segments are compared. These segments each contain genic DNA and non-genic DNA. Furthermore, the genic DNA consists both of segments that appear in the cytoplasmic RNA derived from those genes (exons), and of segments that do not appear in the cytoplasmic RNA derived from those genes (introns; see Chapter 13). In that the data points all lie close to the diagonal, it can be inferred that, within a species, genic and non-genic DNA oligonucleotide frequencies *do not greatly differ*. Indeed, from a study of dinucleotide frequencies John Subak-Sharpe and his colleagues in 1976 deduced that all segments “have evolved in response to the *same* major selection pressure” [15]:

The general design of bulk nuclear DNA is the result of unidentified constraints ... which affect all of the DNA ... DNA fractions of fundamentally different function ... would have arrived at the same general design by virtue of flexibility inherent on the one hand in the genetic code, and on the other hand in *unidentified properties of base composition*.

Oligonucleotide frequency differences can be more precisely evaluated using either the above graphical approaches (Figs. 4.2, 4.3, and 4.4), or principle component analysis. While intronic DNA and non-genic DNA of mammalian genes have similar oligonucleotide frequencies, it is possible to demonstrate small differences from those of exonic DNA. In particular, introns and non-genic DNA obey PR2 better than exons.

This was reported in 1989 for mammalian genes by Tetsuya Yomo and Susumo Ohno [19] who concluded: (i) “coding and noncoding regions of genes are governed by the same universal construction rule,” (ii) “an intrinsic force has been exercising a far greater influence than natural selection,” and (iii) “in spite of evolutionary constraints imposed by natural selection, coding regions might also be maintaining a surprising degree of symmetry with their complementary strands.” In 2003 Emanuele Bultrini and her colleagues in Italy reported similar results for fruit fly and a nematode worm. They pointed out that symmetry operates over a short-range, and considered the possibility that intronic and non-genic DNA sequences might more readily form stem-loop structures than exonic sequences [20]:

A symmetrical trend is apparent on a scale of a few kilobases in individual ... introns. This short-range property of introns is not simply due to their symmetrical base composition,

since it is drastically reduced in randomized introns. Rather, it results from the *preferred* use of reverse complementary oligomers.... It would be tempting to link the above symmetry properties of introns to formation of stem-loop structures.

By this time, even those who had long thought PR2 would be explained solely in terms of “mutational biases” (see below), had come round to the view that genomes contain palindromic sequences that “may be under selective pressure to preserve their palindromic character and therefore follow PR2 (as pure palindromic sequences are effectively base paired)” [21]. Oligonucleotide equipfrequencies do indeed imply a potential of sequences to adopt secondary structures (see Chapter 5), and may even provide a way of locating exons in uncharted DNA sequences (see Chapter 13).

Level of Selection?

All this bears on the question as to whether Nature *first* ‘wrote’ sequences with single base parity, from which the observed parity of complementary dinucleotides, trinucleotides, and higher order oligonucleotides, is an automatic consequence? Phrased evolutionarily, the question is, do selective pressures, be they extrinsic or intrinsic to the organism, operate such as to promote directly single base parity? Alternatively, do the pressures operate at another level so that single base parity is an automatic *consequence* of a primary pressure for parity acting at another (higher) level? These horse-and-cart questions were posed in 1995 as follows [13]:

Did evolutionary forces select for the Chargaff ratios in single DNA strands, with equality of complementary oligonucleotide frequencies being an automatic consequence? Alternatively, did evolutionary forces select for equality of complementary oligonucleotide frequencies, with the Chargaff ratios being an automatic consequence?

If Nature ‘writes’ a sequence with parity at the oligonucleotide level, then parity at the single base level is an automatic consequence. However, the converse does not usually apply. If evolutionary forces cause a sequence to be ‘written’ with parity at the level of single bases, then, however long the sequence, parity at the oligonucleotide level does not necessarily follow. Appendix C shows that PR2 does not operate at the oligonucleotide level in an artificial sequence that is but one of a *large set of sequences* that Nature, while constrained by PR2 at the mononucleotide level (i.e. constant base composition), *could* have ‘written.’ Yet, when we look at natural sequences we find that PR2 operates universally at the oligonucleotide level (*a much narrower set of sequences*). Since natural sequences generally show this parity at the oligonucleotide level, it is probable that they were initially ‘written’ at that level [22].

Thus, it is likely that sequences that are of biological origin, and are of sufficient length, demonstrate parity at the single base level because ‘Nature’ initially ‘wrote’ them under the influence of evolutionary forces that required parity at the oligonucleotide level (perhaps to form stems in stem-loop secondary structures). Nature requires parity primarily at the oligonucleotide level and, by default, there is parity at the mononucleotide level.

This view has been supported by Pierre-François Baisnée and his colleagues using a statistical approach [23]. For the mathematically inclined we can say that, on a random basis, given that $\%A = \%T$ in a sequence, then the probability of **A** (written as $P(A)$) at any position will equal the probability of **T** (written as $P(T)$) at any position. Thus, $P(A) = P(T)$. On a random basis, the probabilities of finding the corresponding dinucleotides would be written as $P(AA)$ and $P(TT)$.

It follows that $P(AA) = P(A) \times P(A) = P(A)^2 = P(T)^2 = P(T) \times P(T) = P(TT)$. In 1995 in the USA, biochemist Noboru Sueoka argued that PR2 follows from mutational biases acting randomly at the mononucleotide level [24]. However, the actual observed frequencies of higher oligonucleotides (e.g. **AA** or **TT**), cannot be accurately predicted simply from the frequencies of the corresponding mononucleotides. Thus, if you accept values for $p(A)$ and $p(T)$ based on the actual observed frequencies of these single bases, you cannot accurately predict the frequencies of **AA** and **TT** by multiplying the probabilities of the single bases. Similarly, if you accept probability values for the four mononucleotides $P(A)$, $P(C)$, $P(G)$, $P(T)$, and for the dinucleotide $P(AA)$, based on actual observed frequencies, then you cannot accurately predict the observed frequencies of the seven trinucleotides containing **AA** (i.e. **AAA**, **AAC**, **AAG**, **AAT**, **CAA**, **GAA**, **TAA**).

By this probabilistic reasoning, technically known as Markov chain analysis, you cannot accurately predict the frequencies of intermediate order oligonucleotides (e.g. 4-mers) from the frequencies of mono-, di-, or trinucleotides (low order oligomers). At these lower k -mer levels this is a one-way-street. You can go down, but not up, unless special conditions prevail (i.e. it is conditional; Appendix C). However, it turns out that you *can* predict frequencies of high order oligonucleotides (e.g. 11-mers) from the frequencies of low and intermediate order oligonucleotides (3-mers and 4-mers). At this level you go up as well as down (with limits set by the actual length of the sequence under study). In short, there is a *random* element at the higher level and a *non-random* element at the lower level. Gregory Phillips and his colleagues concluded from studies in *E. coli* [25]:

The Markov chain rule did not accurately estimate the frequencies of di- and trinucleotides. However, it became relatively accurate at the tetranucleotide level and higher. This implies that constraints affecting nucleotide frequencies occur at the trinucleotide level or lower.

Blaisnée argued persuasively that parity (i.e. symmetry of base composition between two strands in a duplex) derives largely from forces acting at higher oligonucleotide levels [23]:

The underlying assumption is that base composition symmetry results from single point mutations that equally affect complementary strands. ... In addition higher-order symmetry is widely considered, implicitly or explicitly, as the consequence of first-order symmetry. ... [However], reverse-complement symmetry does not result from a single cause, such as point mutation [i.e. mutational bias], ... but rather emerges from the combined effects of a wide spectrum of mechanisms operating at multiple orders and length scales.

At the heart of the problem is the need to identify a distinct selective force operating at one or other level. We shall see in Chapter 5 that, concerning nucleic acid structure, a primary pressure operates at least at the level of complementary dinu-

cleotides (which include the self-complementary dinucleotides; Table 4.1). Indeed, Israeli biochemist Ruth Nussinov even suggested that dinucleotide frequencies are more fundamental than trinucleotide frequencies, so that we should seek to explain the latter in terms of the former, and not *vice versa* [26]. We shall see in Chapter 10 that protein-encoding regions are likely to be particularly informative, since here there is a distinct selective force (“protein-pressure”) that can be presumed not to act in non-protein-encoding regions. Indeed, studies with microorganisms by Fran Supec and his colleagues in Croatia show that genomes are dominated by oligonucleotide frequencies that can *override protein-pressure* to the extent that the latter is reflected in amino acid compositions (AAC) [27]:

We find that **G + C** content, the most frequently used measure of genomic composition, cannot capture diversity in AAC and across ecological contexts. However, di-/trinucleotide composition in intergenic DNA predicts amino acid frequencies of proteomes to the point where very little cross-species variability remains unexplained (91% of variance accounted for). Qualitatively similar results were obtained for 49 fungal genomes, where 80% of the variability in AAC could be explained by the composition of introns and intergenic regions. ... A corollary is that the previously proposed adaptations of proteomes to environmental challenges ... may need to be reinterpreted, while taking into account the evolutionary forces shaping DNA oligonucleotide frequencies.

Inter-chromosomal Accounting

We have seen that, at least by virtue of the stems in DNA stem-loop structures, there would have been an evolutionary pressure for single-strands of DNA to have approximate equivalences of the Watson-Crick pairing bases. Thus, Chargaff’s PR2 is consistent with single strands of DNA having considerable potential for forming secondary structure.

However, loops with some numerically unpaired bases (see sequence 4.4) make up a large part of most structures, so that PR2 for single-stranded nucleic acid would be expected to be much less precise than PR1 for nucleic acid in duplex form. Yet in some organisms PR2 shows remarkable precision (e.g. see the above base composition of Vaccinia virus, which consists of one linear chromosome). Explanations for this precision include the possibility of a need for some base-pairing between loops, which may not necessarily be closely colocalized in a single DNA strand. Thus, we begin to recognize the possibility of long-range base pairing interactions between distant loops [8]. We can refer to this as *intra*-chromosomal accounting, which provokes the question as to whether there could be *inter*-chromosomal accounting?

Fundamental to the problem of inter-chromosomal accounting is chromosome orientation. A linear chromosome can be considered freely rotatable through 180 degrees, and which DNA end is ‘left,’ and which DNA end is ‘right’ is decided by the original human mappers or sequencers of the DNA, often in an *arbitrary* manner. However, orientation affects the sign of Chargaff differences. It follows from PR1 (for duplex DNA) that if **A** > **T** in the top strand, then **T** > **A** in the bottom strand (i.e. the strands will have equal violations of PR2, but with *opposite* signs).

While the human sequencers may have designated one strand of each chromosome as the 'top' strand and entered the corresponding sequences in GenBank, 'Nature' may have chosen a different orientation. Assuming that one strand of a chromosome provides a relatively uniform accounting template, which strand of another chromosome would 'Nature' have picked to co-account with that strand?

An example for a hypothetical three chromosome organism should make this clearer. We consider just the **W** bases (**A** and **T**). If 'Nature' would like PR2 to cover the entire genome, then in the first chromosome the top strand difference (**A** – **T**) might be +100 (i.e. **A** > **T**). In the second chromosome the top strand difference might be –25 (i.e. **A** < **T**). In the third chromosome the top strand difference might be –75 (i.e. **A** < **T**). The total difference for the three chromosomes would be zero ($100 - 25 - 75 = 0$). 'Accounting' would be perfect. But if human sequencers assigned the third chromosome the opposite orientation, the total difference for the three chromosomes would be $100 - 25 + 75 (=150)$. Thus, for any set of chromosomes, unless there are compelling reasons for the orientation displayed in GenBank, we have to explore alternative orientations. Taking this into account, a study of the six chromosomes of the nematode worm suggests that there may indeed be some inter-chromosomal accounting so that PR2 may apply to entire genomes [28].

Base Frequency Interdependencies

Much of the discussion in this chapter might appear laborious to someone with a mathematical turn of mind. For example, some observations (Fig. 4.2) appear as a simple application of the binomial theorem (i.e. of complementary trinucleotides there are 4 pairs in the **S**₃ (**SSS**) group, 12 pairs in the **S**₂**W** group, 12 pairs in the **W**₂**S** group, and 4 pairs in the **W**₃ group). However, non-mathematical members of the species *Homo bioinformaticus* should usually be able to get there, albeit slowly, by the careful use of pencil and paper as indicated in Appendix C. And there are now powerful computers and programs (e.g. principle component analysis) that can take us into realms that previously only the mathematically gifted could contemplate.

Recently, Robert Sinclair in Japan [29] has pointed to more mathematical regularities. A base in a sequence (assumed circular) *must* be both *preceded* and *followed* by any one of the four bases (**N**). In other words, for **G**, **NG** = **GN**. The frequency of the four dinucleotides **NG** must equal the frequency of the four dinucleotides **GN**, and likewise for **NA** and **AN**, etc.. Thus:

$$\mathbf{AG} + \mathbf{CG} + \mathbf{TG} + \mathbf{GG} = \mathbf{GA} + \mathbf{GC} + \mathbf{GT} + \mathbf{GG} \quad (4.6)$$

$$\mathbf{AG} + \mathbf{CG} + \mathbf{TG} = \mathbf{GA} + \mathbf{GC} + \mathbf{GT} \quad (4.7)$$

So the frequency of one 2-mer can be calculated from the frequencies of a few others. The question arises as to whether these mathematical dependencies have been sensed by living systems and serve some biological need, such as error-correction

(see Part VI). Indeed, different mathematical dependences may have arisen sequentially over evolutionary time, so it is possible that the Sinclair dependencies, or others that remain to be discovered, might have preceded those determined by Chargaff (or vice versa) [30–32].

Summary

Prose, poetry and palindromes can be seen as informational devices that trade-off increasing degrees of redundancy in order to increase error-detecting power. The most extreme of these is the palindrome which, if a general restriction, would severely compromise the normal transfer of information between humans, while greatly decreasing the chances of error. That hereditary information in the form of DNA sequences is palindrome-like, suggests that evolutionary pressures for error-detection may be at least as powerful as those for the encoding of primary messages. Given limits on genome space, there are potential conflicts between different forms and levels of information. Classical Darwinian selective forces in the environment acting on an organism's form and function provide *extrinsic* constraint, but genomes are also under *intrinsic* constraint. In palindromes there is a one-to-one pairing relationship between symbols (letters, bases). In DNA this finds expression as Chargaff's second parity rule (PR2), namely Chargaff's first parity rule for duplex DNA (PR1) also applies, to a close approximation, to single-stranded DNA. Selective forces generating the PR2 equivalences may have acted at higher oligomer levels than that of single bases. Diminished PR2 equivalences in protein-coding regions suggests diminished potential for secondary structure (stem-loops) in these regions. The *intra*-chromosomal 'accounting' that generated the mathematical regularities noted by Chargaff, may have been supplemented by *inter*-chromosomal 'accounting.'

References

1. Dickens C (1837) *The Posthumous Papers of the Pickwick Club*. Chapman-Hall, London
2. Moliere (1670) *Le Bourgeois Gentilhomme*. Oliver TE (ed) Ginn, New York (1914)
3. Bök C (2001) *Eunoia*. Coach House Books, Toronto
4. Perec G (1972) *Les Revenentes*. Julliard, Paris
5. Perec G (1969) *La Disparition*. Denoël, Paris [E. V. Wright wrote *Gadsby* (1939) without using the letter 'e'.]
6. Ohno S (1991) The grammatical rule of DNA language: messages in palindromic verses. In: Osawa S, Honjo T (eds) *Evolution of Life*. Springer-Verlag, Berlin, pp 97–108
7. Rudner R, Karkas JD, Chargaff E (1968) Separation of *B. subtilis* DNA into complementary strands. III. Direct analysis. *Proceedings of the National Academy of Sciences USA* 60:921–922
8. Bell SJ, Forsdyke DR (1999) Accounting units in DNA. *Journal of Theoretical Biology* 197:51–61
9. Darwin C (1871) *The Descent of Man and Selection in Relation to Sex*. John Murray, London, p 316

10. Edwards AWF (1998) Natural selection and the sex ratio. *American Naturalist* 151:564–569
11. Forsdyke DR (2002) Symmetry observations in long nucleotide sequences. *Bioinformatics* 18:215–217
12. Prabhu VV (1993) Symmetry observations in long nucleotide sequences. *Nucleic Acids Research* 21:2797–2800
13. Forsdyke DR (1995) Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. *Journal of Molecular Evolution* 41:573–581
14. Josse J, Kaiser AD, Kornberg A (1961) Enzymatic Synthesis of Deoxyribonucleic Acid. VIII. Frequencies of nearest neighbor base sequences in deoxyribonucleic acid. *Journal of Biological Chemistry* 236:864–875
15. Russell GJ, Walker PMB, Elton RA, Subak-Sharpe JH (1976) Doublet frequency analysis of fractionated vertebrate DNA. *Journal of Molecular Biology* 108:1–23
16. Blaisdell BE (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences USA* 83:5155–5159
17. Rogerson AC (1989) A sequence asymmetry of the *Escherichia coli* chromosome appears to be independent of strand or function and may be evolutionarily conserved. *Nucleic Acids Research* 17:5547–5563
18. Rogerson AC (1991) There appear to be conserved constraints on the distribution of nucleotide sequences in cellular genomes. *Journal of Molecular Evolution* 32:24–30
19. Yomo T, Ohno S (1989) Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency – TG/CA excess. *Proceedings of the National Academy of Sciences USA* 86:8452–8456
20. Bultrini E, Pizzi E, Guidice P Del, Frontali C (2003) Pentamer vocabularies characterizing introns and intron-like intergenic tracts from *Caenorhabditis elegans* and *Drosophila melanogaster*. *Gene* 304:183–192
21. Lobry JR, Sueoka N (2002) Asymmetric directional mutation pressures in bacteria. *Genome Biology* 3(10):research 0058.
22. Forsdyke DR, Bell SJ (2004) Purine-loading, stem-loops, and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. *Applied Bioinformatics* 3:3–8
23. Baisnée P-F, Hampson S, Baldi P (2002) Why are complementary strands symmetric? *Bioinformatics* 18:1021–1033
24. Sueoka N (1995) Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *Journal of Molecular Evolution* 40:318–325
25. Phillips GJ, Arnold J, Ivarie R (1987) The effect of codon usage on the oligonucleotide composition of the *E. coli* genome and identification of over- and underrepresented sequences by Markov chain analysis. *Nucleic Acids Research* 15:2627–2638
26. Nussinov R (1981) Eukaryotic dinucleotide preference rules and their implications for degenerate codon usage. *Journal of Molecular Biology* 149:125–131
27. Brbić M, Warnecke T, Kriško A, Supek F (2015) Global shifts in genome and proteome composition are very tightly coupled. *Genome Biology & Evolution* 7:1519–1532
28. Forsdyke DR, Zhang C, Wu J-F (2010) Chromosomes as interdependent accounting units. The assigned orientation of *C. elegans* chromosomes minimizes the total W-base Chargaff difference. *Journal of Biological Systems* 18:1–16
29. Sinclair R (2015) Necessary relations for nucleotide frequencies. *Journal of Theoretical Biology* 374:179–182
30. Rapoport AE, Trifonov EN (2013) Compensatory nature of Chargaff's second parity rule. *Journal of Biomolecular Structure and Dynamics* 31:1324–1336
31. Afreixo V, Rodrigues JMOS, Bastos CA (2015) Analysis of single-strand exceptional word symmetry in the human genome: new measures. *Biostatistics* 16:209–221
32. Yamagishi MEB, Herai RH (2011) Chargaff's 'grammar of biology': new fractal-like rules. arXiv:1112.1528 (<http://arxiv.org/abs/1112.1528>)

Chapter 5

Stems and Loops

I am puzzled by the frequent appearance of palindromes in these sequences. Naturally one would expect occasional short ones by chance, but they seem to occur too often. ... I have wondered if there is any mechanism which would produce them, but have been unable to think of one.

F. Crick, 1974 [1]

The sequence of a single-stranded nucleic acid is referred to as its primary structure. When it forms a duplex by pairing with a complementary single strand it adopts a secondary structure. Chargaff's first parity rule (PR1) for duplex DNA was consistent with the Watson-Crick idea of a base in one strand of the duplex pairing with a complementary base in the other strand of the duplex. This *inter*-strand base pairing stabilized the secondary structure (Fig. 2.1). By the same token, the existence of a parity rule for single strands (PR2) suggested *intra*-strand base pairing. At least by virtue of the composition of the stems in stem-loop secondary structures, there should be approximately equivalent quantities of the classical pairing bases. Do single-stranded nucleic acids have the potential to form such intra-strand secondary structures? If so, is this a chance event, or are adaptive forces involved? These questions began to be addressed when the sequences of various transfer RNAs and bacterial viruses became available in the 1970s. It became evident that nucleic acid structure ('flaps') is a form of information that has the potential to conflict with other forms.

Transfer RNA

Having adopted a stem-loop secondary structure, a single-stranded nucleic acid can participate in more elaborate intra-strand bonding to generate higher-ordered structures (tertiary, quaternary, etc.). However, for many purposes nucleic acid structure can be adequately discussed in terms of secondary structure. This can be simply drawn in two dimensions without showing the pairing strands as double helices. Thus, transfer RNA (tRNA) sequences can be folded to generate stem-loop-containing structures (Fig. 5.1).

In any cell engaged in protein synthesis there must be at least twenty tRNA types, corresponding to the twenty amino acids. The 3' end of each tRNA type is

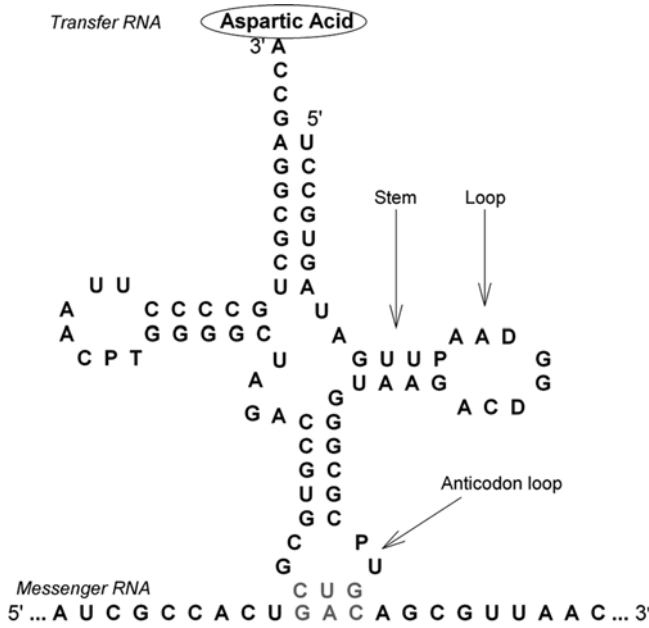


Fig. 5.1 Stem-loop secondary structure of a transfer RNA (tRNA) molecule in simple two-dimensional form. Successive triplet codons in the messenger RNA are underlined (bottom). The mRNA codon for aspartic acid (**GAC**, in gray) interacts with ('kisses') the complementary sequence at the tip of the anticodon loop of the tRNA (**GUC**, in gray). Note that in RNA there is some pairing between **G** and **U** (as well as the more usual **A–U** pairing), and tRNA molecules have a few unusual bases (here designated **P**, **D** and **T**) that are not normally found in mRNA molecules. Some bases may be methylated (see Chapter 18)

attached to a particular amino acid that will be transferred to a growing chain of amino acids (hence “*transfer RNA*”). The chain of amino acids (peptide chain) will give rise to a protein.

At the tip of one of the loops (the anticodon loop) there is a three base sequence (5' **GUC** 3' in Fig. 5.1) that is able to base pair with the complementary sequence (5' **GAC** 3') of a codon in mRNA. This is a brief, reversible interaction, which can be referred to as ‘kissing’ (see Chapter 6). The base pairing follows Watson and Crick in that the pairing strands are anti-parallel – an arrangement applying both to the intra-strand pairing within the tRNA molecule itself, and to its transient inter-strand pairing with the codon sequence in the mRNA. We should note that in RNA, unlike DNA, there is usually the base **U** instead of **T**. Like **T**, **U** pairs best with **A**, but a weaker pairing with **G** can also occur. In passing, we should also note that the base in tRNA before (on the 5' side of) the anticodon triplet is almost invariably **U**, and the base after (on the 3' side of) the anticodon triplet is usually a purine (**R**). The significance of this will be seen in Chapter 7.

The finding of distinctive secondary structures in tRNAs made sense, since a particular tRNA functions cooperatively with a specific enzyme that adds a specific amino acid to the 3' end of the tRNA. The specificity of interaction with the enzyme

(a protein) is largely determined by the structure of the tRNA, which the enzyme distinguishes from other tRNA types. Apart from tRNAs there are other RNAs whose roles are critically linked to their structure. These include RNAs that, like tRNAs, take part in protein synthesis (e.g. ribosomal RNAs; rRNAs). Much of the cytoplasm is concerned with protein synthesis (e.g. it contains many rRNA-containing particles – ribosomes – that act as protein-assembly ‘work-benches’). Accordingly, like other components of the ‘machinery’ for protein synthesis, tRNAs and rRNAs are abundant, collectively making up about 96% of total cell RNA. Their abundance made them very effective decoys distracting biochemists in the 1950s from a minority RNA type that programs the ribosomes – mRNA (see Epilogue). Once over this obstacle, biochemists saw that by examining the proteins made by ribosomes to which mRNAs of known sequence had been added, they could correlate a particular amino acid with a particular sequence in mRNA – the genetic code could be cracked (see Chapter 7).

From the above you might think that mRNAs should not have much secondary structure, which otherwise might interfere with the translation process. Yet, every RNA is a transcript derived by copying (transcribing) the sequence of another nucleic acid (usually DNA) following the Watson-Crick base-pairing rules (see Chapter 6). If it were advantageous for genomic nucleic acid to have a higher ordered structure, then mRNAs might have this structure *by default*. They are landed with it, whether they want it or not. Genomes of viruses should be informative here, since viruses employ their host cell’s machinery for protein synthesis (translating the mRNAs they encode) and often do not have information for tRNAs and rRNAs in their genomes.

The Bacteriophage Paradox

In some viruses that infect bacteria the genome is RNA (e.g. bacteriophages R17 and MS2). In others the genome is DNA (e.g. bacteriophage T4). Genomes of bacteriophage (Greek: *phagein* = to eat) consist of tightly packed genes from which mRNAs are transcribed for translation into proteins by the host’s translation machinery. These proteins of bacteriophage origin assist it to control the host and, late in the cycle of infection, include coat proteins into which new phage genomes are packaged prior to release of virus particles from the infected cell.

In 1947 a pioneer in bacteriophage studies, the biophysicist Max Delbrück, commented on a paper by Niels Bohr that had resolved a paradox in atomic physics by proposing subnuclear particles were in stable orbits [2]:

These were very irrational assumptions, which shocked ... many physicists ... The crucial point ... is the appearance of a conflict between separate areas of experience, which gradually sharpens into a paradox and must then be resolved by a radically new approach.

Bacteriophage were soon to shock similarly. In classical Darwinian terms it was assumed that bacteriophages that encoded optimal proteins were best adapted to their environment. Thus, the environment acting on virus proteins would have selected for survival the ‘fittest’ viruses whose genes encoded optimal proteins. Yet, paradoxically, it was observed that the base sequence seemed to serve the needs

of nucleic acid *structure* just as much as the protein-encoding function. Indeed, the needs of nucleic acid structure sometimes seemed to be served *better* than those of the protein-encoding function. Although not widely perceived at the time, this paradox shattered the Darwinian orthodoxy and pointed to the need for “a radically new approach.”

At what level might structure be of most selective importance – at the level of mRNAs themselves, or of the genes that encoded them, or even of the entire genome? US molecular biologist Winston Salser in 1970 opted for importance at the mRNA level [3]:

RNA phage R17 has very extensive regions of highly ordered base pairing. It has seemed likely that this might be necessary to allow phage packaging. ... [We] were therefore somewhat surprised to find that T4 messengers [mRNAs], which do not have to be packaged, also have a very large amount of secondary structure. ... Our results suggest that a high degree of secondary structure may be important in the functioning of most mRNA molecules. ... The possible functions of such extensive regions of base-pairing are unknown.

In 1972 microbiologist Andrew Ball went further [4]:

The selection pressure for specific base pairing in a messenger RNA severely limits its coding potential ... [so that] ... there is a pressure for some amino acid sequences to be selected according to criteria which are distinct from the structure and function of the proteins they constitute.

On the other hand, chemist Harold White and his colleagues, referring to the “neutral hypothesis” that had been proposed in Japan by Motoo Kimura, opted for the importance of structure at the DNA level [5]:

If hairpin loops do indeed exist in mRNA, recent assertions that mutations in redundant bases (synonymous mutations) are selectively neutral may have to be modified. ... The DNAs from which they [mRNAs] are transcribed may also possess this potential property in both strands. Normally such loops in DNA would be less stable than the linear, fully paired double helix unless an unwinding force were applied to the DNA strand. In such a case the strain of an untwisted helix could be relieved or accommodated by the formation of hairpin loops.

The paradox was solved when better methods for calculating RNA secondary structure became available. For many mRNA sequences, the energetics of the folding of natural sequences are *better* than those of sequences derived from those natural sequences by shuffling the order of the bases. A natural sequence must be shuffled and then refolded many times to arrive, by chance, at a structure approaching the stability of the original natural sequence [6–8]. Yes, ‘the hand of evolution’ has arranged the *order* of bases to support structure, sometimes at the expense of the protein-coding function.

Genome-Level Selection

Although abundantly present in cells, tRNAs and rRNAs are encoded by a relatively small part of genomes. In microorganisms (e.g. bacteria) the sequences of mRNAs are more representative of the sequences of the corresponding genomes as a whole. As suggested above by White, if many mRNAs have significant secondary structure, then the corresponding genomic regions (i.e. the genes from which the mRNAs

are transcribed) should also have this potential. The primary selection pressure for the elaboration of mRNA secondary structure might have been at the *genome* level rather than at the mRNA level. If so, regions of a genome that are *not* transcribed into mRNAs might also demonstrate potential for secondary structure.

When folding programs were applied to the sequences of individual DNA strands, it was found that there is indeed considerable potential for secondary structure *throughout* genomes. Stem-loop potential in DNA is not restricted to regions encoding mRNA (or rRNA or tRNA). It is present in intergenic DNA and, in the case of intron-containing organisms, also in introns (indeed, it is greater in introns than exons; see Chapter 13). Stem-loop potential is widely distributed throughout the genomes of all species (Fig. 5.2).

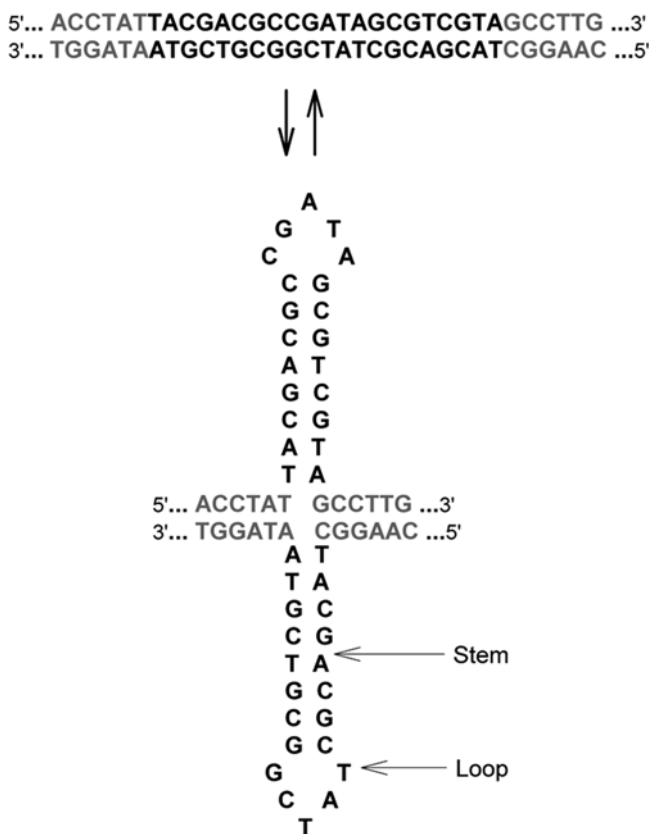


Fig. 5.2 Reversible symmetrical extrusion of stem-loop secondary structures from duplex DNA. There is classical Watson-Crick pairing in the stems. However, just as G and U can pair weakly in RNA (Fig. 5.1), so some extruded DNA structures will have weak pairing between G and T. Since the process is reversible, then, unless otherwise restrained (e.g. by interactions with DNA-binding proteins) a segment of DNA *vibrates* between the two conformations. The frequency should be lower for GC-rich DNA where structures are more stable, than for AT-rich DNA where structures are less stable. It follows that sequences of a given (G+C)% should have matching vibration frequencies and, according to Muller's model (Chapter 2), should therefore have the potential to seek each other out (see Chapter 10). Since secondary structure stabilities are also affected by base order (see later), when two sequences also have similar base orders they should have even closer matching frequencies

Calculation from Single Base Pairs

With knowledge of the Watson-Crick base-pairing rules, given the sequence of a tRNA, one should be able to arrive at the structure shown in Figure 5.1 with pencil and paper by a process of trial and error. This would be quite laborious and with longer sequences becomes impracticable. Enter computers. For present purposes it is not necessary to understand fully the programs (algorithms) that allow computers to calculate elaborate secondary structures, but the principles are relatively straightforward. Even so, it is conceptually quite tricky. Don't worry if it doesn't click the first time through. It took one of the masters of folding, Michael Zuker, several decades of exclusive focus to find the still far-from-perfect algorithm.

The energetics of nucleic acid folding can be viewed simply in terms of *base composition* and *base order*. In other words, the total folding energy can be decomposed into two components – a base composition-dependent component and a base order-dependent component. The latter is not determined directly, but by subtracting the base composition-dependent component from the total energy.

First a computer gives you a value for the total folding energy of a natural sequence. The tricky part is then getting the base composition-dependent component. To do this you have to destroy the base order-dependent component by shuffling to get a randomized sequence with the same base composition. Then you get the computer to fold this sequence and calculate the folding energy.

The trouble here is that, on a chance basis, the randomized version of the sequence that the computer first generates might just happen to have a particular base order that generates an idiosyncratic energy value. To allow for this, you get the computer to shuffle and refold the original sequence, several more times. Each time you recalculate the folding energy. Each successive randomized version has a particular base order different from the original base order of the natural sequence. But it *retains* the original base composition.

Thus, the average folding energy of the set of different randomized versions of the natural sequence reflects this commonality – their base composition. In short, the base composition-dependent component is determined by shuffling and refolding a sequence several times, thus destroying the base order-dependent component, and then taking the *average* folding energy of the resulting structures.

To begin, following Ignacio Tinoco and colleagues [9], we assign "stability numbers" of 1 and 2 to **A-T** and **G-C** base pairs, respectively. This acknowledges, in crude form, the greater strength of the pairing interaction between **G** and **C**, compared with that between **A** and **T** (Table 2.1). A series of hypothetical stem-loop structures that might be extruded from duplex DNA are shown in Figure 5.3. To simplify, we allow only a single stem, and disallow sliding of strands relative to each other. Furthermore, the role of the loop is ignored.

The first 'stem' consists of two sets of consecutive **A** residues, which do not complement each other. The stem would never form. Thus, if this were part of a natural duplex DNA, either the sequence would remain unextruded from the duplex (i.e. the two sets of consecutive **A** residues would remain paired with two sets of

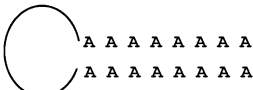
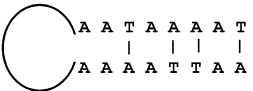
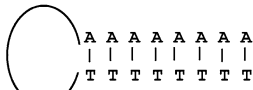
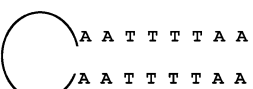
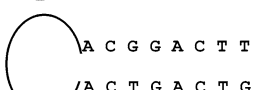
		SCORE			
		Total	=	Base composition-dependent	+ Base order-dependent
1		0	=	0	+ 0
2		4	=	3	+ 1
3		8	=	4	+ 4
4		0	=	4	+ -4
5		8	=	3	+ 5
6		0	=	3	+ -3
		FONS	=	FORS-M	+ FORS-D

Fig. 5.3 Relative contributions of base *order* and base *composition* to the stability of stems in DNA stem-loop structures. The weakly bonding A-T base-pairs (the W bases) are assigned a score of one. The strongly bonding G-C base-pairs (the S bases) are assigned a score of two. Absence of base-pairing scores as zero. The stability of a stem is measured as the total stability score, which is contributed to both by base order and by base composition. The terms FONS, FORS-M and FORS-D refer to folding energy values as described in the text

consecutive T residues on the complementary strand), or a large loop (rather than a stem) would form.

In each limb of the second stem a quarter of the bases are T residues. With the value 1 assigned to an A-T base pair, the total stability of the stem can be scored as 4, with a high score meaning high stability. Since T residues are infrequent, it is likely that Ts will be opposite As, rather than opposite Ts (i.e. if As and Ts in the proportions 3:1 were allowed to randomly combine, there would be more A-Ts than T-Ts). Thus base *composition*, rather than base *order*, can be considered to make the major contribution to the score. This is a fundamental point to which we will return.

The third stem has a set of consecutive As and a set of consecutive Ts. The total possible number of A-T base-pairs (8) is formed. The average number of A-T pairs that would be formed if 8 As and 8 Ts were randomly mixed would be 4 (try it by shaking the letters in a bag and picking out pairs). Thus, the contribution to the score attributable to base composition alone is 4. By subtraction we can determine that the base order-dependent component of the score is 4.

The fourth stem has the *same* base composition as the third, so that the potential contribution of base composition to the stability remains at 4. However, the *distribution* of As and Ts is such that no A-T base-pairs form. Since there is no base-pairing, a stem would never form. Thus, the total score is zero, and by subtraction we determine that the contribution of base order is minus 4. The bases are ordered so as to *oppose* the random tendency for sequences of 8 As and 8 Ts to form 4 A-T intrastrand base-pairs. Stem-loop extrusion from duplex DNA would be opposed. Alternatively, if associated with an extruded structure, the order of bases would favor loop, rather than stem, formation.

The fifth and sixth stems contain equal proportions of all four bases (i.e. their base compositions are identical). In the fifth stem only four complementary pairs are present. These are G-C pairs, to each of which we assign the score of 2, for a total score of 8. The relative contributions of base composition and base order are 3 and 5, respectively (as will be shown below). The sixth stem has no base pairing, so the total score is zero. The potential contribution of base composition remains at 3 so that, by subtraction, the contribution of base order is minus 3.

In summary, Figure 5.3 shows that the correct bases may be present in the correct proportions, but if base order is inappropriate there may be no stem (zero total score). Stem stability depends on base composition in two ways. Complementary bases must be present in equal proportions, and the more G-C pairs there are, the higher will be the stability. Conversely, the more A-T pairs there are, the lower will be the stability.

Role of Base Composition

If the fifth stem in Figure 5.3 were a natural sequence, then we could call the total score (which sums to 8) the "folding of natural sequence" (FONS) value. How do we calculate the relative contributions of base composition and base order to that score? The stem has a unique characteristic, its base order, and two other characteristics that it shares with a *large set* of other possible DNA sequences, its length and base composition. The natural sequence is but one member of a hypothetical set of sequences that share length and base composition. Any *average* characteristic of this hypothetical set must reflect their shared lengths and base compositions. This is another, quite fundamental, point.

By keeping length constant, we can focus on the role of base composition. If the order of bases is randomized (shuffled), keeping the length constant, then members of the set which differ only in base order, are obtained. The sixth stem in Figure 5.3 is one member of the set. Figure 5.4 shows 10 other members of the set. These happened, by chance, to be those *first* generated by consecutively shuffling base order in the fifth stem in Figure 5.3. Associated with each member is a total score. Each score is a "folding of randomized sequence" (FORS) value for the fifth stem in Figure 5.3.

It is seen that most members score less than the natural sequence. The mean score of the set of 10 is 3.3, which approximates to 3. This is the "folding of randomized

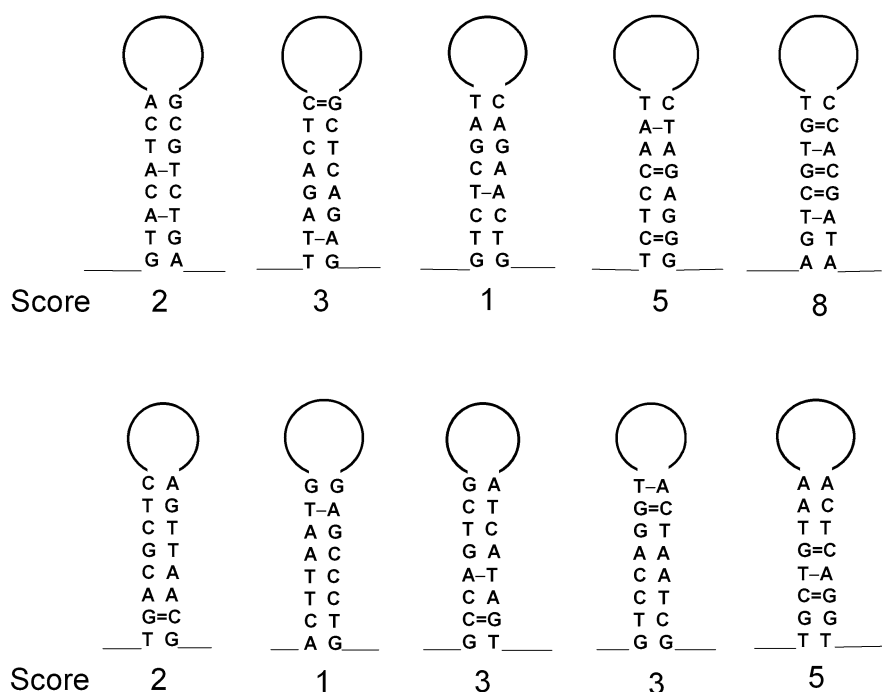


Fig. 5.4 A set of stem-loop structures that happened to be generated *first* by successively randomizing (shuffling) the order of bases in the fifth stem in Figure 5.3, and then refolding. An A–T pair scores 1. A G–C pair scores 2

sequence mean" (FORS-M) value, which is an *average* characteristic of the set, and thus should be *base order-independent*. The value provides a measure of the contribution of base composition to the FONS value for the 5th stem of Figure 5.3. By subtraction from the FONS value, the contribution of base order is found to be 5. This is the "*folding of randomized sequence difference*" (FORS-D) value, which provides a measure of the contribution of the primary sequence (base order) to the stability of stem-loop structures in a natural sequence of given length. Thus, $FONS = FORS-M + FORS-D$. In the present case, the FORS-D value is positive, and is significantly different from zero. This makes it likely that the sequence of the fifth stem in Figure 5.3 had accepted mutations to increase the formation of complementary base pairs, thus enhancing its stem potential.

It will be noted that four of the stems derived by randomization (Fig. 5.4), have lower total scores than 3. An extreme example of this is the sixth stem in Figure 5.3. To generate *by chance* a stem with no base-pairs would usually require many, many, more randomizations than the ten used to generate Figure 5.4. Thus, the sixth stem of Figure 5.3 is *highly improbable*. If the sixth stem were part of a real natural sequence, then the low FORS-D value would imply that base order had been *working against* base composition in determining the total score [6].

Calculation from Dinucleotide Pairs

If the bases of complementary Watson-Crick base-pairs just paired with each other, and neighboring bases had no influence on this, then refinement of computer folding programs would mainly require finding precise values to replace Tinoco’s two crude “stability numbers.” However, when double helices are formed, the flat bases *stack* above each other (like a pile of coins, or “rouleau”). This displaces water molecules. The increased freedom given to the water molecules (see Part V) makes a major contribution to the energetics of helix formation. The actual value depends on the nature of the two flat bases that form the ‘sandwich’ from which water molecules are liberated.

Hence, sequence (i.e. base-order) is of much importance in determining the stability of a nucleic acid duplex. The base order-dependent component can be accounted for by moving from Tinoco’s two stability numbers for single base pairs, to ten numbers for complementary dinucleotide pairs (Table 5.1). Each of these ten numbers has been determined using chemical methods [10].

For most purposes there is no need to go to higher sequence levels (e.g. use numbers corresponding to the 32 complementary trinucleotide pairs). Thus, a sequence can be decomposed into a set of overlapping dinucleotides (Figure 5.5) and the corresponding numbers summed to determine the total folding energy. A computer is able repeatedly to fold a single-stranded nucleic acid, each time calculating the score and discarding folding patterns corresponding to poor scores, until it arrives at

Table 5.1 The ten sets of complementary dinucleotides in DNA (see Table 4.1) with approximate values for pairing energies (in negative kilocalories/mol.). Values for DNA and RNA are different, but of the same order. Note that the **W** base-rich pairs have less negative values (indicating weak pairing) than the **S** base-rich pairs. Since **G** can pair very weakly with **T** (or **U**), precise analysis of stem-loop potential also requires values for complementary base pairs that include this non-classical pairing [10]

Six complementary dinucleotide pairs	A A	T T	-0.9
	A C	G T	-2.1
	A G	C T	-1.7
	C A	T G	-1.8
	C C	G G	-2.9
	G A	T C	-2.3
Four self-complementary dinucleotide pairs	T A	T A	-1.1
	G C	G C	-3.4
	C G	C G	-2.0
	A T	A T	-0.9

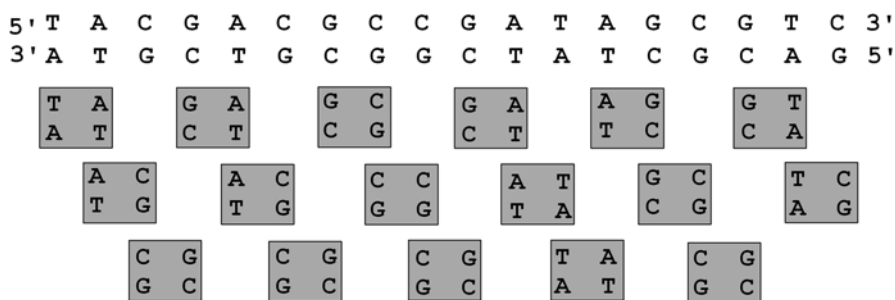


Figure 5.5 Decomposition of a duplex (which might be part of a long stem in a stem-loop structure), into individual members of the set of 10 complementary dinucleotides. Each of these can be scored using values from Table 5.1. The sum of the scores provides an overall energy value for the stem (in negative kilocalories/mol)

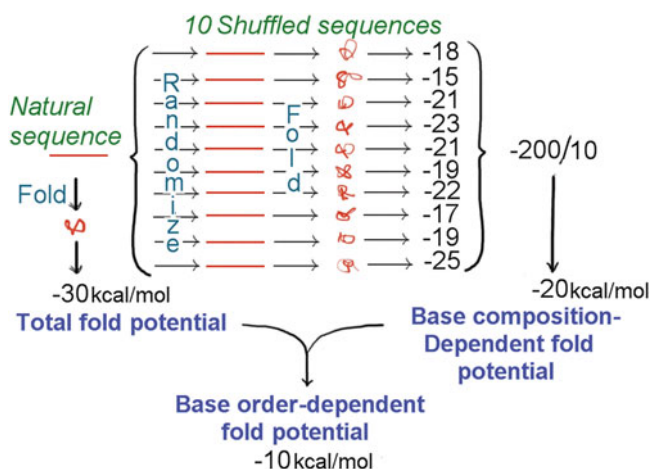


Fig. 5.6 Summary of steps in determination of base order-dependent fold potential. A natural sequence (horizontal red line at left) is computer-folded to give a structure of stability -30 kcal/mol (total fold potential). Base order is randomized to produce ten shuffled sequences that share with the natural sequence only their base composition. These are each folded to obtain corresponding stability values. Idiosyncracies due to each distinctive base order are averaged out (at right) to find the contribution of base composition to the total fold potential (-20 kcal/mol). The contribution of base order is found by subtraction

a folding pattern the score of which cannot be improved upon. For a summarizing example see Figure 5.6.

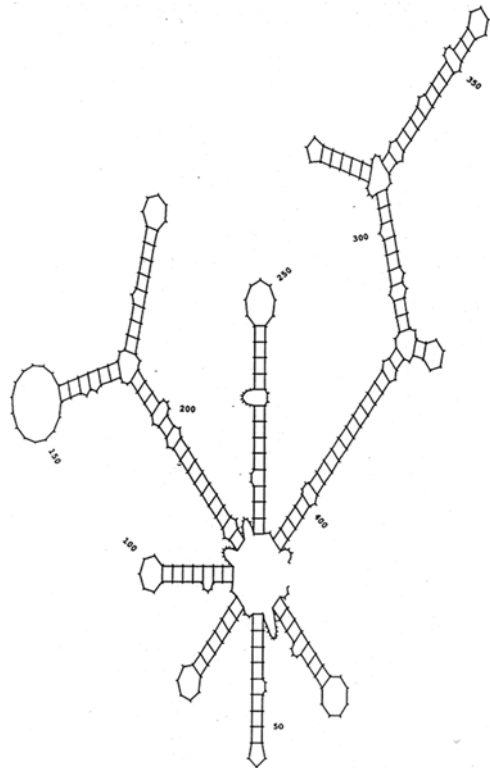
How do we decide that a calculated structure is like the actual structure the nucleic acid adopts within a complex intracellular environment? Support for calculated structures has been obtained in cases where nucleic acids have been crystallized and structures determined by X-ray crystallography. Also there are enzymes (nucleases) that recognize specific features of nucleic acid structure. A good calculated structure allows prediction as to which nucleases a nucleic acid will be susceptible.

RNA Structure and Conflict

All mRNAs have a coding region – a region containing a sequence of non-overlapping codons that must be ‘read,’ one at a time, by corresponding tRNAs, for the synthesis of a specific sequence of amino acids. As linear sequences, mRNAs may begin with a 5′ non-coding region that is followed by the coding region, and then end with a 3′ non-coding region (see Chapter 13).

Figure 5.7 shows the calculated structure of the 408 base sequence of the human mRNA encoding the cancer-related protein (oncoprotein) cFes [11]. This structure has a calculated stability of about –250 kilocalories/mol. Do not worry what the energy units mean, but note that the energy value is negative, as are the energy values for complementary dinucleotide pairs shown in Table 5.1. Greater negativity corresponds to greater stability. This is because, by convention, the stability of a structure is assessed by the energy made available (released) when the structure forms. The structure itself loses energy in this process, and it is this net energy loss (loss being negative, gain being positive) that is taken as a measure of stability. This contrasts with the positive scoring scheme used earlier in the chapter when calculating structures using the Tinoco numbers.

Fig. 5.7 Calculated structure of the mRNA for the oncoprotein c-fes. The laddering indicates individual base pairings. For present purposes all structures are “stem-loop,” but a few would be considered “bulge loops.” The structure includes some weak pairings between **G** and **U** (not shown). Extruded DNA segments corresponding to this sequence (exons) might have some pairings between **G** and **T**. When non-genic and intronic sequences are folded likewise, similar, highly significant, structures are generated



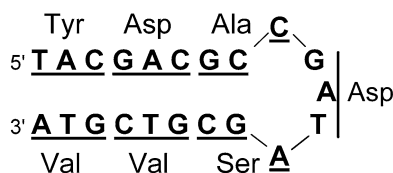


Fig. 5.8 Potential conflict between the need of a nucleic acid for a particular structure and its need to encode a particular amino acid sequence. Sequence 3.1 of Chapter 3 is here folded into a stem-loop structure. If it were part of the protein-coding region of an mRNA, it would encode the amino acid sequence **TyrAspAlaAspSerValVal**. If stem-formation is essential, then there is less flexibility in the encoding of amino acids. For example, the codon **TAC** encoding tyrosine, **Tyr**, *must* exist opposite the codon **GTA** encoding valine, **Val**. If the protein ‘wants’ the tyrosine, then it *must* accept the valine, and vice versa. Otherwise, the encoding nucleic acid must weaken, or dispense with, the stem

If one were to shuffle the natural c-fes mRNA sequence and then calculate the energetics of the resulting structure, the chances would be great that the stability *value* would be considerably ‘higher’ (i.e. would be much *less* negative) than that of the natural sequence (i.e. the stability itself would be *less*). One would have to test-shuffle and calculate folding values *many times* in order to obtain, by chance, a structure of energy value similar to that of the natural sequence. Thus, it is extremely improbable that the structure shown in Figure 5.7 would be arrived at by chance. The structure of the natural sequence is of high statistical significance.

It should be noted that the structure encompasses all parts of c-fes mRNA, including the coding region. Thus, there is a potential conflict between the ‘need’ of the nucleic acid sequence for secondary structure and its ‘need’ to encode an optimally functioning protein (Fig. 5.8).

There is also the possibility of a conflict between the ‘needs’ of the *genes* encoding tRNAs and rRNA for DNA secondary structures appropriate for their functions *in the nucleus*, and the ‘needs’ of the tRNAs and rRNAs themselves to have structures appropriate to their functions *in the cytoplasm*. The genes (DNA) have essentially the same sequences as their RNA products. Yet, the structures of tRNAs and rRNAs, arrived at in response to selective pressures operating during protein synthesis in the cytoplasm, may not optimally serve the ‘needs’ for structure of their corresponding genes. These genes may be presumed to have arrived at their structures in response to selective pressures operating during their function in the nucleus (or in the cytoplasm in organisms, or cell cycle stages, when there is no defined nucleus). Thus, the sequences of these nucleic acid molecules may be a compromise between competing demands operating at genomic (i.e. DNA) and cytoplasmic (i.e. RNA) levels. There is a potential two-fold conflict between needs at the levels of gene and of RNA.

By the same token, there is a potential three-fold conflict between needs at the levels of gene, RNA and protein. One solution is to edit RNAs *en route* from the site of transcription to the site of translation, so that their sequences differ from the sequences of the corresponding genes [12]. The positions in mRNAs where such editing occurs can be informative; editing may be in regions corresponding to the

active sites of their encoded proteins. So the needs of the encoded proteins, rather than the needs of mRNA structure, would seem to dominate [13]. Potentially harmful structure at the 3' end (that is not protein-encoding) may also require editing (see Chapters 15, 18).

That structure for mRNAs is seldom critical is shown from studies in thermophiles where rRNAs and tRNAs (which both depend on structure) have increased (G+C)% values (which enhance their stabilities), but mRNAs do not (see Chapter 11). So it is likely that mRNAs have much structure by default, because the corresponding genes require structure. When structure has meaning for an mRNA it often locates to the 5' end [14, 15].

PR2 Revisited

The high precision of PR2 appears paradoxical in that, for example, in Figure 5.7 there are also many unpaired loops. And often two sequences that could form an inverted repeat are widely separated from each other [16]. Folding into more elaborate structure than secondary can bring loops to interact with other loops by base pairing, either in close proximity or, in principle, at a distance. Evidence suggesting some form of 'accounting' *between* chromosomes (Chapter 4) makes this possibility seem plausible.

If you look carefully at Figure 5.7 you will see that there are bulges due to mismatches that interrupt what might have been long stems. The longest uninterrupted stem is 11 base pairs in length. If the entire sequence were presented in unfolded linear form you would see many inverted repeats of which the longest would be 11 base pairs (an 11-mer). If the probability of a particular base (A, C, G or T) at a particular position in a sequence were $\frac{1}{4}$, then the probability of a given 11 base sequence would be $(\frac{1}{4})^{11}$ (i.e. $\frac{1}{4} \times \frac{1}{4} \times \frac{1}{4} \dots$ repeated 11 times). This works out at one copy for every 4.2×10^6 bases (see Appendix D). So our haploid genome (3×10^9 bases) could contain around 1000 copies of our 11-mer. On average, each of these 11-mers would be separated by 4.2×10^6 bases from an identical 11-mer. Likewise, there could be 1000 copies of the inverse complementary 11-mer, similarly separated. The occurrence in the *c-fes* gene of the 11-mer, with its 11-mer inverse complement *close by*, can hardly be a chance event. Thus, it is tempting to suppose that selective pressures for inverted repeats in DNA explain Chargaff's PR2 for all orders of oligonucleotide up to a limit dependent on genome size.

On average, stem (stack) lengths tend to be around 4 base pairs, extending to 6 or 7 base pairs for very AT-rich or very GC-rich sequences [17]. As noted in Chapter 4, if we take equifrequency of complementary 4-mers as the result of selection for stems in nucleic acid structures, then PR2 for lower order k-mers would seem to be explained, and high order k-mers could arise from further mixing of intermediate and low order k-mers. In the case of the *c-fes* gene it would seem that such mixing with subsequent selection would account for the 11-mer.

There is also the possibility that some k-mers might ‘hitch-hike’ on others. For example, inclusion in a text of the phrase “important history,” implies selection for two words that the author brings together. Accompanying them are several shorter ‘words’ (imp, port, or, ant, etc.) that the author’s mind did not specifically select. Also, by virtue of their closeness, the word “this” has also come along for the ride. And it is not fussy about its vehicles – “important Hispanic” or “both islands” suit “this” fine!

In Figure 5.8 twelve 8-mers intervene between the 8-mer **TACGACGC** and its complementary partner (**GCGTCGTA**). These 8-mers (**ACGACGCG**, **CGACGCCG**, etc.) do not need a complementary partner in close proximity and need not themselves have arisen in response to selection. They could reflect probabilistic associations between the short oligonucleotides (e.g. 3-mer and 5-mer) which reside within them. These short oligonucleotides could have arisen by selection for structure, or might have hitch-hiked on higher order oligonucleotides that were selected. Thus, the frequencies of higher order oligonucleotides can sometimes be predicted from those of lower order oligonucleotides (using what are known as first order and second order Markov models). Indeed, using an equation based on tetra-nucleotide frequencies, in 1988 Arnold and his colleagues were able to predict for yeast the frequencies of higher order oligonucleotides [18].

By what mechanism would palindromic inverted repeats arise? In agreement with Nussinov [19], in 1984 Henry Krisch proposed that inverted repeats arose on a large scale through “rearrangements involving the frequent incorporation of inverted segments” [20]. But how segments of DNA would so obligingly invert themselves remains controversial [21]. It seems more likely that PR2 precision would usually have been achieved a base-pair at a time (see Chapter 7). High stability structures would grow from those of lower stability. What began as a ‘quasipalindrome’ would grow, in stepwise fashion, to a full palindrome [22].

Symmetrical Extrusions

The apparent symmetry of extrusion (Fig. 5.2) would predict that the folding energies of a sequence window (‘top strand’) and its inverse complementary sequence (‘bottom strand’) in a biological sequence would be similar. Indeed, this is the rule rather than the exception. Figure 5.9 shows fold values of top and bottom strands in a long DNA segment of a nematode worm [23]. Despite wide fluctuations, whichever parameter is chosen (FONS, FORS-D, FORS-M), the values (red and blue lines) are very close.

Peaks in FONS value as extreme as -100 kcal/mol (Fig. 5.9a) are due more to *the way bases are ordered* (see FORS-D values; Fig. 5.9b), than to base composition (see FORS-M values; Fig. 5.9c). Whereas FONS and FORS-M values are always negative, some FORS-D values are positive. Thus, sometimes base order (FORS-D) and base composition (FORS-M) work together to generate the total folding energy (FONS), and sometimes they work in opposition.

How do such profiles relate to the locations of genes and other genetic elements in a DNA segment? Does a peak correspond to a specific element or to a boundary

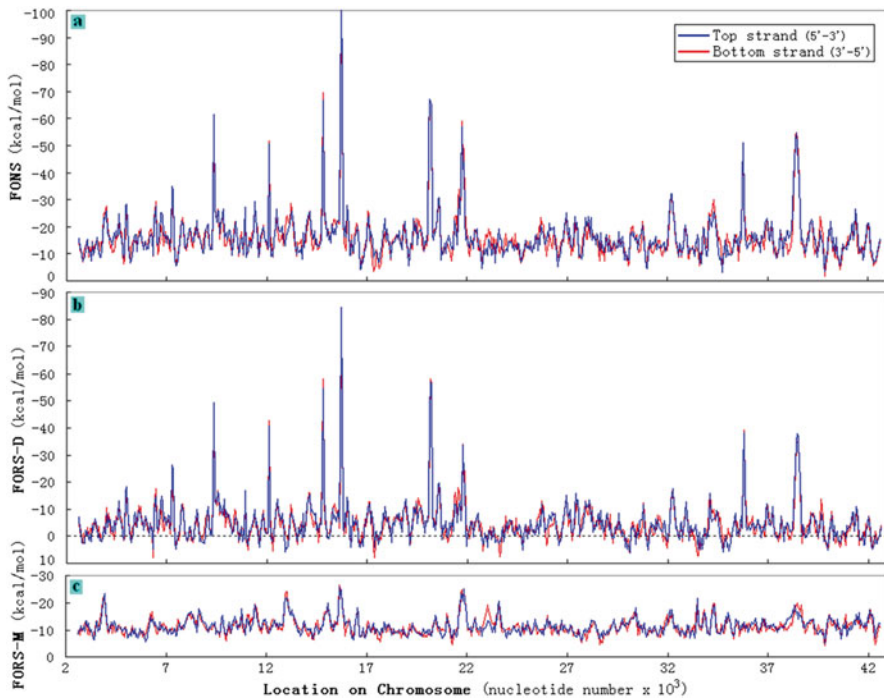


Fig. 5.9 The distribution of values for (a) FONS (total energy of folding), (b) FORS-D (the base order-dependent component), and (c) FORS-M (the base composition-dependent component), in the top strand (blue line) and bottom strand (red line) of a 40 kb segment of chromosome 1 of *C. elegans* (nucleotides 2500 to 42500). Folding energies were calculated for sequence windows (200 nt) moved in 50 nt steps [23]

between elements (see Chapter 13)? Furthermore, there are regions rich in bases that do not form standard Watson-Crick pairs. **A** and **C** in the *same* strand do not pair, so stem-loop extrusion is minimal. **G** and **T** in the *same* strand pair only weakly, but some extrusion is possible.

So extrusion symmetry is violated in regions rich in **A** and **C**, which have complements **G** and **T** on the *opposite* strand of the duplex. Such regions with the potential to asymmetrically extruding stem-loops are found at chromosome ends (telomeres) and at the attachments points (centromeres) for the fibers that separate chromosomes when cells divide (see Chapter 14).

DNA as Substrate

Proteins are macromolecules that often act as enzymes catalyzing changes in other molecules – their specific substrates – which may be micromolecules (e.g. glucose or fatty acids), or macromolecules like themselves (e.g. DNA).

In the course of evolution there are changes in proteins as a consequence of changes (mutations) in the corresponding regions of the DNAs that encode those proteins (Fig. 1.1). Substrates do not usually participate in this directly. For example, glucose is a preferred energy source in many organisms. As organisms evolve, specific enzymes may change and so improve the utilization of glucose. That's natural selection. *But glucose itself has no say in this.* It does not change to lighten, or impede, the task of the enzymes.

For some enzymes, however, DNA is a substrate. DNA both encodes the enzymes (locally in the regions of their genes), and is either their local or general substrate. We have seen here that, by virtue of changes in the order or composition of its bases, DNA can change a character such as stem-loop potential. Thus, it has a means of lightening or impeding the tasks of the enzymes that act on it. More than this, the *constancy* of glucose means that the enzymes of glucose metabolism can count on glucose being the *same* from generation to generation. The *limited constancy* of DNA means that many of the proteins (e.g. enzymes) that react with it are on an evolutionary treadmill. They must change as their substrate changes (see Chapter 12).

Summary

Transcription of DNA generates single-stranded RNAs that may operate by virtue of their structures (tRNAs, rRNAs), or by virtue of their encoding of proteins (mRNAs). The former consist of a small number of types which, while quantitatively abundant, are collectively encoded by a *small* part of genomes. The latter consist of a large number of types which, while each is quantitatively sparse, are collectively encoded by a *larger* part of genomes. All types of RNA fold into highly significant (non-random) 'stem-loop' structures. While the selective forces affecting tRNA and rRNA structures mainly relate to their roles in the cytoplasm, those affecting mRNA structures largely relate to the need for structure (stem-loop potential) of the genes from which they were transcribed. This need may conflict with the need to optimize the sequence of an encoded protein. Since non-genic regions of genomes also have potential for highly significant stem-loop structure, the potential is likely to be a *genome-wide* response to some selective force. In 1972 Harold White proposed that the strain of an unwinding force acting on regular duplex DNA would be relieved by extrusion of stem-loop structures. The potentials for such higher ordered structures in single-stranded nucleic acids may be calculated from the base-pairing energies of overlapping dinucleotides, which are fundamental units of nucleic acid structure. Contributions to the energetics of such structures decompose into base composition-dependent and base order-dependent components that may either support or oppose each other. The latter, most important, component is determined by subtracting the base composition-dependent component from the total folding energy. The base composition-dependent component is itself determined by shuffling and refolding a sequence several times – thus destroying the

base order-dependent component – and then taking the average folding energy of the resulting structures. Extrusion symmetry between the two strands of a DNA duplex is violated in regions rich in **A** and **C** (with complements **T** and **G**). Selection for nucleic acid structure provides an explanation for PR2.

References

1. Crick F (1974) Letter to H. G. Khorana, 28th June. Archives of National Library of Medicine, Washington, DC
2. Delbrück M (1949) *Transactions of the Connecticut Academy of Arts and Sciences* 38:173–190
3. Salser W (1970) Discussion. *Cold Spring Harbor Symposium in Quantitative Biology* 35:19
4. Ball LA (1972) Implications of secondary structure in messenger RNA. *Journal of Theoretical Biology* 36:313–320
5. White HB, Laux BE, Dennis D (1972) Messenger RNA structure: compatibility of hairpin loops with protein sequence. *Science* 175:1264–1266
6. Forsdyke DR (1998) An alternative way of thinking about stem-loops in DNA. *Journal of Theoretical Biology* 192:489–504; Einert TR, Netz RR (2011) Theory for RNA folding, stretching, and melting including loops and salt. *Biophysical Journal* 100:2745–2753 [An early model of Poland and Scheraga did not take into account the possibility that stem-loop extrusion from duplex DNA could be either *assisted* or *impeded* by the order of bases.]
7. Seffens W, Digby D (1999) mRNAs have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Research* 27:1578–1584
8. Forsdyke DR (2007) Calculation of folding energies of single-stranded nucleic acid sequences: conceptual issues. *Journal of Theoretical Biology* 248:745–753 [For many purposes, the control randomization of sequences can be performed at the 1-mer level. This retains base compositions, but not nearest-neighbor associations.]
9. Tinoco I, Uhlenbeck OC, Levine MD (1971) Estimating secondary structure in ribonucleic acids. *Nature* 230:362–367
10. Allawi HT, SantaLucia J (1997) Thermodynamics and NMR of internal GT mismatches in DNA. *Biochemistry* 36:10581–10589 [Since the Table 5.1 values for DNA were not then available, in my early studies I assumed that the corresponding values for RNA (later shown to closely approximated the DNA values), would suffice for my purposes.]
11. Zuker M (1990) Predicting optimal and suboptimal secondary structure for RNA. *Methods in Enzymology* 183:281–306
12. Hundley HA, Bass BL (2010) ADAR editing in double-stranded UTRs and other non-coding RNA sequences. *Trends in Biochemical Science* 35:377–383
13. Yura K, Sulaiman S, Hatta Y, Shionyu M, Go M (2009) RESOPS: A database for analyzing the correspondence of RNA editing sites to protein three-dimensional structures. *Plant and Cell Physiology* 50:1865–1873
14. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–258
15. Gu W, Zhou T, Wilke CO (2010) A universal trend of reduced mRNA stability near the translation initiation site in prokaryotes and eukaryotes. *PLOS Computational Biology* 6:e1000664
16. Zhang S-H, Huang Y-Z (2010) Limited contribution of stem-loop potential to symmetry of single-stranded genomic DNA. *Bioinformatics* 26:478–485
17. Fontana W, Konings DAM, Stadler PF, Schuster P (1994) Statistics of RNA secondary structures. *Biopolymers* 33:1389–1404
18. Arnold J, Cuticchia AJ, Newsome DA, Jennings WW, Ivarie R (1988) Mono- through hexanucleotide composition of the sense strand of yeast DNA: a Markov chain analysis. *Nucleic Acids Research* 16:7145–7157

19. Nussinov R (1982) Some indications for inverse DNA duplication. *Journal of Theoretical Biology* 95:783–791
20. Alff-Steinberger C (1984) Evidence for a coding pattern in the non-coding strand of the *E. coli* genome. *Nucleic Acids Research* 12:2235–2241
21. Kong S-G, Fan W-L, Chen H-D, Hsu Z-T, Zhou N, Zheng B, Lee H-C (2009) Inverse symmetry in complete genomes and whole-genome inverse duplication. *PLOS One* 4:e7553
22. Noort V van, Worning P, Ussery DW, Rosche WA, Sinden RR (2003) Strand misalignments lead to quasipalindrome correction. *Trends in Genetics* 19:365–369
23. Zhang C, Xu S, Wei J-F, Forsdyke DR (2008) Microsatellites that violate Chargaff's second parity rule have base order-dependent asymmetries in the folding energies of complementary DNA strands and may not drive speciation. *Journal of Theoretical Biology* 254:168–177

Chapter 6

Chargaff's Cluster Rule

Another consequence of our studies on deoxyribonucleic acids of animal and plant origin is the conclusion that at least 60% of the pyrimidines occur as oligonucleotide tracts containing three or more pyrimidines in a row; and a corresponding statement must, owing to the equality relationship [between the two strands], apply also to the purines.

Erwin Chargaff (1963) [1]

A nucleic acid sequence has three fundamental characteristics – its length, its base composition, and its sequence. If you know the sequence then you can calculate length and base composition with great precision. However, before the emergence of sequencing technologies in the 1970s there were biochemical methods that could provide values for length and base composition, albeit less precisely. Furthermore, Chargaff and his colleagues developed a method for evaluating a particular sequence characteristic – base clustering – that could distinguish DNA samples on the basis of sequence differences.

Base Clusters

It is a chemical property of DNA that under acid conditions it loses its purines (**R**), but retains its pyrimidines (**Y**). The loss of purines makes it easier to break the molecule chemically at the sites from which purines have been stripped. This leaves free segments (oligonucleotides) containing clusters of pyrimidines. The following two lines show a sequence with the pyrimidine clusters that would result if the sequence were broken in this way:

$$\begin{array}{ccccccc} \text{YYRYYYYYRRRRRYRYRRYRRRRYYYYYYYYRY} & & & & & & (6.1) \\ \text{YY YYYYY} & & \text{Y Y} & & \text{Y} & & \text{YYYYYYY YY} \end{array}$$

Pyrimidine clusters can be separated on the basis of size differences to generate a distinct profile for any particular DNA molecule. From sequence 6.1 we obtain 3 **Y**s, 2 **YY**s, 0 **YYY**s, 1 **YYYY**, 0 **YYYYY**s, 0 **YYYYYY**s and 1 **YYYYYYY**.

It is easy to dismiss such clusters as trivial. After all, in an average audience you may see several members of the same sex together in a row. Males and females are

often not evenly distributed. There are statistical tests for telling whether clustering is likely to be a chance phenomenon. Chargaff found that clustering of pyrimidines was more than expected on a chance basis. Furthermore, with knowledge that DNA was usually in duplex form – pyrimidines in one strand pairing with purines in the other – it was easy to infer that purine clusters must complement pyrimidine clusters. Thus the above single-strand would appear in duplex form as:



Chargaff's observation has since been found, like his first and second parity rules (PR1 and PR2), to apply to many genomes, and so can be considered a general feature of DNA.

Although the clusters themselves are not random, are they randomly distributed in the genome, or are they more prevalent in certain regions? Assuming that, like many things on this planet, random distribution (chaos) is the default state (see Chapter 15), what selective forces might have allowed clustered bases to persist? A cluster, by definition, locally violates PR2, but base clustering does not necessarily imply an extensive local conflict with PR2. A cluster of pyrimidines might be locally matched by purines (clustered or relatively unclustered), so the violation might not extend beyond several bases.. For example, a run of **T** residues might be accompanied by a corresponding number of dispersed **A** residues, so that $A\% \cong T\%$. On the other hand, if clusters were themselves clustered – pyrimidine clusters clustering with pyrimidine clusters, and purine clusters clustering with purine clusters – violations might be more extreme.

Clusters of Clusters

The cluster observation was extended by work from Wacław Szybalski's laboratory in the 1960s. He showed that clustering of clusters in microorganisms is most evident in transcriptionally active regions, and that the nature of the clustering of clusters (purine or pyrimidine) relates to transcription direction [2]. The 'top' strand of part of a DNA duplex that is transcribed contains pyrimidine clusters if transcription is to the left of the promoter (where the enzyme RNA polymerase initiates transcription), and purine clusters if transcription is to the right of the promoter (Fig. 6.1). Likewise, for human genes, Oliver Smithies and his colleagues showed that there are distinct local deviations from PR2, which again correlate with transcription direction [3]. Thus, clustering of clusters can result in local deviations from PR2 in favor of the clustered bases.

When transcription of mRNA is to the left, the top strand is the "mRNA template" strand (pyrimidine-rich), and the bottom strand is the "mRNA synonymous," or "coding," strand (purine-rich). When transcription of mRNA is to the right, the top strand is the coding strand (purine-rich), and the bottom strand is the template

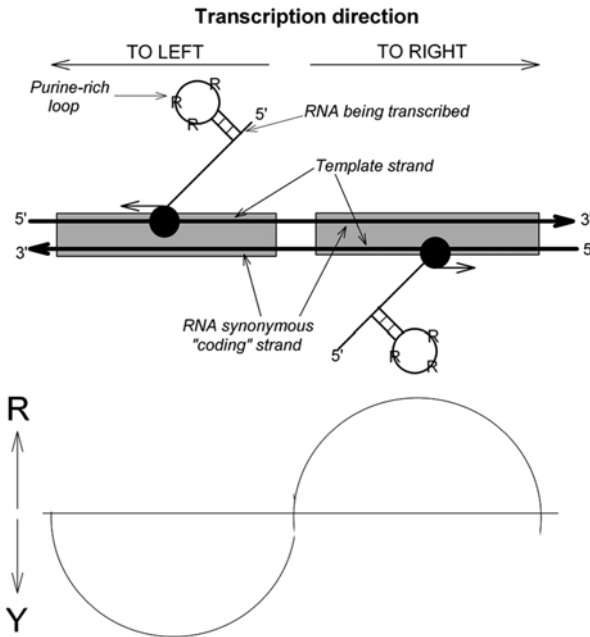


Fig. 6.1 Purine-loading of loops in RNAs is reflected in deviations from Chargaff's PR2 in the DNA regions from which the RNAs are transcribed. Heavy black horizontal arrows refer to the 'top' and 'bottom' strands of duplex DNA. Two genes (gray boxes) are transcribed from a central promoter (site of binding of RNA polymerases), one to the left, and one to the right. Black balls with thin arrows (indicating the direction of transcription) are RNA polymerases that have moved either left or right from the promoter transcribing linear RNA molecules. These are shown projecting at an angle from the horizontal. As the RNA molecules lengthen they assume stem-loop structures with **R**-rich loops (purine-loading). The graphical representation shows that in a region of leftward transcription the "Chargaff difference" for the top (template) strand favors pyrimidines ("Y-skew"). In a region of rightward transcription the "Chargaff difference" for the top (RNA-synonymous) strand favors purines ("R-skew"). If the RNA is mRNA, then RNA-synonymous strands are also "coding strands"

strand (pyrimidine-rich). It follows that, whether arising from a gene transcribed to the left or to the right, the transcribed RNAs themselves tend to be purine-rich ("purine-loaded"). As will be considered in Chapters 11 and 12, purine-loading is extreme in thermophiles (Fig. 6.2).

Smithies used non-overlapping 'windows' of a constant size (0.1 kilobase) to examine base composition in a variety of species. Bases were counted in a window and then the window was moved 0.1 kilobases along the sequence, and the count repeated. Thus, a base composition profile for a genomic region was constructed. When sequences of large genomes became available, it was found that window sizes between 0.5 kilobase and 1 kilobase (overlapping or non-overlapping) were optimal for determining the locations of genes (open reading frames; ORFs) and their transcriptional orientation in previously uncharted DNA [4–6]. In most cases,

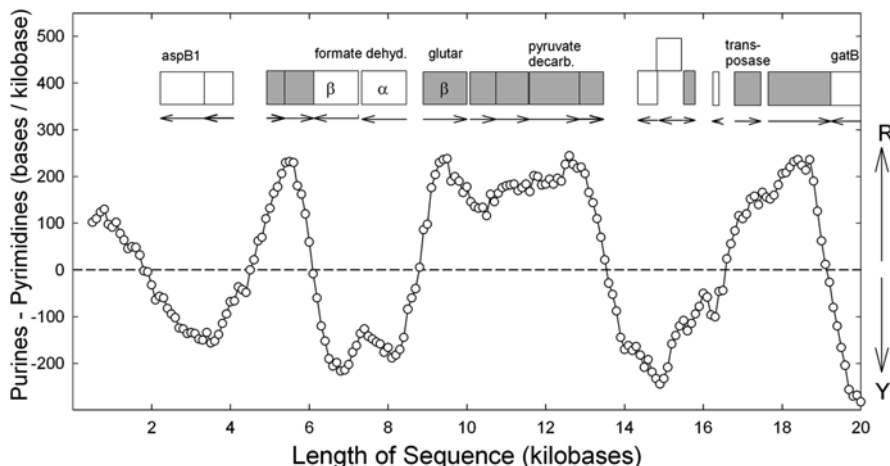


Fig. 6.2 Purine or pyrimidine excess in the top strand of the first 20 kb of the genomic sequence of the thermophile *Methanococcus jannaschii*, a bacterium-like microorganism classified as a species of archaea. A 1 kilobase sequence window was moved in steps of 0.1 kilobase and base compositions were determined in the window at each step. Total pyrimidines were subtracted from the total purines to give the recorded values (in bases/kilobase). When there are equal quantities of purines and pyrimidines, the difference (“Chargaff difference”) between purines and pyrimidines is zero. Thus, points above the dashed horizontal line correspond to purine excess, and points below the line correspond to pyrimidine excess. The locations of genes (open reading frames) are shown as boxes (white when transcribed to the left and gray when transcribed to the right). Transcription directions are also marked by horizontal arrows. Some genes are identified (see their abbreviated names above the boxes)

the transcripts were mRNAs that could be translated into proteins. However, the purine-loading phenomenon is a characteristic of all transcripts, whether they be templates for a protein product (mRNAs), or exist in their own right (e.g. tRNA, rRNAs and other RNAs with various functions) [7].

Polarity

The text you are now reading was written in the direction you are reading it – most helpful if you had wanted to read it as it was being written. Looking over my shoulder, you could have begun reading before the sentence was complete. But the nature of the present medium is such that there is a considerable delay between the time of composition and the time of reading. So the sentence could have been written backwards for all you could care. Or the middle words could have been written first, and then words at the ends added later. All you want is that the final product be in the left-to-right order with which you are familiar.

We write nucleic acid sequences from left to right beginning with the 5′ end. Because we have, by convention, chosen to write them this way, it does not neces-

sarily follow that nucleic acid sequences are assembled in biological systems in the same way. Yet, it is a chemical property of the nucleotide ‘building blocks’ from which nucleic acids are composed (Fig. 2.2), that the 5′ end of one is joined to the 3′ end of the nucleotide that has already been added to the growing chain (Figs. 2.3, 2.4). In biological systems nucleic acid synthesis has polarity, beginning at the 5′ end and proceeding to the 3′ end. This applies not only for replication – the copying of DNA information into DNA information, but also for transcription – the copying of DNA information into RNA information.

Furthermore, a nucleic acid sequence is ‘read’ *biologically* in the same direction as it has been ‘written’ *biologically*. Translation – the decoding of mRNA information to form a sequence of amino acids – proceeds from 5′ to 3′. Proteins have their own polarity, beginning and ending at what, for chemical reasons, are known as the amino-terminal, and carboxy-terminal ends, respectively. Proteins are ‘written’ with this polarity. But proteins fold into unique structures, so are not ‘read,’ in the same way as they are synthesized, by the molecules with which they interact three-dimensionally to generate the phenotype (Fig. 1.4). Similarly, nucleic acid transactions that require secondary and higher ordered structures (Chapter 5), are usually uninfluenced by the way a nucleic acid sequence was assembled within cells [8].

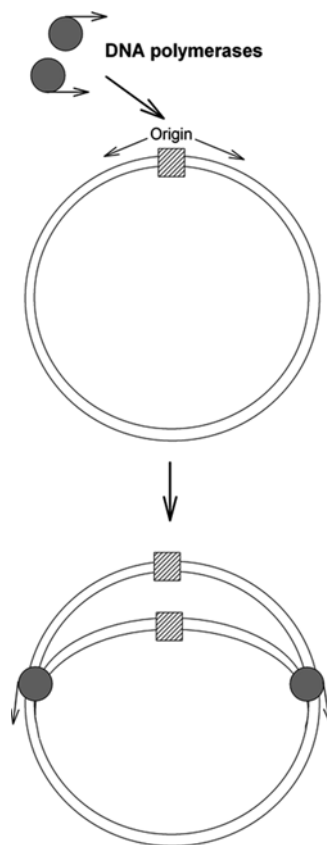
However, in some cases there are alternative, sometimes energetically equivalent, folding patterns. Folding can begin at the end of a protein or nucleic acid that is synthesized first, before the synthesis of the other end is complete. This early folding can influence the folding path adopted by a later synthesized part of a molecule. Thus, the order of synthesis of the primary sequence can influence the folding pattern, and hence the subsequent three-dimensional interactions, of a macromolecule. A macromolecule that is synthesized rapidly may not be able to follow slow folding paths. In this case there may be selection for codons that are translated slowly (codon bias; see Chapter 12), perhaps because of a deficiency (in quantity or quality) of the corresponding transfer RNAs. This slows synthesis, so facilitating slow folding processes. Even so, the selection of a particular folding pathway may need help from other molecules (‘molecular chaperones;’ see Part V).

Origin of Replication

The enzyme RNA polymerase initiates transcription at distinct sites in the genome (promoters) that have certain general features (e.g. often there is the base sequence **TATA**, known as the “TATA box”). The RNA transcript is then synthesized sequentially by the addition of ribonucleotides (nucleotides with ribose as the pentose sugar), one at a time, until appropriate termination signals are encountered (again, often there is a distinctive base sequence). When there is a need to transcribe a particular RNA at a particular time, the enzyme usually begins and ends transcription at distinct sites where transcription can be regulated.

However, since an entire chromosome (often circular in microorganisms) has to be replicated, in theory the enzyme DNA polymerase could initiate DNA replication

Fig. 6.3 Bi-directional replication of a circular chromosome from a single origin (box) by DNA polymerases (gray balls) that attach at the origin of replication. One DNA polymerase travels anticlockwise and the other DNA polymerase travels clockwise, so that at any moment there are two Y-shaped replication forks as shown in Figure 2.3. It is possible that there is more than one polymerase at each replication fork – at least one for the ‘leading’ strand and at least one for the ‘lagging’ strand (see Fig. 6.6.)



at any site by joining together deoxyribonucleotides (nucleotides with deoxyribose as the pentose sugar) according to the dictates of the parental DNA template. The enzyme could then proceed from that starting point round the circular genome until returning to the starting point.

Yet, presumably to facilitate regulation, there is usually one distinct origin of replication in a circular chromosome (say 12 o'clock), from which DNA replication proceeds leftwards and rightwards, bidirectionally. Thus, when the chromosome is half-replicated there are at least two DNA polymerases tracking in different directions (say at 9 o'clock and 3 o'clock; Fig. 6.3). They complete replication by meeting at a point in the circle approximately opposite the starting point (say, at 6 o'clock). This is the point of termination.

In a short linear chromosome, the origin is usually at, or near, one end. In a long linear chromosome there may be multiple origins so that segments of the chromosome may be replicated simultaneously by independent DNA polymerases, and the newly synthesized segments then joined together by other enzymes ("ligases").

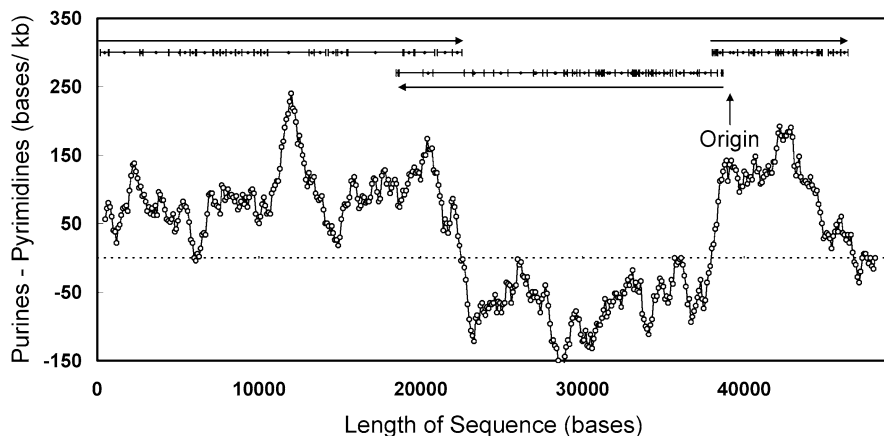


Fig. 6.4 Purine or pyrimidine excess in the top strand of the 48 kilobase genomic sequence of lambda phage. Base frequencies were determined in 1 kilobase windows, which were moved in steps of 0.1 kilobase. Since numbers of pyrimidines are subtracted from numbers of purines, positive values indicate top-strand purine-loading, and negative values indicate top-strand pyrimidine-loading. Horizontal lines with vertical cross-hatching show the location of genes. Long continuous horizontal arrows show the collective transcriptional orientation of groups of genes. The vertical arrow refers to the origin of replication within the interval 38686-39585 bases. From this interval replication proceeds bidirectionally to terminate around base 21000. Thus, the horizontal rightward arrow at top right, corresponding to rightward movement of a DNA polymerase around the circular chromosome, extends to the horizontal rightward arrow at top left

Origins and terminations of replication often have distinctive sequence characteristics, but these alone usually do not allow us to determine origins and terminations by inspection of the DNA sequence. However, Szybalski and his colleagues found for the circular lambda phage genome, that genes to the left of the origin of replication were transcribed to the left, and genes to the right of this origin were transcribed to the right [9]. Since leftward-transcribing and rightward transcribing genes are distinguishable by top strand pyrimidine-loading and purine-loading, respectively, it follows that the switch from pyrimidines to purines, and vice-versa, provides a method of determining origins of replication and sites of termination (Fig. 6.4).

The possible general nature of this became evident from studies in organisms considered higher in the evolutionary scale [3]. In the circular genome of Simian virus 40 (SV40), which infects monkeys, genes are transcribed to the left, to the left of the origin of replication, and here $C\% > G\%$; genes are transcribed to the right, to the right of the origin of replication, and here $G\% > C\%$. When further sequences became available the relationship to the origin of replication was found to be a feature of many microbial genomes [10]. Instead of determining the origin of replication by differences between purines and pyrimidines in moving sequence windows ("skew analysis"), an interesting variation is to allow the differences to summate ("cumulative skew"; Fig. 6.5).

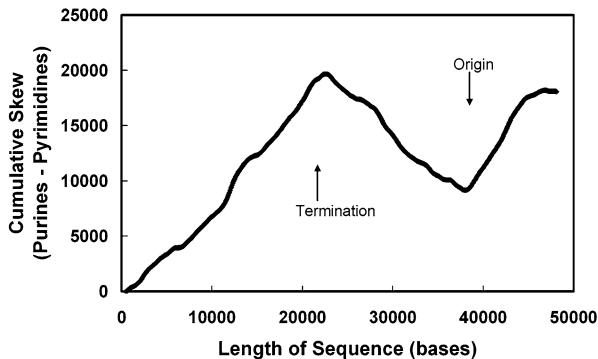


Fig. 6.5 Cumulative skew presentation of the lambda phage data shown in Figure 6.4. Instead of plotting directly, values for consecutive windows from left to right are progressively summed so that the curve slopes upwards in regions of rightward transcription and downwards in regions of leftward transcription. Origin and termination regions for DNA synthesis are indicated by inflections around 21000 bases (termination) and 39000 bases (origin)

Gene Movements

Yet, although often highly informative, for some genome sequences, skew plots (Chargaff difference plots) or cumulative skew plots (cumulative Chargaff difference plots) give ambiguous results. While reasons for this are uncertain, one factor causing ambiguity is the movement of genes, or groups of genes, from one part of a genome to another (transposition). This movement provided an opportunity to address the chicken-and-egg question as to whether replication direction can determine transcription direction, or vice-versa. If a rightward-transcribing gene to the right of the origin of replication were transposed to the left of the origin of replication, but kept its orientation so it was still rightward-transcribing, would it retain its purine-loading (i.e. would transcription overrule replication)? The answer seems to be that it begins to accept mutations from purines to pyrimidines (i.e. replication overrules transcription). The new mutations can sometimes change amino acids, so that a gene transposed in this way has increased opportunities for functional change [11].

Another example of gene movement is in mitochondria (see later), where there is a dramatic violation of PR2 with ‘top’ and ‘bottom’ strands differing greatly in base composition. When a mitochondrial DNA segment turns head-over-heels (inversion), the ‘top’ strand of the genes it contains moves from one strand to the other and, over evolutionary time, adopt the base composition of the new host strand [12]. Similarly, when a gene translocates from a chromosome to one with a different base composition, it adopts the base composition of the new host chromosome [13].

Why should there be a relationship between replication direction and transcription direction? One suggestion is that it is necessary not to interrupt transcription when replication is occurring in the same part of the genome. Perhaps there is less disruption if RNA polymerase and DNA polymerase move in the same direction on a strand of DNA that is acting as a template for synthesis of complementary anti-parallel strands of RNA or DNA, respectively [14, 15]. One possibility is that if DNA polymerase and RNA polymerase were not moving in the same direction, they might collide ‘head-on.’ A result of such a collision might be that DNA in the region could end up irreversibly knotted [16]. However, there is a caveat. Many of these studies were based on data from microorganisms. In 2009 Laurent Duret and his colleagues in France noted for human cells [17]:

Origins of replication are not randomly distributed as they display significant overlap with [transcriptional] promoter regions and CpG islands [see Chapter 18]. The hypothesis of a selective pressure to avoid frontal collisions between replication and transcription polymerases is not supported by experimental data as we find no evidence for gene orientation bias in the proximity of origins of replication.

Leading and Lagging Strands

The facts that (i) nucleic acid synthesis is directional, from 5′ to 3′, and (ii) the two strands of the DNA duplex are antiparallel (Fig. 2.4), have the interesting implication that one of the two child strands at the Y-shaped replication fork (Fig. 2.3) has to be synthesized in the opposite direction to the direction in which the replication point itself is moving. This is shown in Figure 6.6 in which the two strands are labeled as “leading” and “lagging,” depending on whether DNA synthesis is continuous (“leading strand”), or discontinuous, being synthesized in fragments that are later joined up (“lagging strand”). As drawn, this does not seem very elegant. Surely Nature could do better? Nature probably does. A three dimensional model with a multiplicity of mobile molecular actors might do her more justice.

In passing we should recall that the two parental strands in the stem of the Y are actually coiled around each other as a double helix (Fig. 2.1). Furthermore, when DNA synthesis is complete the two new duplexes go their separate ways. For child strand synthesis the parental strands have first to be uncoiled (unwound). If the entire chromosome is circular, this means that for every uncoiling step there has to be, first a transient strand breakage, then an uncoiling, and then a rejoining of the broken ends. Long linear DNA molecules are anchored in such a way that they usually also require this. The uncoiling is carried out by enzymes (topoisomerases) which, when a DNA duplex is not being replicated, tend to uncoil it slightly without strand breakage. This “negative supercoiling” places the helix under stress, a stress that can be partly relieved by the extrusion of stem-loops (see Fig. 5.2). Thus, stem-loop extrusion is normally favored energetically by the negatively supercoiled state of DNA duplexes.

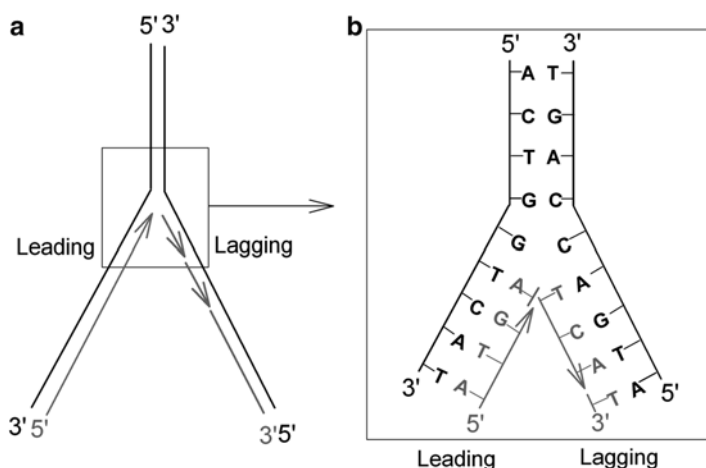


Fig. 6.6 Leading and lagging strand synthesis of DNA at the inverted Y-shaped point of DNA replication. Since parent (black) and child (gray) strands are antiparallel, on the left limb of the replication fork the “leading” child strand can be synthesized in the same direction as overall replication is occurring (accompanied by the separation of the two parental strands; see Fig. 2.3). On the right limb of the replication fork, the “lagging” strand is synthesized in a direction opposite to that in which overall replication is occurring. Discontinuous child strand DNA synthesis generates fragments that are joined up to make a continuous strand, so that overall replication occurs in the 3' to 5' direction. The box in the large scale figure without bases (a), is expanded in (b) to show bases and the initial fragmentary nature of the lagging strand

Purine loading

Chargaff's main interest in the base cluster phenomenon was that, prior to modern sequencing technology, it provided some measure of the uniqueness of the base order of a particular nucleic acid. Szybalski's main interest was that the clustering might play a role in the control of transcription. This implied an evolutionary selection pressure for clustering so that organisms with clusters would better control transcription than organisms that did not have clusters. However, following a better understanding of Chargaff's PR2 as a reflection of nucleic acid secondary structure (Chapters 4 and 5), a case could be made that a selection pressure for clustering had arisen at the level of complete RNA transcripts (i.e. after transcription had occurred), and was related to the secondary structure of individual RNA molecules.

An important implication of what may be called “Szybalski's transcription direction rule” is that RNAs, in general, tend to be purine-loaded (Fig. 6.7). This was suggested by Chargaff's work on the base composition of total RNA extracted from various species, but his data would have mainly reflected the compositions of the most abundant RNA type, the ribosomal RNAs (rRNAs) [18, 19].

Given that RNAs in general tend (i) to be loaded with runs of purines, and (ii) to have elaborate secondary structures, where in the structures would purine clusters be found? Since, for base-pairing in duplex stems, purine clusters must be matched with complementary pyrimidine clusters, and since pyrimidine clusters are scarce in

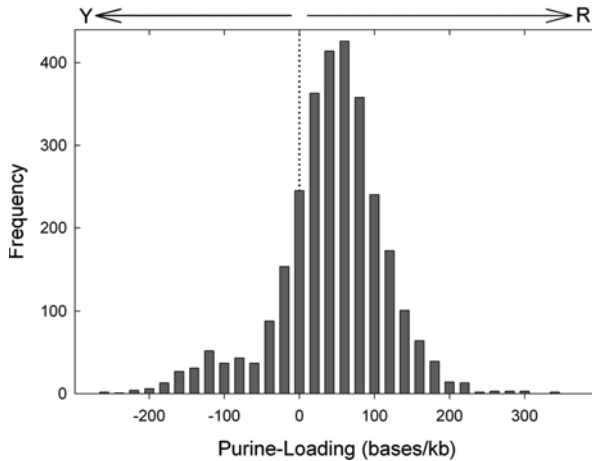


Fig. 6.7 Distribution of purine-loading among 3000 species. Positive X-axis values indicate purine-loading ($R > Y$). Negative values indicate pyrimidine-loading ($Y > R$). The purine-loading value for all human genes (excluding mitochondria) is 42 bases/kilobase, so that, on average, there are 42 more purines than pyrimidines for every kilobase of coding sequence. The shoulder with negative purine-loading values (i.e. pyrimidine-loading) corresponds mainly to mitochondria. Purine-loading of coding regions was calculated from codon usage tables for all species represented in the GenBank database (in 1999) by four or more genes. The purine-loading value (bases/kilobase) for a species is calculated as $1000[(G-C)/N] + 1000[(A-T)/N]$, where G , C , A , and T correspond to the numbers of individual bases, and N corresponds to the total number of bases in the codon usage table for that species. This measure of the purine-loading of RNAs disregards 5' and 3' non-coding sequences in mRNAs. When purine-loading is zero, $(A+G)\%$ is 50. This may be calculated from the formula: $\text{Purine-loading} = 20[(A+G)\%] - 1000$

RNAs, purine clusters should occupy the unpaired regions of RNA secondary structures – the loops. Indeed, this is where they are found in calculated mRNA and rRNA structures [5, 20, 21]. Why should it be advantageous for an RNA to ‘load’ its loops with purines?

A possible answer derives from the studies of Jun-ichi Tomizawa on the way a ‘sense’ RNA molecule and its complementary ‘antisense’ RNA molecule interact, prior to forming a double-strand duplex molecule (dsRNA) [22, 23]. In many organisms such sense-antisense interactions are important in regulation and immune responses (see Chapter 15). Indeed, small antisense RNAs and DNAs have long been used experimentally to bind to complementary mRNAs and block their translation [24].

Tomizawa found that sense sequences in RNA search out complementary antisense sequences through ‘kissing’ interactions between the tips of the loops of stem-loop structures. If Watson-Crick base-pairing between the loops is achieved, then the formation of a length of dsRNA becomes feasible (Fig. 6.8). Thus, if RNA loops were purine-loaded, ‘kissing’ interactions would decrease (because purines pair poorly with purines), and hence the probability of forming dsRNA would decrease. Why would *failure* to form dsRNA be of selective advantage? Or, in other words, why would formation of dsRNA be disadvantageous?

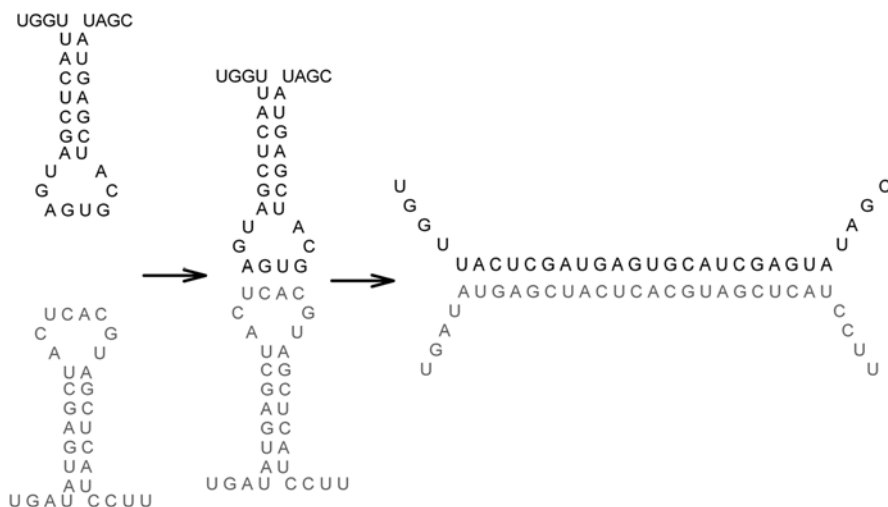


Fig. 6.8 A stem-loop ‘kissing’ model for the initiation of hybridization between two RNA species (black and gray base letters). The physico-chemical state of the crowded cytosol is conducive to intramolecular interactions, so that single-stranded RNAs in solution rapidly adopt energetically stable stem-loop structures. These structures must be disrupted if two free RNA species are to form a mutual duplex. Accordingly, there are two steps in duplex formation. The two RNAs first interact transiently at the tips of stem-loops (‘kissing’), mainly by conventional G-C and A-U base-pairing. If this pairing is sufficiently stable, it propagates progressively, and the stem-loop structures unfold to generate a duplex. In the present case, the pairing cannot proceed beyond 22 base-pairs (about two helical turns), since beyond this limit base-complementarity is lost. Here, as in Figure 5.2, all duplex structures are dynamic, undergoing vibratory partial disruptions (‘breathing’), which, as a critical temperature is approached, result in increasing degrees of strand separation. Thus, the reverse process probably occurs in one step (i.e. there might not be a distinct multistep path). For this, the duplex would have to be heated to a critical temperature when the two strands would suddenly separate and each would immediately adopt the stable stem-loop configurations shown at the left

One explanation arises from the physico-chemical state of the fluid part of the cytoplasm (the cytosol) where macromolecules are highly concentrated. The ‘crowded’ cytosol is an environment that, besides encouraging interactions between different parts of a molecule (*intramolecular* interactions), also encourages interactions between molecules (*intermolecular* interactions; see Chapter 16).

That RNA-RNA interactions would be affected can be deduced from first principles. Cytosolic characteristics such as salt concentration and hydrogen ion concentration (i.e. degree of acidity or alkalinity) should have been fine-tuned over evolutionary time to support the rapid and accurate interactions between codons in mRNA sequences and the anticodons of tRNA molecules, which are so vital for protein synthesis. As a by-product of this, interactions between RNA molecules *in general* would also have been facilitated. However, when so tied up in RNA-RNA interactions, mRNAs might be less available for association with ribosomes where protein synthesis occurs. Purine-loading should militate against mRNA-mRNA

interactions, while not compromising mRNA-tRNA interactions. And under some circumstances dsRNAs prove harmful to cells (see Chapter 15). To avoid this, some RNAs are edited shortly after their transcription [25] (see Chapter 18).

Whatever the merit of these arguments, which remains to be fully investigated [26], it would follow that RNAs might have achieved the same result by adopting pyrimidine-loading rather than purine-loading (since pyrimidines pair poorly with pyrimidines). *All* the RNAs in a given cytosol should adopt the *same* strategy (purine-loading or pyrimidine-loading). This can be compared with the decision made within a country for all vehicles to drive either on the left side or on the right side of the road. Yet, within the RNAs of a species there is variation in the extent of purine-loading. Indeed, some ‘maverick’ RNAs, such as those in mitochondria, show pyrimidine-loading.

Pyrimidine-Loading in Mitochondria

From the average distribution of loading (purine versus pyrimidine) among species (Fig. 6.7) it can be seen that, although most species purine-load, some pyrimidine-load. Mitochondria, the intracellular organelles concerned with energy generation, generally pyrimidine-load their RNAs. A cell may have its cytosolic RNAs ‘driving’ on the purine side of the road, but its mitochondria, which exist in the same cytosol (but have their own genomes, cytosols, and RNA types), ‘drive’ their RNAs on the pyrimidine side. Thus, some aspect of the intra-mitochondrial environment may have been particularly conducive to pyrimidine-loading of mitochondrial RNAs. On the other hand, during evolution the first mitochondrial ‘Eve’ may just have happened to have had its RNAs pyrimidine-loaded, and this feature was retained in all subsequent mitochondrial lineages.

Since transcription in mitochondria tends to be unidirectional (i.e. most genes are orientated together in one strand), there is usually a strong purine/pyrimidine asymmetry between the two strands of mitochondrial DNA (i.e. PR2 is strongly violated). The ‘top,’ pyrimidine-loaded, mRNA-synonymous, strand is likely to be less vulnerable to mutation than the ‘bottom,’ purine-loaded, strand, which, in its role as template, is likely to be more exposed to the alkaline intra-mitochondrial environment. Could this explain the evolution of pyrimidine-loading? Alkali favors deamination of cytosine (see Chapter 18), while militating against depurination. Thus, it would appear adaptively advantageous to assign pyrimidines to the less exposed top strand. A purine-rich template strand should be less vulnerable to mutation (or perhaps easier to repair), than a pyrimidine-rich template strand. But the PR2 violation suggests less single-strand secondary structure (fewer stem-loops) which, while perhaps facilitating transcription, would seem to make the strand more vulnerable to mutation. One solution is to protect the DNA strand by transiently pairing with its corresponding transcript, so forming an RNA-DNA duplex [27].

Mitochondria generate a highly mutagenic witch's brew of toxic oxygen derivatives (superoxides and hydrogen peroxide). It is perhaps for this reason that there are multiple 'back-up' copies of the genome (at least ten in each human mitochondrion), that may facilitate DNA repair (see Chapter 17). The maintenance of mitochondrial function is of fundamental importance, not only because of their role in energy generation (ATP synthesis), but also because mitochondria can release proteins that instigate "programmed cell death" (apoptosis; see Chapter 16). A failure to maintain the integrity of mitochondrial DNA would imperil mitochondrial function and hence the life of their host cell. It is perhaps for this reason (among others) that there are multiple 'back-up' mitochondria per cell, and even suggestive evidence that mitochondrial DNA, or perhaps intact mitochondria, can transfer between cells [28]. Indeed, just before their host cell begins DNA synthesis, mitochondria temporarily fuse together so that their genomes share one common mitochondrial cytosol [29, 30].

Since male gametes usually have little cytoplasm (Fig. 1.3), children usually inherit mitochondria from their mother, although paternal 'leakage' sometimes occurs. Mitochondria interact with the products of their host nuclear genome and incompatibilities between the nuclear genes of one parent and the mitochondrial genes contributed by the other, have the potential to assist the development of reproductive isolation and hence divergence into new species (see Part III) [31].

Mitochondrial genomes are attractive subjects for bioinformatic investigation. Generally they are quite small (16.6 kilobases in humans) and many sequences are available. Furthermore, wherever their host cells go, there must mitochondria go. Thus, if a host cell has adapted to grow in an extreme environment, it is likely that its mitochondria have similarly adapted. Mitochondrial DNA includes a short segment that shows a high degree of species specificity; hence, it has found use as a 'bar code' for rapid species identification (see Chapter 12).

In some species groups there is a reciprocal relationship between the degree of purine-loading and $(G+C)\%$. When purine-loading is high (i.e. pyrimidine-loading is low), $(G+C)\%$ is low. Conversely, when purine-loading is low (i.e. pyrimidine-loading is high), $(G+C)\%$ is high (see Chapter 12). This does not apply to mitochondria. Across all species examined, mitochondria have pyrimidine-loading (top strand) *and* low $(G+C)\%$ (both likely to decrease vulnerability to deamination of cytosine).

Chargaff drew attention to a "curious consequence" of his cluster rule – it placed a limit on the number of base arrangements (permutations) [1]:

At least 70% of the pyrimidines, and therefore 70% of the purines, occur in tracts of at least three of one kind, ... This has a curious consequence in that it restricts the number of possible permutations, though this figure is still enormous.

This was to prove of much significance (see complexity of DNA; Chapter 14). We should also note that, while most viruses are purine-loaded like their host cells, some are strongly pyrimidine-loaded, and these also have *high* $(G+C)\%$ values (e.g. HTLV1 see Chapter 11; EBV see Chapter 14). This will be further considered in Chapter 15 where the role of double-stranded RNA (dsRNA) as an intracellular alarm signal is reviewed.

Summary

The two DNA strands of a gene are the mRNA-synonymous “coding” (codon-containing) strand and the mRNA-template strand. Clusters of clusters of purines are general features of coding strands of DNA, which usually contain an excess of purines (“purine-loading”). Accordingly, since purines pair with pyrimidines, complementary clusters of clusters of pyrimidines are general features of the corresponding template strands, which usually contain an excess of pyrimidines (“pyrimidine-loading”). The clustering of clusters within a gene locally violates Chargaff’s PR2. This permits prediction of transcriptionally active regions in uncharted DNA and, in some cases, the origins and termination sites of DNA replication. RNA transcription and DNA replication appear to proceed optimally when the enzymes performing these functions (polymerases) are moving in the same direction along the DNA template. Since the stems of nucleic acid stem-loop structures require equal numbers of complementary bases, clusters tend to occupy loop regions. Here they would militate against the loop-loop ‘kissing’ interactions that precede the formation of nucleic acid duplexes. Thus, the loading of loops with bases that do not strongly base-pair with each other should decrease unproductive interactions between nucleic acids, so leaving them freer to engage in productive interactions, such as those between mRNAs and tRNAs that are critical for protein synthesis.

References

1. Chargaff E (1963) *Essays on Nucleic Acids*. Elsevier, Amsterdam, pp 124, 148
2. Szybalski W, Kubinski H, Sheldrick P (1966) Pyrimidine clusters on the transcribing strands of DNA and their possible role in the initiation of RNA synthesis. *Cold Spring Harbor Symposium in Quantitative Biology* 31:123–127
3. Smithies O, Engels WR, Devereux JR, Slightom JL, Shen S (1981) Base substitutions, length differences and DNA strand asymmetries in the human G-lambda and A-lambda fetal globin gene region. *Cell* 26:345–353
4. Saul A, Battistutta D (1988) Codon usage in *Plasmodium falciparum*. *Molecular Biochemistry and Parasitology* 27:35–42
5. Bell SJ, Forsdyke DR (1999) Deviations from Chargaff’s second parity rule correlate with direction of transcription. *Journal of Theoretical Biology* 197:63–76
6. Lao PJ, Forsdyke DR (2000) Thermophilic bacteria strictly obey Szybalski’s transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Research* 10:1–20
7. Schattner P (2002) Searching for RNA genes using base-composition statistics. *Nucleic Acids Research* 30:2076–2082
8. Mahen EM, Watson PY, Cottrell JW, Fedor MJ (2010) mRNA secondary structures fold sequentially but exchange rapidly in vivo. *PLOS Biology* 8:e1000307
9. Szybalski W, et al. (1969) Transcriptional controls in developing bacteriophages. *Journal of Cellular Physiology* 74, supplement 1:33–70
10. Frank AC, Lobry JR (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238:65–77

11. Tillier ERM, Collins RA (2000) Replication orientation affects the rate and direction of bacterial gene evolution. *Journal of Molecular Evolution* 51:459–463
12. Asakawa S, Kumazawa Y, Araki T, Himeno H, Miura K-I, Watanabe K (1991) Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. *Molecular Biology and Evolution* 32:511–520
13. Galtier N (2004) Recombination, GC-content and the human pseudoautosomal boundary paradox. *Trends in Genetics* 20:347–349
14. Brewer BJ (1988) When polymerases collide. *Cell* 53:679–686
15. French S (1992) Consequences of replication fork movement through transcription units *in vivo*. *Science* 258:1362–1365
16. Olavarrieta L, Hernández P, Krimer DB, Schwartzman JB (2002) DNA knotting caused by head-on collision of transcription and replication. *Journal of Molecular Biology* 322:1–6
17. Necseulea A, Guillet C, Cadoret J-C, Prioleau M-N, Duret L (2009) The relationship between DNA replication and human genome organization. *Molecular Biology & Evolution* 26:729–741
18. Chargaff E (1951) Structure and function of nucleic acids as cell constituents. *Federation Proceedings* 10:654–659
19. Elson D, Chargaff E (1955) Evidence of common regularities in the composition of pentose nucleic acids. *Biochimica Biophysica Acta* 17:367–376
20. Wang H-C, Hickey DA (2002) Evidence for strong selective constraints acting on the nucleotide composition of 16S ribosomal RNA genes. *Nucleic Acids Research* 30:2501–2507
21. Forsdyke DR, Bell SJ (2004) Purine-loading, stem-loops, and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. *Applied Bioinformatics* 3:3–8
22. Eguchi Y, Itoh T, Tomizawa J (1991) Antisense RNA. *Annual Reviews of Biochemistry* 60:631–652
23. Brunel C, Marquet R, Romby P, Ehresmann C (2002) RNA loop-loop interactions as dynamic functional motifs. *Biochimie* 84:925–944
24. Cristillo AD, Heximer SP, Forsdyke DR (1996) A 'stealth' approach to inhibition of lymphocyte activation by oligonucleotides complementary to the putative G0/G1 switch regulatory gene *G0S30/EGRI/ZFP6*. *DNA and Cell Biology* 15:561–570
25. Liddicoat BJ et al. (2015) RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science* 349:1115–1120
26. Paz A, Mester D, Baca I, Nevo E, Korol A (2004) Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *Proceedings of the National Academy of Sciences USA* 101:2951–2956
27. Holt IJ, Jacobs HT (2014) Unique features of DNA replication in mitochondria: a functional and evolutionary perspective. *BioEssays* 36:1024–1031 [The RNA-DNA duplex structure is known as an "R-loop."]
28. Spees JL, Olson SD, Whitney MJ, Prockop DJ (2006) Mitochondrial transfer between cells can rescue aerobic respiration. *Proceedings of the National Academy of Sciences USA* 103:1283–1288
29. Mitra K, Roysam B, Lin G, Lippincott-Schwartz J (2009) A hyperfused mitochondrial state achieved at G₁-S regulates cyclin E buildup and entry into S phase. *Proceedings of the National Academy of Sciences USA* 106:11960–11965
30. Chen H, Vermulst M, Wang YE, Chomyn A, Prolla TA, McCaffery JM, Chan DC (2010) Mitochondrial fusion is required for mtDNA stability in skeletal muscle and tolerance of mtDNA mutations. *Cell* 141:280–289
31. Louis EJ (2009) Origins of reproductive isolation. *Nature* 457:549–550

Part III
Variation and Speciation

Chapter 7

Mutation

Variation, whatever may be its cause, and however it may be limited, is the essential phenomenon of Evolution. Variation, in fact, is Evolution. The readiest way ... of solving the problem of Evolution is to study the facts of Variation.

William Bateson 1894 [1]

Members of a biological species vary. We acknowledge as a “variation” a character in an organism that is different from the character seen in most other members of its species. Hence the organism would be called a “variant.” The difference may be in *number*, or in *arrangement*, or in an individual *character itself*. Long before the nature of mutation at the DNA level was understood, Bateson drew attention to these distinctions.

Variations in the number or arrangement of body parts, he deemed “meristic” and “homeotic” variations, respectively. Variations in the substances of which the parts were composed, he deemed “substantive” variations. Thus, a variant organism with an extra finger (Fig. 7.1) would be a meristic variant (Greek: *meros* = part). A variant organism with rearrangement of body parts (e.g. an insect’s leg where its antenna should be, or a flower’s stamen where a petal should be) would be a homeotic variant (Greek: *homeo* = same). In Bateson’s words: “The essential phenomenon is not that there has merely been a change, but that something has been changed into the *likeness* of something else.” A variant organism with loss of normal coloration (e.g. an albino) would be a substantive variant.

It was difficult for Bateson to imagine that variations in number or arrangement could have originated in the same way as substantive variations. He could not separate the cause of a variation (now regarded as a ‘spontaneous’ or induced mutation in DNA) from the actual variation observed in the organism itself. He could not separate the mutational (variant-generating) *process* from the *result*.

We now know that mutations at the DNA level may sometimes produce no observable effect. Among mutations producing an observable effect, some happen to result in meristic variations, some happen to result in homeotic variations, and some happen to result in substantive variations. It is true that there are different types of mutation (changes in DNA), but a given type of mutation does not correspond to a particular type of observed variation. A single base change from, say, **T** to **G** (“micro-mutation”) can produce extreme changes in structure or physiology. A deletion of a large segment of DNA (“macromutation”) can result in no phenotypic change.

Fig. 7.1 A meristic variation (from William Bateson's 1894 text)



Note that, in the scientific literature, a mutation at the DNA level, and the observed result of that mutation (a “variant” or “mutant” organism) may both be called “a mutation.” So whether the word refers to cause or effect must be determined by context. By the same token, “micromutation” and “macromutation” may refer, as above, to an actual mutation at the DNA level, or to the extent of a change in phenotype. Again, context is important since Goldschmidt used these terms to distinguish fundamentally different mutational processes (Table 3.1). A mutation at the DNA level may be repaired, sometimes within seconds. In this case it will usually be both chemically and biologically undetectable (i.e. with current technology we can only hypothesize that such mutations occur). If not repaired, it will be an “accepted mutation.” This acceptance initially occurs within an individual and then, with breeding, may spread, sometimes by positive selection (see below), and sometimes randomly (e.g. ‘drift’ in small populations), to other members of a species.

The Genetic Code

We refer to *the* genetic code because, with a few exceptions (minor differences), all species use the same code. The code defines a relationship between amino acids and groups of nucleic acid bases (Table 7.1). Each position in a triplet codon can be occupied by one of four bases. There are four alternatives for the first position, four alternatives for the second position, and four alternatives for the third position.

Table 7.1 The genetic code as it would appear in mRNA. At the DNA level **T** would replace **U**. Amino acids are represented as three letter abbreviations. **UAA**, **UAG** and **UGA** are stop codons that terminate a growing chain of amino acids. The amino acids also have single-letter abbreviations, usually corresponding to their first letter. Exceptions are **Phe** (F), **Tyr** (Y), **Trp** (W), **Gln** (Q), **Arg** (R), **Asn** (N), **Lys** (K), **Asp** (D), and **Glu** (E)

		Second Position			
		U	C	A	G
First Position	U	UUU Phe	UCU Ser	UAU Tyr	UGU Cys
		UUC Phe	UCC Ser	UAC Tyr	UGC Cys
		UUA Leu	UCA Ser	UAA Stop	UGA Stop
		UUG Leu	UCG Ser	UAG Stop	UGG Trp
	C	CUU Leu	CCU Pro	CAU His	CGU Arg
		CUC Leu	CCC Pro	CAC His	CGC Arg
		CUA Leu	CCA Pro	CAA Gln	CGA Arg
		CUG Leu	CCG Pro	CAG Gln	CGG Arg
	A	AUU Ile	ACU Thr	AAU Asn	AGU Ser
		AUC Ile	ACC Thr	AAC Asn	AGC Ser
		AUA Ile	ACA Thr	AAA Lys	AGA Arg
		AUG Met	ACG Thr	AAG Lys	AGG Arg
	G	GUU Val	GCU Ala	GAU Asp	GGU Gly
		GUC Val	GCC Ala	GAC Asp	GGC Gly
		GUA Val	GCA Ala	GAA Glu	GGA Gly
		GUG Val	GCG Ala	GAG Glu	GGG Gly

Thus, there are $4 \times 4 \times 4 = 4^3 = 64$ possible triplets combinations (see Table 4.2). Sixty-one code for the twenty amino acids normally found in proteins. The remaining three provide alternative signals for stopping the synthesis of a protein.

Since there are more triplet codons than amino acids, and all codons are potentially employable, then the code is described as “degenerate.” There is not a reversible, one-to-one, relationship between codon and amino acid. If you know the codon you know the amino acid, but if you know the amino acid you do not know the codon. The only exceptions are **UGG**, which usually encodes only the amino acid tryptophan, and **AUG**, which usually encodes only the amino acid methionine. The latter codon is special since, in the appropriate context, **AUG** provides a signal for methionine to begin the sequence of a protein. Thus, **AUG** and the three chain-termination triplets (**UAA**, **UAG**, **UGA**) provide punctuation signals to a cell’s protein synthesizing machinery, so that it ‘knows’ when to start and stop translating the sequence of an mRNA template.

The degeneracy of the code mainly relates to the third position of codons. Thus, **GGN** is a generic codon for glycine. The third base (**N** = any base) can be any of the

four bases and the encoded amino acid will still be glycine. A mutation from **GGU** to **GGC** would be a non-amino-acid-changing mutation, otherwise known as a “synonymous” mutation (i.e. the mutation might occur, but the encoded amino acid would be the same, and so would have the same name; Greek: *syn* = same, *onyma* = name). Mutations in the first and second codon positions are usually amino-acid-changing mutations, otherwise referred to as “non-synonymous” mutations (i.e. the encoded amino acid would be different, and so would have a different name).

A nucleic acid sequence can display some flexibility in that it can change considerably with respect to third codon positions while still encoding the same protein (see sample sequences 3.1 and 3.2 in Chapter 3). However, this flexibility has its limits. Extreme shifts in base composition can affect the amino acid composition of a protein (see Chapter 12). Amino acids are divided into various groups on the basis of their physical and chemical properties (e.g. acidic, basic, hydrophilic, hydrophobic). For a position in a protein sequence that must have a basic amino acid, but is otherwise flexible, a genome rich in **A** and **T** (i.e. low **(G+C)%**) will tend to have the codons **AAA** or **AAG**, which encode the basic amino acid lysine; for the same position, a genome rich in **G** and **C** (i.e. high **(G+C)%**) will tend to have one of the **GC**-rich codons for the basic amino acid arginine (**CGN**).

One Base at a Time

Mutation is usually a random process and, within a short time frame, two mutational events are unlikely to occur in close proximity in DNA (see later). Thus, one base in a triplet codon is likely to change at a time. This one-step mutational process means that a mutation in the first codon position will change an amino acid to one vertically related to it in Table 7.1 (e.g. **CUU** to **GUU** shifts coding from leucine to valine). A mutation in the second codon position will change an amino acid to one horizontally related to it (e.g. **CUU** to **CAU** shifts coding from leucine to histidine). A mutation in the third codon position may change the amino acid only if there is more than one type of amino acid in the same box (e.g. **CAU** to **CAA** shifts coding from histidine to glutamine).

This implies that a particular amino acid is related, through mutation, to a certain subset of other amino acids; in the case of mutations affecting first and third codon positions, the amino acid replacements often show some similarity to the original amino acid in their physical and chemical properties. Thus, leucine and valine (related through first position mutation) are both classed as ‘water-fearing’ (hydrophobic) amino acids, which are likely to locate to the interior of a protein. Histidine and glutamine (related through third position mutation) are both ‘water-loving’ (hydrophilic) amino acids, which may locate to the surface of a protein. To some extent, the genetic code would appear to buffer the immediate effects of a mutation, in that an original amino acid would be replaced with an amino acid with somewhat similar properties. By this token, mutations involving second codon positions are the most likely to be lethal (e.g. leucine-histidine interchange).

Table 7.2 Distribution of non-amino-acid-changing (synonymous) and amino-acid-changing (non-synonymous) mutations among the three codon positions, as calculated from the genetic code. There are many technical problems in arriving at accurate values for the numbers of synonymous and amino-acid-changing differences between two similar nucleic acid sequences. Mathematically-inclined biologists spend much time on this. However, for many purposes it suffices to assume that amino-acid-changing mutations involve first and second codon positions, and synonymous mutations involve third codon positions (gray boxes)

Mutations	Codon Position			
	1	2	3	All
Total	183	183	183	549
Synonymous	8	0	126	134
Amino acid-changing	175	183	57	415

When the actual mutational changes undergone by proteins are examined, it is seen that many of them fit the one-step pattern, but there are also many amino acid changes (e.g. leucine to aspartic acid) that would require more than one step (e.g. **CUU** to **GUU** or **CAU**, and then either **GUU** to **GAU**, or **CAU** to **GAU**). Thus, the mutational process may take more than one path, and this may require that an organism with an intermediate mutant form (**GUU** or **CAU** in this example) be viable for a number of generations to provide time for a second mutation to occur. Each of the twenty amino acids can mutate to one of the nineteen others (total 380), which implies a total of 190 (380/2) possible reciprocal interchanges. Of these, 75 require only single base substitutions. The remaining 115 require substitutions in two or three bases.

Since there are three bases in a codon and each base can be substituted by one of three other bases, a given codon is related by one-step mutation to nine other codons. Given 61 codons for amino acids, there are 549 (i.e. 61×9) possible codon base substitutions (183 for each codon position). Of these 549, it can be determined by inspecting the genetic code (Table. 7.1) that 134 would be non-amino acid-changing (synonymous) and 415 would be amino acid-changing (non-synonymous). The relationships of different types of mutations to codon positions are shown in Table 7.2. Mutations may be from purine to purine, or from pyrimidine to pyrimidine (both “transition mutations”), or may be between purine and pyrimidine (“transversion mutations”). Thus **A** to **G** is a transition, and **A** to **T** is a transversion.

RNY Rule

A codon in mRNA, and its anticodon at the tip of a tRNA loop, transiently ‘kiss’ to form a short helical duplex (Fig. 5.1). The specificity of this interaction appears to depend on complementary base pairing of three or less of the bases in a triplet

codon. However, codon-anticodon interactions are more than triplet-triplet interactions. Neighboring bases are drawn in, so there is a quintuplet-quintuplet interaction. This is because a major component of the interaction energy between two complementary bases derives from the stacking interactions with their neighbors (see Chapter 5). Thus, the bases in mRNA on the 5' side of the first base of a codon and on the 3' side of the third base of a codon influence the recognition of the codon by the corresponding anticodon.

Whereas the triplet anticodon sequences in tRNAs vary according to the amino acid specificity of the tRNA, the flanking bases that can affect tRNA-mRNA interactions show regularities [2, 3]. On one side of tRNA anticodons there is a relatively invariant pyrimidine and on the other side there is a relatively invariant purine. Accordingly, it would seem advantageous for a codon in mRNA when interacting with its anticodon to be *preceded* in the mRNA by an **NNY** codon (so that the **Y** in the codon can pair with the invariant purine in the tRNA), and to be *followed* in the mRNA by an **RNN** codon (so that the **R** in the codon can pair with the invariant pyrimidine in the tRNA). Thus, if we express our central pairing codon as **NNN** (where **N** is any base), the sequence of codons should read ..., **NNY**, **NNN**, **RNN**, ... Since, codons are translated successively, it follows that codons that, while encoding amino acids correctly, can *also* obey an “**RNY** rule,” will be preferred. Thus, all central codons and their two flanking codons, should ideally read ..., **RNY**, **RNY**, **RNY**, ... [4]. The tendency of codons to adopt a generic **RNY** form is so evident that it can assist detection of protein-coding regions (ORFs) in uncharted sequences [5]. The mRNA codons in Figure 5.1 have been arranged to be consistent with the rule.

The sequence of a codon in mRNA also exists, at the level of the corresponding gene, in duplex DNA. Here, in the mRNA-synonymous (“coding”) DNA strand, **RNY** codons base-pair with their complements (**RNY** anticodons) in the other DNA (mRNA “template”) strand. Thus, the **RNY** pattern is present in *both* strands (Fig. 7.2).

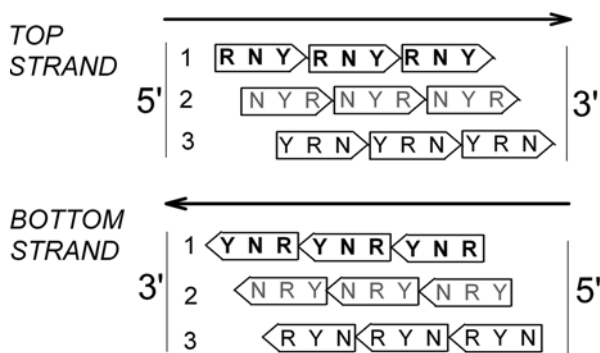


Fig. 7.2 The **RNY** pattern is in both DNA strands. Since the genetic code is a non-overlapping triplet code, there are three potential reading frames in the top strand of DNA, and three potential reading frames in the bottom strand of DNA. The pattern **RNY** can only correspond to one of the three frames and, since **R** always pairs with **Y** in duplex DNA, whichever frame **RNY** occupies in the top strand, it must also occupy in the bottom strand (here frame 1). Arrowed boxes indicate the polarity of coding triplets

In the case of amino acids whose codons contain **Y** in the first position or **R** in the third position, “**RNY**-pressure” on codons is in potential conflict with protein pressure (the pressure to insert a particular amino acid in a particular position in a protein). Also to be accommodated are **GC**-pressure (see Chapter 10) and **AG**-pressure (purine-loading pressure; Chapter 6). Indeed, in species under high **AG**-pressure third codon positions tend to acquire purines and the **RNY** rule may no longer apply (see Chapter 12).

Negative and Positive Selection

If reproductive success is impeded by a mutation, then selection of organisms with the mutation is negative. If reproductive success is promoted then the selection is positive. These are two, mutually exclusive, consequences of a mutation in a nucleic acid sequence. The extreme imperative of negative selection is: if you mutate, you die. Thus, the broad population of non-mutators remains and the few mutators die. The extreme imperative of positive selection is: if you do *not* mutate, you die. Thus, the broad population of non-mutators dies, and the few mutators flourish (i.e. there is a population ‘bottle-neck’ from which only the mutators emerge). Occupying the middle-ground between these extremes are ‘neutral’ mutations, and mutations that may lead to either weak positive or weak negative selection. In the latter cases there will, by definition, be effects on the number of descendents, but only in the long-term. Whatever the time frame, rather than being random, reproductive success tends to follow *some form of selection* (Table 3.1). Our focus here is natural selection.

Whether a base mutation will lead to negative or positive selection depends on the part of a gene-product (usually a protein) that it affects. A mutation affecting the active site of an enzyme will usually disturb enzyme function and this may impair the function of an organism to the extent that it will produce fewer descendents than will organisms without the mutation (i.e. its fitness to reproduce its kind is impeded). In the extreme, this is ensured by the death of the organism. On the other hand, a mutation affecting an antigen at the surface of a pathogen may allow it to evade the immune defenses of its host. A pathogen mutating in this way may produce more descendents than will pathogens without the mutation (i.e. its fitness to reproduce its kind is enhanced). In the extreme this is ensured by the death of pathogens that do not have the mutation.

Nucleic acid bases that are evolving slowly (i.e. they are conserved among related organisms) are likely to affect functions subject to negative selection (i.e. organisms with mutations in the bases are functionally impaired). Nucleic acid bases that are evolving very rapidly (i.e. they are not conserved among related organisms) are likely to affect functions subject to positive selection (i.e. organisms with mutations in the bases are functionally improved). So a determination of evolutionary rate can assist the distinction between bases under positive selection, and bases under negative selection, and hence can provide information on the function of a nucleic acid segment. For this, base differences between sequences can be calibrated against

some temporal scale (e.g. the period from evolutionary divergence of sequences from a common ancestral sequence to the present). Accumulation of a large number of differences (base substitutions) in a short time can be taken to indicate positive selection. However, accurate temporal calibration is difficult.

A more direct way to evaluate the mutational rate of a nucleic acid segment can be understood by considering a nucleic acid in the same way as we consider a language with different levels of information that may be subject to different forms of selection (Chapter 3). A public speaker conveys both a message (primary information) and an accent (secondary information). Imagine requiring each member of a group of competing speakers, one at a time, to read a given text to a large audience. The speakers are informed that they will be timed to determine the slowest, and that the audience will be polled to determine the most incomprehensible. The slowest and the most incomprehensible speakers will then be eliminated. The performance will be repeated, after which the slowest and most incomprehensible speakers will again be eliminated. This process will continue until a winner emerges.

Initially, each speaker relays both the text (primary information) and an individual accent (secondary information). However, under pressures both to speak rapidly and to be understood, speakers with more deviant accents are soon eliminated. Speakers are under strong pressure to eliminate personal idiosyncrasies of accent (i.e. to mutate their secondary information). The pressures for fast and coherent speech will progressively decrease the diversity of the secondary information among surviving group members. The final sound of the text will probably be the same for any large group of speakers exposed to the same large audience. Thus, the divergent accents of the initial multiplicity of speakers converge on a single accent to which the hearing of the average member of the audience is best attuned.

In a competition where there is no pressure for speed, idiosyncrasies of accent are less likely to interfere with comprehension (i.e. the diversity of secondary information is tolerated). The speakers could then also compete on the basis of emphasis, charm, body-language, or what you will (other forms of information).

Viewed from this perspective, it can be seen that a nucleic acid segment which is evolving rapidly with respect to its primary information (e.g. the sequence of a protein) may not be able to accommodate some of the other forms of information ("accents") that it might otherwise be carrying. These other forms, evolving leisurely at their own pace, include the ordering of bases to support stem-loop potential and purine-loading. *Thus, sequences under positive natural selection are also likely to be sequences where one or more forms of secondary information are impaired.* On this basis, whether natural selection is positive or negative can be evaluated without temporal calibration and *without a comparison with sequences in other genomes* [6].

There is another way to evaluate whether selection is negative or positive. The further apart two genes, the more likely they are to be separated by recombination. The frequency of recombination between them may be sufficiently high to allow separation of a nucleic acid segment containing a gene that is evolving *slowly* from other segments in the region. There can be gene shuffling (see Chapter 11). Thus, in one individual within a population a slowly evolving conserved gene may occur in association with a particular neighboring gene. In another individual the

conserved gene may occur in association with a variant form of the neighboring gene (especially likely if the neighboring gene belongs to a polymorphic gene family). Among members of the population there is sequence *diversity* in the region.

However, in the case of a nucleic acid segment containing a gene that is evolving *rapidly* (i.e. individuals containing it are being positively selected), there may not be time for separation from neighboring segments through recombination. Thus, neighboring genes of particular types will tend to remain attached (there is strong linkage). They will ‘hitchhike’ through the generations with a positively selected gene. Variant forms of neighboring genes will be lost from the population in the course of this ‘selective sweep.’ Consequently, among members of a population, variation in the region of a positively selected gene is decreased. Sequence diversity is lost locally, due to the imposition of a ‘bottle-neck’ on the population by the positively selected gene. This way of evaluating whether selection is negative or positive requires that sequences from different individuals be compared.

Neutral Mutations

The idea of neutral mutations was attractive to those seeking an internal frame of reference both for evaluating mutation rates, and for determining whether a mutation would impede or promote reproductive success. Mutations in third positions of codons often do not change the nature of the encoded amino acid, and hence do not change the corresponding protein and any characters that depend on that protein (Tables 7.1, 7.2). It was tempting to consider such synonymous mutations as “neutral” (see Chapter 5).

The most obvious advantage of the use of a particular codon, rather than a synonymous one, is that some codons can be translated more rapidly, or more accurately, than others. This is indeed of evolutionary significance for certain unicellular organisms where the rate of protein synthesis is critical (bacteria, yeast). Here synonymous mutations may not be neutral. But in many organisms the rate of protein synthesis is not critical. Here it was tempting to believe that mutations in third codon positions might be neutral.

To provide a relatively time-independent, internal, frame of reference for determining the form of natural selection (negative or positive), it was found convenient to compare in a gene the ratio of amino-acid-changing (non-synonymous) base substitution mutations (d_a) to non-amino-acid-changing (synonymous) base substitution mutations (d_s). For this, it was usual to align similar (orthologous) sequences from members of different species. At points where the bases were not the same, the differences were scored as either amino-acid-changing or synonymous. The procedure assumed that synonymous mutations were adaptively neutral, and hence reflected a ‘background’ rate of accepted mutation in a genomic segment of interest. Thus, the ratio (d_a/d_s) within a nucleic acid segment seemed capable of providing an index of the rate at which that segment was evolving. A high ratio suggested that the segment was under positive selection. A low ratio suggested that the segment was under negative selection.

Indeed, in many cases values for synonymous mutations (d_s) being significantly above zero, determinations of ratios agree with biological expectations. This suggests that third codon position mutations can indeed be neutral. Yet it is not unusual to find values for synonymous mutations *at or close to zero*. This is particularly apparent with certain genes of the malaria parasite, *Plasmodium falciparum*. Some interpreted this as revealing a recent population bottle-neck (i.e. a shrinking of the population, the surviving members of which become founders for a subsequent population expansion). Following the bottle-neck (i.e. a loss of population diversity) there would have been insufficient time for new synonymous mutations, to accumulate. Thus, at the extreme, existing species members could have been derived from one ‘Eve’ of recent origin.

However, others proposed that zero values for synonymous mutations could result from *high conservation* of bases that, while not determining the nature of an encoded amino acid, *do determine something else* – unspecified secondary information under some form of negative selection [7]. This would violate the neutral assumption, so that both the recent origin argument, and calculations based on ratios (d_a/d_s), would be invalid [8, 9]. Favoring this view, it was shown that, more than most genomes, that of the malaria parasite is sensitive to some of several non-classical selective factors, which affect third codon positions and collectively constitute the “genome phenotype.” Some examples of this will be given in Chapter 14.

Genome Phenotype

Sometimes an amino acid can be encoded from among as many as six possible synonymous codons (Table 7.1). In 1974 this prompted US geneticist Walter Fitch to remark that “the degeneracy of the genetic code provides an enormous plasticity to achieve [nucleic acid] secondary structure without sacrificing specificity of the message” [10]. Yet, sometimes this “plasticity” (flexibility) is insufficient, so that, with the exception of regions of genes under strong positive natural selection pressure (see Chapter 13), genomic pressures can ‘call the tune.’

When this happens, genomic pressures modify the amino acid sequence (non-synonymous codon changes), sometimes at the *expense* of protein structure and function. A protein has to adapt to the demands of the environment, but it also has to adapt to pressures that have derived, not from the conventional environment acting upon the conventional (‘classical’) phenotype, but from other environments, including what may be called the “internal reproductive environment” acting on the “genome phenotype.” Thus, Italian biochemist Giorgio Bernardi noted in 1986 [11] that:

The organismal phenotype comprises two components, the classical phenotype, corresponding to the ‘gene products,’ and a ‘genome phenotype’ which is defined by [base] compositional constraints.

To further illustrate this, let’s take a set of homologous (orthologous) genes that carry out essentially the same function in mice and rats – two species that diverged from a common ancestral species some ten million years ago. Although they have

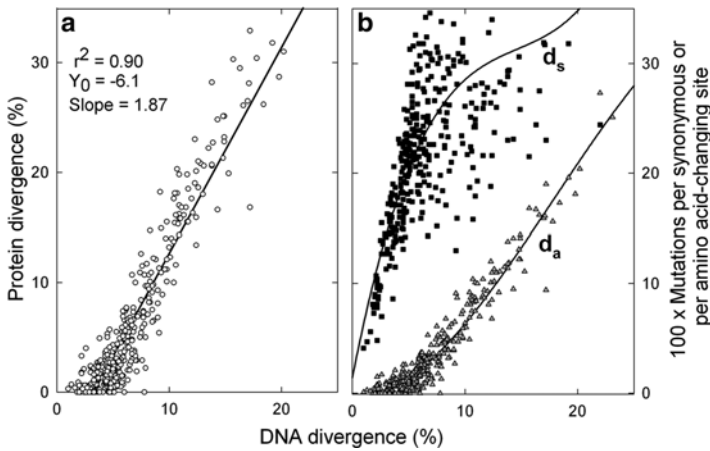


Fig. 7.3 Divergence between 363 genes of mouse and rat. For each gene, the relationship is shown between DNA divergence (bases %) and either (a) protein divergence (amino acids %), or (b) mutations per synonymous site (black squares) and per amino-acid-changing site (gray triangles). For convenience, in (b) each value is multiplied by 100 and points are fitted to third order regression lines. Note from Table 7.2 that more bases in the DNA of coding regions are available for amino-acid-changing mutation than for synonymous mutation. Thus, in (b) values are expressed as mutations relative to the number of sites (bases) available for that class of mutation. Since there are fewer sites available for synonymous mutation, when mutation rates are high (or over long time scales), synonymous sites are more likely to become saturated, and so are unable to accept more forward mutations. This might be responsible for the curve of d_s to the right in (b). Data for preparing this figure were kindly provided by Kenneth Wolfe

many similar characters, mice are anatomically and physiologically different from rats, and we would expect to find these differences reflected in differences in some of their proteins (i.e. differences in the classical phenotype). For the protein-coding regions of individual genes of the set, geneticists Kenneth Wolfe and Paul Sharp in Ireland compared base substitutions (DNA divergence) with amino acid substitutions (protein divergence) [12].

In Figure 7.3 each point corresponds to an individual gene and, since divergence increases with time, the X-axis (percentage DNA divergence) can be conceived as a time axis. Plotted against the percentage DNA divergence between mouse and rat for each gene are the corresponding protein divergences (Fig. 7.3a), and the corresponding amino-acid-changing (d_a) and synonymous (d_s) divergences (Fig. 7.3b). Genes differ dramatically in divergence. While some proteins (bottom left of Fig. 7.3a) have not diverged at all, the corresponding DNA sequences have diverged a little. Some proteins (top right of Fig. 7.3a) have diverged more than 20% and the corresponding DNA sequences are also highly diverged. The two divergences are linearly related. But there is an intercept on the X-axis that is significantly different from zero.

So it appears that for ten million years some proteins (bottom left of Fig. 7.3a) have remained unchanged (conserved). Presumably organisms with mutations in these proteins have died before they could reproduce their kind (i.e. they have been

negatively selected), and so no mutations are found in modern organisms. On the other hand, organisms with synonymous mutations that do not change the nature of an encoded amino acid would not have been counter-selected so severely. Thus, for a gene whose protein is unchanged there is a DNA divergence of about 5.4% (intercept on X-axis), implying that some synonymous mutations have been accepted. This is shown in Figure 7.3b where, as expected, the plot for amino acid-changing mutations (d_a) resembles that in Figure 7.3a, and the plot for synonymous mutations (d_s) extrapolates back close to zero.

The linear relationship shows that amino-acid-changing mutations, and synonymous mutations, are *correlated*. If a gene has zero or a few amino-acid-changing mutations (i.e. it is likely to have been under negative selection), then it will have few synonymous mutations and will display a low overall DNA divergence. If a gene has many amino-acid-changing mutations (i.e. it is likely to have been under some degree of positive selection), then it will also have many synonymous mutations and will display a high overall DNA divergence. This is a general observation and is not confined to the mouse-rat divergence [6]. Why should there be this correlation? Why is the synonymous DNA divergence so low in genes with low protein divergences, and so high in genes with high protein divergences?

It seems likely that, within the group of codons that encode a particular protein, the demands of the conventional phenotype and of the genome phenotype interrelate. An accepted amino-acid-changing mutation often cannot help change one or more aspects of the genome phenotype; it invokes (makes more acceptable) secondary compensatory mutations, mainly synonymous, to correct the change. By the same token, an accepted mutation that primarily changes the genome phenotype may happen to be amino-acid-changing and so may also change the conventional phenotype; it invokes further compensatory mutations, mainly non-synonymous, to correct the change.

For example, a primary mutation from the codon **ACA** to **AAA** may be to the selective advantage of an organism because it causes a lysine to be substituted for a threonine; but the mutation might also marginally affect nucleic acid conformation (stem-loop potential; Chapter 5), purine-loading (Chapter 6) and **(G+C)%** (see Chapter 10). A primary mutation from the codon **AGC** to **AGG**, while not affecting **(G+C)%**, causes an arginine to be substituted for serine, which may be of selective advantage to an organism; but the mutation might also marginally affect nucleic acid conformation, purine-loading, and **RNY**-pressure. A pressure to purine-load might change **ACA** to **AAA**, or **AGC** to **AGG**, thus conferring adaptive advantages; but the mutations could also cause proteins to encode lysine or arginine ‘against their wills’ (i.e. purine-loading would ‘call the tune’).

Most importantly, it should be noted that *genes with similar degrees of divergence do not necessarily colocalize in a genome*. A gene encoding a protein with zero divergence may be the neighbor of a gene encoding a protein with 20% divergence. Thus, *the borders of a region within which amino-acid-changing mutations and non-amino-acid-changing mutations can coadapt may approximate to the borders of an individual gene*. In this respect each gene is a ‘selfish’ autonomous entity in the sense implied by Hans Kalmus and George Williams (see Chapters 10, 11, and 12).

Compensatory Mutations

We should now qualify the above assertion that, mutation being essentially a random process, within a short time frame two mutational events are unlikely to occur in close proximity in DNA. If by “in close proximity” we included codons within a common gene, then the chances of a second mutational event being favored by a prior mutational event in the same gene are quite high. For example, the above amino-acid-changing mutation from **ACA** to **AAA** may have been accepted *primarily* because it changed a threonine to a lysine. As a by-product the purine content of the gene increased. This may have made it propitious for the subsequent acceptance of a synonymous purine-to-pyrimidine substitution mutation elsewhere in the gene, thus restoring the original degree of purine-loading without further changing amino acids. Currently, it is not clear how short the time frame between the two mutations would usually be, or how close within a gene the compensatory second mutation would usually be to the first.

Sometimes the second mutation will be a back-mutation at the same site. But a back-mutation is unlikely if the prior forward-mutation is lethal. Thus, the glycine codon **GGA** can forward-mutate to **TGA**, which is a stop codon and so is likely to be immediately lethal. When a **GGA** codon exits a genome in this way (because organisms with the mutation are no longer around to reproduce), it is not readily restored. For this reason the frequency in the population of **GGA** relative to the other three glycine codons should diminish. Although glycine in a certain position might be optimal for the function of a protein, the amino acids substituted for glycine when the latter three glycine codons mutate might be consistent with viability for a time sufficient for a back-mutation to glycine to occur, so restoring optimal protein function.

It can be seen in Figure 7.3b that as overall DNA divergence increases, the plot for protein divergence (d_a) increases linearly, whereas the plot for synonymous divergence (d_s) curves to the right. Thus, at high degrees of divergence, in many, but not all, genes the d_a/d_s ratio increases. This might be interpreted as revealing that many widely diverged genes have been under positive selection pressure (acting on the conventional phenotype). However, it should be noted that the ratio increase is due more to a *decline* in d_s than to an increase in d_a . So the plots show that, at high degrees of divergence, synonymous mutations tend to be constrained in some genes, but not others. Bases at synonymous sites may be conserved, or the sites may have become mutationally ‘saturated,’ so that no more forward mutations can be accepted. Thus, there are pressures on the genome phenotype that resist change (decrease d_s) despite concurrent pressures on the conventional phenotype that promote change (increase d_a).

RNA viruses, because they evolve rapidly (see Chapter 10), are very suitable for studies of this type. Isabelle Novella and her colleagues studied the *parallel evolution* of various virus strains as they adapt to new conditions, an adaptation presumed to be primarily due to the acceptance of amino-acid-changing mutations [13]. However, concurrent synonymous mutations did not occur randomly as classical

neutral theory would predict. The *same* synonymous mutations were accepted independently in *different* evolving strains. Thus, either synonymous mutations independently contribute to strain adaptations (e.g. affecting RNA structure), or they are secondary to (compensatory for) primary adaptations at amino acid-changing sites (or vice-versa). Further independent evidence for conservation of bases at synonymous sites derives from the uniformity of codon-bias among orthologous genes in vertebrates (“coincident codons”). This was attributed to selection acting coincidentally not only on proteins but also on RNA structure [14].

From all this it can be seen that the synonymous mutation rate (d_s), although often correlated with the amino-acid-changing mutation rate (d_a), has a life of its own. Its utility as a frame of reference for evaluating changes in d_a (i.e. the d_a/d_s ratio) may be somewhat limited [6–8]. We have here focused on regions encoding proteins. However, mutations that may be categorized as compensatory can also occur in close proximity in non-coding regions [15].

Were First Mutations Synonymous?

Let’s return to the mouse and rat (Fig. 7.3). Can we infer from the sequences of these two modern genomes the likely nature of the first mutations that distinguished the mouse and rat lineages ten million years ago? More simply, can we infer whether those mutations were likely to have been amino-acid-changing (non-synonymous), so implying that changes in the conventional (genic) phenotype had been instrumental in initiating the divergence process (i.e. classical Darwinian natural selection)? Alternatively, were the first mutations synonymous, so implying that changes in the genome phenotype had been instrumental in initiating the divergence process?

Given that each gene has a specific mutation rate, with a proportionate number of synonymous and amino acid-changing mutations (Fig. 7.3), the problem is set out in simple terms in Figure 7.4. Given a *sustained and uniform* proportionality within a gene, the plots should extrapolate back to zero. But if *at the outset* one type of mutation had made a disproportionate contribution, then there should not be extrapolation to zero. Hence, the nature of the mutations, that by their accumulation contributed most to the *initial* divergence from a common ancestor, is hinted at in Figure 7.3.

Viewing the X-axis as a time axis, and extrapolating back to zero divergence (zero time), it is seen that synonymous mutations (i.e. third codon position mutations) are the most likely candidates (Fig. 7.3b). Synonymous differences begin to appear at zero time, whereas amino-acid-changing differences (involving first and second codon positions) emerge later. In other words, the divergence of proteins (Fig. 7.3a) is likely to be *secondary* to whatever initiated the divergence process. From the bottom line (“known” in Figure 7.4) we see that gene A has high conservation that is likely to reflect some important phenotypic function (e.g. histone proteins). But, as indicated by the lack of synonymous mutations in A (relative to B and C), there has been a *proportionate* high conservation of functions related to the genome phenotype.

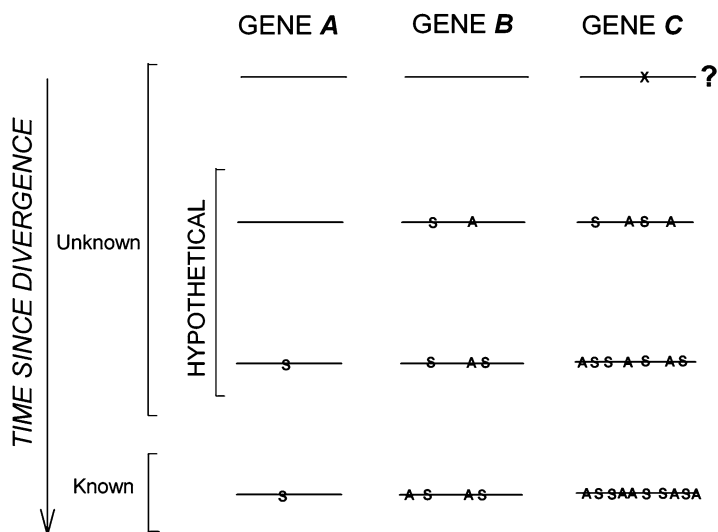


Fig. 7.4 Were the first mutations (X) associated with the mouse-rat divergence, synonymous (S) or amino-acid-changing (A; non-synonymous)? Genes A, B and C have low, intermediate, and high mutation rates, respectively. This is inferred from the number of known substitutions that distinguish the modern genes (bottom three lines). Under the assumption that rates of synonymous and amino-acid-changing mutations are positively correlated (i.e. proportionate), hypothetical intermediate substitution levels at progressive intervals following the divergence are shown, with proportionate numbers of synonymous (S) and amino-acid-changing (A) mutations in each gene

It may not be immediately apparent that the X-axes in Figure 7.3 (DNA divergence) can be viewed as conceptual time axes. The following metaphor may help. When at cruising speed the efficiency of a vehicle's fuel usage (kilometres/litre) is constant. However, when initially accelerating to that cruising speed, efficiency is less (i.e. less distance per unit of fuel). Thus, a plot of distance travelled (kilometres) against fuel usage (litres) would appear like the plots of protein divergence (d_a) against DNA divergence in Figures 7.3a and 7.3b. Around the time of the divergence from an ancestral species (i.e. 'acceleration' of a new species 'to cruising speed'), synonymous mutations (d_s) would have been *differentially* accepted ('low distance travelled relative to fuel usage'), and these would primarily account for the DNA divergence. After the divergence ('attainment of cruising speed'), amino-acid-changing mutations ('distance traveled') and synonymous mutations ('fuel usage'), would have been accepted *proportionately* (Fig. 7.4).

We should note that synonymous mutations in coding regions, being non-amino-acid-changing, are useful indicators of concomitant mutations that occur *outside* protein-encoding regions (i.e. both extragenically, and within introns; see Chapter 13). Synonymous, extragenic, and intronic mutation rates (all three) are often found to correlate closely. Although we focus here on synonymous mutations, all such non-protein-encoding mutations could collectively have driven the divergence process.

Phylogenetic Trees

The divergence into mouse and rat lineages (Fig. 7.3) can be represented as part of a continuously branching process from few ancestral life forms to the multiplicity of modern species. Researchers construct phylogenetic trees to model this branching into species over time. Organisms showing similar characters are placed close on the trees. Organisms showing different characters are placed at greater distances on the trees. Some characters turn out to be more useful for tree construction than others. Naturally, when nucleic acid sequences became available, researchers used these for tree construction. The fewer the mutations between them, the closer were considered two sequences, and hence the closer were considered the corresponding species. In similar fashion, one can construct language trees. The fewer the differences between two languages the closer they are placed on the language tree.

Various methods for multiple alignments of long strings of DNA bases were introduced. However, this approach was empirical and did not take into account the possibility that *some aspects of sequences, other than long strings of bases, might better reflect the underlying evolutionary process that caused the species to diverge in the first place*. For example, when spoken languages begin to diverge there is first a difference in accents. In this case, aligning long texts would necessitate the inclusion of much redundant information. Some measure of *accent difference* might *more accurately* display the relationship between the languages, because redundancies would be eliminated.

DNA sequences being comparable to spoken languages (Chapter 3), their base compositions can be regarded as their ‘dialects’ or ‘accents’ (see Chapter 11). Thus a DNA sequence, while having a distinct *base order* that might encode a protein, also has another property that, by virtue of the degeneracy of the genetic code, is to some extent independent – its *base composition*. Indeed, emerging pathogen species can be distinguished in clinical samples solely on the basis of base composition differences [16]. Furthermore, once the ‘noise’ of the primary base sequence has been eliminated, phylogenetic trees derived from 1-mer differences (differences in base composition) or, better still, from differences between sets of higher order k-mers (see Chapter 4), can be quite accurate [17–21].

The above deduction (Fig. 7.4) that synonymous mutations ‘fuelled’ the initial divergence between mice and rats, implies that, in a given period of evolutionary time, phylogenetic lineages with many branch points will tend to accumulate more synonymous mutations than lineages with fewer branch points. Indeed, phylogenetic studies of Chris Venditti and Mark Pagel in England support this [22, 23]. They note that a dependence of branching on synonymous mutations seems to *exclude natural selection* as a general initiator of species divergence:

Species do not so much ‘run in place’ as simply wait for the next sufficient cause of speciation to occur. Speciation is freed from the gradual tug of natural selection; there need not be an ‘arms race’ between the species and its environment, nor even any biotic effects. To the extent that this view is correct, the gradual genetic and other changes that normally accompany speciation may often be *consequential* to the event that promotes the reproductive isolation, rather than causal themselves.

Likewise, Blair Hedges and coworkers in the USA [24], from analyses of more than 50,000 eukaryotic species, infer “an uncoupling of speciation from adaptation,” and conclude that “adaptive change that characterizes the phenotypic diversity of life would appear to be a separate process from speciation.” This view, reminiscent of that of Romanes and Bateson (see Chapter 9), is not only incongruent with that of many modern evolutionists, but also with that of Charles Darwin.

Limits of Natural Selection

Despite Darwin’s belief that “external conditions ... cause mere variability” [25], the generation of DNA mutations (whether synonymous or non-synonymous) is essentially a spontaneous random process independent of any Darwinian environmental factor. Organisms (“mutants,” “variants”) with variant characters usually result from non-synonymous DNA base changes. Favorable or disfavorable variant characters favor or disfavor the organisms bearing them, so their descendents accordingly increase or decrease in the population by natural selection. For present purposes we include sexual selection (e.g. peahens prefer the most ornate peacocks) as a form of natural selection (i.e. peahens and peacocks constitute a mutual *external* reproductive environment). Whatever the selection, be it positive or negative, it is the classical *natural* selection of Darwin, who used the term to contrast with the *artificial* selection carried out by man in his role as breeder or horticulturalist [26].

Yet, within the constraints of a species, the ability of Nature to ‘invent’ new phenotypes is limited. As discussed in Chapter 3, Shakespeare could explore the theme of personal ambition in a castle in Scotland in one play, but other themes required that new plays be written. For real novelty – and the enhanced reproductive success this implies – just as Shakespeare’s works diverge, so must Nature’s. The only figure in Darwin’s great book, *On the Origin of Species by Means of Natural Selection*, was a phylogenetic tree. Here he distinguished linear and branching (divergent) evolution. His title suggested that both might depend on natural selection, although he cautioned [27]: “I am convinced that natural selection has been the most important, but not the exclusive, means of modification.”

Natural selection is selection acting on gene-products (e.g. proteins) – a process that can be referred to as “genetic” selection. There is little disagreement that natural genetic selection plays an important role in the process of linear, non-branching, evolution (“species survival”). However, throughout the nineteenth and twentieth centuries its role in the branching that creates two linear growing arms from one stem (“species arrival”) was contentious. Was species arrival due to business-as-usual natural selection, or was there a need for an underlying ‘genetic revolution,’ which Goldschmidt equated with a ‘pattern change’ that would affect the *internal* reproductive environment (non-genetic selection)? Overcoming constraints in human thought sometimes requires powerful paradigm shifts. Perhaps, in similar fashion, constrained species need to be periodically jolted out of their complacency?

If so, what could be the ultimate molecular basis of this genetic kick-in-the-pants paradigm shift? Chargaff provided an early clue [28]:

It is quite possible that our biochemical thinking will have to acquire an entire new dimension. ... If there is any meaning in the concept of biological information, the existence of the regularities of the sort pointed out by me ... may furnish us the only chemical clue to systems through which information is preserved or – and this is more important – transmitted [preserved through the generations]. Compositional regularities in themselves are not sufficient; but they will, I believe, have to be taken into account in any eventual solution of this problem.

It has been suggested here that synonymous mutations and the base composition ‘accent’ of DNA are highly relevant to the problem. Since a biological species is defined as such because its members are reproductively isolated from the members of another species – namely, that their DNAs do not normally recombine – then base composition (1-mer) *dissimilarity* and inability to recombine would seem to be positively associated. Conversely, base composition *similarity* may associate with ability to recombine. Since 1-mers may result from evolutionary pressures at higher k-mer levels (Chapter

), we can hence infer a positive association between k-mer dissimilarity and inability to recombine. Phylogenetic studies reveal growing recognition of the role of reproductive isolation (inability to recombine) in species divergence, not as a consequence of, but as the ‘spark’ which initiates, that divergence [22–24].

Summary

Anatomical or physiological variations that are inherited are due to inherited changes (mutations) in base sequences of DNA. Mutations that change genes can affect the *conventional phenotype* resulting in linear within-species evolution, often under the influence of natural selection (species survival). These changes associate with amino-acid-changing (non-synonymous) mutations in the first or second bases of triplet codons. DNA mutations can also result in changes in the *genome phenotype*. These changes associate with synonymous (non-amino-acid-changing) mutations, usually in the third bases of codons. Each gene in a genome has distinctive rates of acceptance of amino-acid-changing and synonymous mutations, which are positively correlated. A gene with few amino-acid-changing mutations also has few synonymous mutations. A gene with many amino-acid-changing mutations also has many synonymous mutations. Two genes may be closely located but differ greatly in their mutation acceptance rates. Thus, each gene is an independent mutational entity. Synonymous mutations, and correlated mutations in regions that do not encode amino acids, may be important for changing the ‘pattern’ of a genome, so sparking the onset of branching evolution (species arrival). By eliminating redundant information, oligonucleotide frequency patterns should provide rapid and more sensitive indices of species differences than direct sequence comparisons.

References

1. Bateson W (1894) *Materials for the Study of Variation Treated with Especial Regard for Discontinuity in the Origin of Species*. Macmillan, London, pp 6, 85, 573
2. Bossi L, Roth JR (1980) The influence of codon context on genetic code translation. *Nature* 286:123–127
3. Simonson AB, Lake JA (2002) The transorientation hypothesis for codon recognition during protein synthesis. *Nature* 416:281–285
4. Eigen M, Schuster P (1978) The hypercycle. A principle of natural self-organization. Part C. The realistic hypercycle. *Naturwissenschaften* 65:341–369
5. Shepherd JCW (1981) Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. *Proceedings of the National Academy of Sciences USA* 78:1596–1600
6. Forsdyke DR (2006) Positive Darwinian selection. Does the comparative method rule? *Journal of Biological Systems* 15, 95–108
7. Forsdyke DR (2002) Selective pressures that decrease synonymous mutations in *Plasmodium falciparum*. *Trends in Parasitology* 18:411–418
8. Nugal CF, Wolf JBW, Kaj I (2013) Why time matters: codon evolution and the temporal dynamics of dN/dS. *Molecular Biology & Evolution* 31:212–231
9. Reis M dos, Yang Z (2013) Why do more divergent sequences produce smaller nonsynonymous/synonymous rate ratios in pairwise comparisons? *Genetics* 195:195–204
10. Fitch WM (1974) The large extent of putative secondary nucleic acid structure in random nucleotide sequences or amino acid-derived messenger-RNA. *Journal of Molecular Evolution* 3:279–291
11. Bernardi G, Bernardi G (1986) Compositional constraints and genome evolution. *Journal of Molecular Evolution* 24:1–11 [Orgel, Crick and Sapienza (Selfish DNA (1980) *Nature* 288:645–647) had pointed to the need to distinguish the “organismal phenotype” from the “intragenomic phenotype.”]
12. Wolfe KH, Sharp PM (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *Journal of Molecular Evolution* 37:441–456 [DNA divergence has a relationship to time, but some think that the relationship to increasing organismal complexity facilitates better analytic insight (see Huang S (2015) New thoughts on an old riddle. What determines genetic diversity within and between species. *arXiv* 1510.05918).]
13. Novella IA, Zarate S, Metzgar D, Ebendick-Corpus, BE (2004) Positive selection of synonymous mutations in vesicular stomatitis virus. *Journal of Molecular Biology* 342:1415–1421
14. Bains W (1987) Codon distribution in vertebrate genes may be used to predict gene length. *Journal of Molecular Biology* 197:379–388 [A study of “coincident codons”.]
15. Terekhanova NV, Bazykin GA, Neverov A, Kondrashov AS, Seplyarskiy VB (2013) Prevalence of multinucleotide replacements in evolution of primates and *Drosophila*. *Molecular Biology and Evolution* 30:1315–1325
16. Sempath R *et al.* (2007) Global surveillance of emerging influenza virus genotypes by mass spectrometry. *PLOS One*: 2(5): e489
17. Blaisdell BE (1986) A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences USA* 83:5155–5159
18. Bronson EC, Anderson JN (1994) Nucleotide composition as a driving force in the evolution of retroviruses. *Journal of Molecular Evolution* 38:506–532
19. Qi J, Wang B, Hao BL (2004) Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *Journal of Molecular Evolution* 58:1–11
20. Richter M, Rosselló-Móra R (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proceedings of the National Academy of Sciences USA* 106:19126–19131.
21. Yang K, Zhang L (2008) Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction. *Nucleic Acids Res.* 36(5):e33

22. Venditti C, Pagel M (2009) Speciation as an active force in promoting genetic evolution. *Trends in Ecology & Evolution* 25:14–20
23. Venditti C, Meade A, Pagel M (2010) Phylogenies reveal new interpretation of speciation and the Red Queen. *Nature* 463:349–352
24. Hedges SB, Marin J, Suleski M, Paymer M, Kumar S (2015) Tree of life reveals clock-like speciation and diversification. *Molecular Biology & Evolution* 32:835–845
25. Darwin C (1856) Letter to J. D. Hooker. In: Darwin F (ed) *Life and Letters of Charles Darwin*. Volume 1. Appleton, New York (1887) p 445
26. Forsdyke DR (2001) *The Origin of Species, Revisited. A Victorian who Anticipated Modern Developments in Darwin's Theory*. McGill-Queen's University Press, Montreal
27. Darwin C (1872) *On the Origin of Species by Means of Natural Selection*. 6th Edition. John Murray, London [By the 6th edition, Darwin was more inclined to the view that some acquired characters were inherited (Lamarckism). This is perhaps why he then questioned the sufficiency of natural selection as an explanation for evolutionary advance.]
28. Chargaff E (1963) *Essays on Nucleic Acids*. Elsevier, Amsterdam, p 95

Chapter 8

Species Survival and Arrival

We have all grown up in the greatest confidence that we all knew ... what Darwin meant. I am very tired of having some excessively loosely expressed truism, such as that 'all defective deer must be devoured by tigers', put forward as 'the ordinary Darwinian argument'.

Ronald Fisher (1930) [1]

If it had inherited a mutated gene that impaired its ability to run, then, relative to its companions, a deer would be defective and so subject to selective devourment by predators. This is Tennyson's "Nature, red in tooth and claw." In 1862, when reading Darwin's *Origin of Species by Means of Natural Selection* while breeding sheep in New Zealand, the 27 year old Butler appeared to accept "the ordinary Darwinian argument," albeit with some important caveats [2]:

That the immense differences between the camel and the pig should have come about in six thousand years is not believable; but in six million years it is not incredible... . Once grant the principles, once grant that competition is a great power in Nature, and that changes in circumstances and habits produce a tendency to variation in the offspring (no matter how slight that variation may be), and unless you can define the possible limit of such variation during an infinite series of generations, unless you can show that there is a limit, and that Darwin's theory oversteps it, you have no right to object to his conclusions.

Yet in 1930 the great statistician Ronald Fisher began *The Genetical Theory of Natural Selection* with: "Natural Selection is not Evolution" (Fisher's capitalization) [3]. In so doing he drew attention to a question that had much concerned Darwin. To what extent can evolution be due to natural selection? Seeming to look to causes beyond natural selection, Fisher placed two quotations in a prominent position at the opening of his first chapter. The first was from an 1856 letter by Darwin that downplayed the evolutionary role of "external conditions" (e.g. tigers) [4]:

At present ... my conclusion is that external conditions do *extremely* little, except in causing mere variability. This mere variability (causing the child *not* closely to resemble its parent) I look at as *very* different from the formation of a marked variety or new species. [Darwin's italics]

The second was from a paper by William Bateson [5]:

As Samuel Butler so truly said: "To me it seems that the 'Origin of Variation,' whatever it is, is the only true 'Origin of Species'."

However, despite the caution implied by these quotations, Fisher set out to show, using mathematical procedures referred to by biologist Ernst Mayr as “bean bag genetics,” that natural selection was, in essence, evolution. Natural selection was responsible not only for species *survival*, but also for species *arrival*. Thus, Fisher fanned the classical Darwinist flame long tended by Alfred Wallace and August Weismann, that had flickered but never faded under the attacks of Romanes and Bateson in the late nineteenth and early twentieth centuries. With the support of contemporary geneticists in the United States (Dobzhansky and Muller), and despite the efforts of Goldschmidt, Fisher’s genic viewpoint came to reign for much of the twentieth century.

To better appreciate the arguments of the contending schools we here consider the roles of natural selection and reproductive isolation in species formation. For this we must first review some basic genetics as set out by the man who, in the early 1860s, while Butler was breeding sheep and reading Darwin in New Zealand, was breeding peas and reading Darwin in Moravia – Gregor Mendel.

Blending or Discrete Inheritance?

Do the characters of two parents blend in their children, or do children inherit certain discrete characters from one parent and other discrete characters from the other? If blending were the rule, then variation in a population would be constantly decreasing unless opposed by some differentiating force (i.e. fresh mutations). We can imagine the mixing of black and white paints giving a uniform gray from which the originals would not be recoverable. On the other hand, if characters were inherited discretely, those that were transmitted would be undiluted and fully recoverable. Variation within a population would tend to remain constant depending on rates of gain and loss, through mutation and selection. In 1857, two years before the publication of his *Origin of Species*, Darwin favored such discrete inheritance [6]:

I have lately been inclined to speculate, ... that propagation by true fertilization will turn out to be a sort of mixture, and not true fusion, of two distinct individuals, or rather of innumerable individuals, as each parent has its parents and ancestors. I can understand in no other view the way in which crossed forms go back to so large an extent to ancestral forms.

Consistent with this, his friend the botanist Joseph Hooker noted in 1860 [7]:

A very able and careful experimenter, M. Naudin, performed a series of experiments at the Jardin des Plantes at Paris, in order to discover the duration of the progeny of fertile hybrids. He concludes that the fertile posterity of hybrids disappears, to give place to the pure typical form of one or other parent.

Shortly thereafter the suspicions of Darwin, Hooker and Naudin were confirmed by Mendel’s breeding studies with peas [8]. When Mendel crossed pure tall plants with pure small plants all the hybrid offspring were tall. The small character had disappeared. When these offspring were crossed among themselves, the small character re-emerged in the next generation. On average, there were three tall plants to one small plant. The tall character was considered “dominant” to the small character

(“recessive”). At that time (1865) chromosomes had not been discovered and Mendel’s work remained largely unnoticed until 1900. Someone who may have paid attention was Romanes. He cited Mendel in 1881 and in 1894 described his own unpublished Naudin-like results obtained over several years by crossing varieties of rats or rabbits [9]:

One may breed scores and hundreds of first crosses between different varieties, and never get ... any resemblance to one side of the house. Yet, if the younger are subsequently crossed *inter se* (i.e. brothers and sisters, or first crossings) the crossed parentage at once repeats itself. *Ergo*, even if the pups which are to be born appear to give a negative result, keep them to breed from with one another.

With hindsight the ability of characters to retain their identities through generations of breeding is rather obvious. Your father was male and your mother was female, yet you have inherited this characteristic – sex – quite discretely. You are either male or female. You are not a mixture (a hermaphrodite). Despite their mothers being female, boys usually do, in time, become as masculine as their fathers. Despite their fathers being male, girls usually do, in time, become as female as their mothers. Darwin made this point in a letter [10]:

My dear Wallace ... I do not think you understand what I mean by the non-blending of certain varieties. ... An instance will explain. I crossed the Painted Lady and Purple sweet peas, which are very differently coloured varieties, and got, even out of the same pod, both varieties perfect but none intermediate. Something of this kind I should think must occur at least with your butterflies and the three forms of *Lythrum*; tho’ these cases are in appearance so wonderful, I do not know that they are really more so than every female in the world producing distinct male and female offspring.

In his usual colorful way, Butler made the same point in 1877 [11]:

The memory [stored information in the embryo] being a fusion of its recollections of what it did, both when it was its father and also when it was its mother, the offspring should have a very common tendency to resemble both parents, the one in some respects, the other in others; but it might also hardly less commonly show a more marked recollection of the one history than of the other, thus more distinctly resembling one parent than the other. And this is what we observe to be the case ... so far as that the offspring is almost invariably either male or female, and generally resembles rather the one parent than the other.

Mendelism

The height of a pea plant mainly depends on the length of its stem that, in turn, depends on one gene, two types of which (T and t) exist among members of the species. Individual diploid plants with two copies of each chromosome can be either pure TT , or Tt , or pure tt . The pure forms are referred to as homozygotes (Greek: “*homo*” = same). Each Tt individual will have inherited a T gene from one parent and a t gene from the other. So a particular chromosome inherited from one parent will contain T at a particular ‘slot’ or ‘address’ (locus). The same chromosome inherited from the other parent will contain t at the same ‘slot’ or ‘address’ (locus). Thus T and t are defined as allelic genes, or “alleles,” which are mutually exclusive

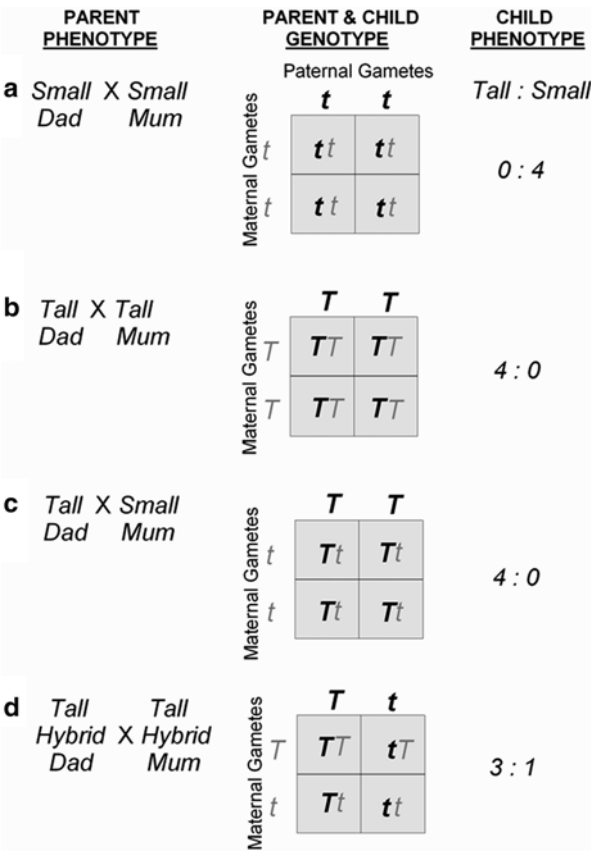


Fig. 8.1 Distribution of allelic genes for tallness (T) and smallness (t). Grey boxes show the genotypes of four children that would be produced, on average, if four maternal gametes and four paternal gametes met. When male and female homozygotes are of the same kind, all children are of the parental type, as in (a) and (b). When male and female homozygotes are of different kinds as in (c), all children are heterozygotes (Tt). If T is dominant to t , then all heterozygotes appear tall. If a male heterozygote is crossed with a female heterozygote as in (d), gametes of type T and t are produced in equal quantities by both parents and, of four children, on average three appear tall (TT , Tt , tT) and one appears small (tt). These are the famous Mendelian ratios

alternatives with respect to location on a chromosome. If two Tt parents mate, then, with random mixing of chromosomes during production of gametes, the probability of a TT child is 25%, of a child with both T and t (i.e. Tt or tT) is 50%, and of a tt child is 25% (Fig. 8.1). Tt and tT ("reciprocal crosses") are equal heterozygotes (Greek: "hetero" = different); the order relates to the convention of placing the genome of paternal origin first.

So a t -bearing chromosome borne by a haploid spermatozoon from one Tt parent might happen to meet a t -bearing chromosome in a haploid ovum from another Tt parent. We should note that, on a chance basis ('random drift'), this might happen successively. So *all* their children (diploid) would be tt . The T genes the parents had

been carrying through the generations would have been lost. Of course, fresh T genes could be recreated by fresh mutations in future generations. If the difference between the T and t alleles were a single base in DNA causing a difference in an amino acid at a particular position in an encoded protein, then t would retain the potential to mutate to T , so allowing generation of a new Tt individual. Such mutations being rare (sporadic), only one allele would mutate at a time. Thus tt would have to first mutate to tT , which could then mutate to TT . So pure tt parents, already most unlikely to produce Tt offspring by mutation, would be even more unlikely to produce TT offspring by mutation.

Multigenic Characters

All this implies a one-to-one relationship between an allelic gene and a character. Since genes are usually inherited intact, there is non-blending inheritance. However, in some circumstances there may be blending inheritance. When neither character is dominant there may be an interaction between the products of two parental alleles to generate a new character. Sometime this will be a mean between the characters of the two parents. Yet the new character is not sustained since, when children bearing it are old enough to be crossed among themselves then, in Romanes' words "the crossed parentage at once repeats itself" in their offspring.

Many characters are multigenic, resulting from the interactions of the products of several *non-allelic* genes. For example, one parent may contribute genes A , B , C and D , which positively determine height in an animal – each gene controlling the length of a particular part. The other parent may contribute the corresponding allelic genes A' , B' , C' and D' , which negatively determine height. If genes for tallness (A , B , C and D) and smallness (A' , B' , C' and D') are otherwise neutral in their effects, then members of the population will have all possible combinations, and their children will inherit these alleles in all possible combinations. Depending on the various dominance-recessive relationships, among a large number of children there could then be a smooth (binomial) distribution of heights between the rare extremes of 'pure' tall ($ABCD$) and of 'pure' small ($A'B'C'D'$). In 1894, six years before Mendel's work reemerged, Bateson had some feeling for this [12]:

An error more far-reaching and mischievous is the doctrine that a new variation must immediately be swamped, ... This doctrine would come with more force were it the fact that as a matter of experience the offspring of two varieties, or of variety and normal, does usually present a mean between the characters of its parents. Such a simple result is, I believe, rarely found among the facts of inheritance. ... Though it is obvious that there are certain classes of characters that are often evenly blended in the offspring, it is equally certain that there are others that are not.

In all this we are still able only to quote case against case. No one has [yet] found general expressions differentiating the two classes of characters, nor is it easy to point to any one character that uniformly follows either rule. Perhaps we are justified in the impression that among characters which blend or may blend evenly, are especially certain quantitative characters, such as stature; while characters depending on differences of number, or upon qualitative differences, as for example colour, are more often alternative in their inheritance.

Following the discovery of Mendel's work, Bateson and his colleague Rebecca Saunders noted in 1902 [13]:

In, for example, the stature of a ... race of man, a typically continuous character, there must certainly be on any hypothesis more than one pair of possible allelomorphs [alleles]. There may be many such pairs, but we have no certainty that the number of such pairs, and consequently of the different kinds of gametes, are altogether *unlimited* ... If there were even so few as, say, four or five pairs of possible allelomorphs, the various homo- and heterozygous combinations might, on seriation, give so near an approach to a continuous curve [i.e. blending], that the purity of the elements would be unsuspected [Bateson's italics].

Mendel himself had reached this conclusion when studying beans (*Phaseolus*), rather than his favorite plant, the pea (*Pisum*):

Apart from the fact that from the union of a white and a purple-red (crimson) colouring a whole series of colours results, from purple to pale violet and white, it is also striking that among 31 flowering plants only one received the recessive character of the white colour, while in *Pisum* this occurs on the average in every fourth plant. ... Even these enigmatic results, however, might probably be explained by the law governing *Pisum* if we might assume that the colour of the flowers and seeds of *Ph. multiflorus* is a combination of two or more entirely independent colours, which individually act like any other constant character in the plant.

The 'pure' combinations $ABCD$ and $A'B'C'D'$ could correspond to the "pure lines" of Johannsen (see Chapter 3). If bred with their own kind (e.g. an $ABCD$ male gamete meets an $ABCD$ female gamete), they would not be diluted or 'swamped' (i.e. they would retain their positions near the surface of Jenkin's sphere; see Fig. 3.2). However, if crossed (e.g. an $ABCD$ male gamete meets an $A'B'C'D'$ female gamete), then their rare distinctive characters (extremes of tallness and smallness) might not be seen among the children. There would be a variety of distinctive blends (e.g. $A'BCD$, $AB'CD'$, etc.), and consequently a regression towards the mean would be observed (Fig. 3.3).

Reproductive Isolation

When one line diverges into two lines the conditions are propitious for the 'arrival' of new species. The divergence may be cut short either by blending (Fig. 8.2b), or by natural selection, which brings about reproductive isolation in the most effectual way possible – by eliminating potential breeding partners (Fig. 8.2c). In these cases, although evolution is essentially linear, a new species may eventually emerge. However, if not cut short, branching into two lines can continue (Fig. 8.2d). Members of one line are unable to reproduce (produce fertile offspring) with members of the other and, on this basis, can be defined as members of independent species.

For reproductive purposes, members of a species *assort* – they can be said to positively assort with their own kind (members of the same species), and to negatively assort with other kinds (members of other species). Thus, members of one species cannot swap genes or other segments of their DNAs with members of another species. So a character encoded by a gene tends to remain within its species as it moves 'vertically' through the generations (Fig. 1.1). But if its species does not

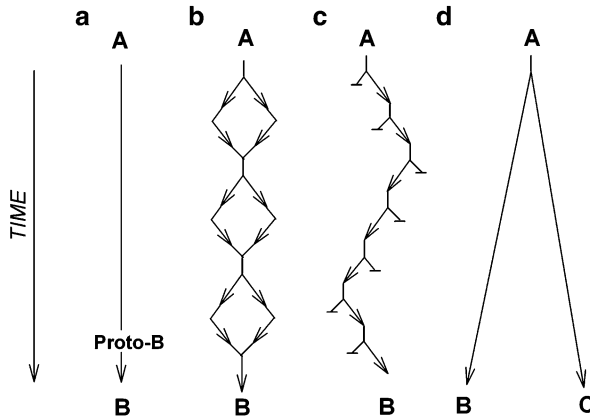


Fig. 8.2 Differences between linear (a, b, c) and branching (d) evolution. (a) Change from A, by way of various intermediate forms (“proto-B”), to arrive at form B. Examples of this are (b) and (c). In (b) there is branching of varieties of *equal fitness*. But the varieties are not reproductively isolated and their characters tend to blend in their offspring. Branching cannot be sustained. There can be differentiation so that B comes to differ from A, but evolution is basically linear. In (c) there is *unequal fitness* of varieties, so one is more vulnerable to elimination (negative natural selection). Reproductive isolation occurs when the more vulnerable is eliminated. In this context, natural selection acts as a reproductive barrier. Again, there can be differentiation so that B comes to differ from A, but still evolution is basically linear. (d) Varieties of *equal fitness* can branch from each other when there is an appropriate reproductive barrier

fare well, from a gene’s perspective all is not lost. Although threatened by adverse natural selection, a gene can ‘hope’ to escape to a new species if its host species diverges. And, in certain circumstances [14], it can sometimes transfer ‘horizontally’ to another species. Here it is likely to find itself in an ‘alien’ environment to which, under special conditions, it may adapt.

Cycling Through the Generations

The problem of species arrival through branching evolution boils down to the problem of identifying the barriers that initiate and sustain reproductive isolation [15]. We can represent evolution as an arrow of time departing from ancestral forms of species A to arrive at contemporary forms of species B (Fig. 8.2a). Yet, until a species becomes extinct, reproduction of its members is actually a cyclic process that, in ‘higher’ forms, can be depicted as proceeding from gametes, to zygotes, to adults with gonads, where new gametes are produced to continue the cycle (Fig. 8.3).

When branching occurs, the single cycle gives rise to two cycles that cannot merge to reform a single cycle. The cycles are isolated. For this, the original cycle must be interrupted *at some point*. The problem of the origin of species is primarily that of determining what form the interruption usually takes, and at what point in the cycle it usually occurs. There are three fundamental barriers (interruption points) to

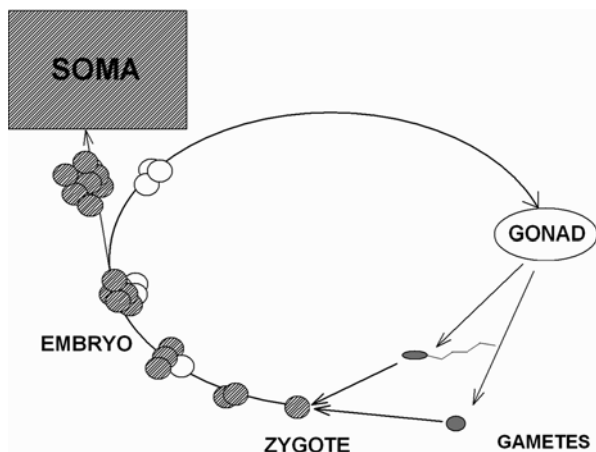


Fig. 8.3 The eternal cycling of the germ-line. The mortal soma provides support to the gonad, but is discarded in each generation. For simplicity male and female forms are here considered as one (as in hermaphrodites)

cycle operation, one operating to prevent gametes meeting – the transmission (prezygotic) barrier – and two operating after the gametes have met to form the single-celled zygote (fertilized egg). This then divides repeatedly to produce a multicellular embryo and eventually a mature, gonad-bearing, adult. The Montagues and the Capulets had only the gamete transfer barrier to keep apart Romeo and Juliet. Nature also has the two post-zygotic barriers. Nature may permit Romeo to fertilize Juliet, but then, cruelly, may not allow the two parental sets of chromosomes to cooperate effectively, either for development of the embryo or child (“hybrid inviability”), or for the pairing process (meiosis) within the adult gonad that is required for the formation of gametes for the next generation (“hybrid sterility”).

Transmission, development and gametogenesis – all three – depend on gene products. Ongoing gene cataloguing projects permit identification of gene subsets corresponding to these three barriers. It appears likely that there are far more genes in the development subset than in the transmission and gametogenesis subsets. After all, the latter subsets are restricted to a few specialized organ systems, whereas development involves all organ systems. Assuming that the mutation rate per gene is the same in each subset, then a mutation in the development subset is more likely to occur than mutations in the other subsets. This means that if a genic difference underlies cycle interruption, then over *short* evolutionary time scales that difference is most likely to be in the development subset. Thus, hybrid inviability keeps parents reproductively isolated because there are more mutational targets in the development subset.

On the other hand, transmission *always trumps inviability* since without transmission there can be no zygote to develop into an embryo. And if there is no embryo there can be no adult for gametogenesis. Yet, over *long* evolutionary time scales,

mutations in the transmission subset are *bound to occur*. Thus, if a genic difference underlies cycle interruption, that defect is most likely to *finalize* in the transmission subset. As far as reproductive isolation is concerned, any preexisting differences in the post-zygotic subsets that may have sustained reproductive isolation then become irrelevant. However, mutations in those subsets may remain and increase – free to fulfill other roles.

Jurassic Park

To better understand this we return to linear evolution. In Figures 8.2a-c the arrows illustrate progressive evolution within a line of organisms that, potentially, could progress from species to species. It just so happens we have not been able to cross an A with a B to demonstrate their reproductive incompatibility, thus proving them to be independent species. Paleontologists sometimes find samples of A and its descendents in the fossil record. From such fossils it is sometimes possible to isolate DNA fragments for sequencing. It is then possible to reconstruct the entire DNA ‘blueprint’ of the organism as it once existed. As portrayed by Crichton in *Jurassic Park* [16], it then becomes theoretically possible to reconstruct the entire organism.

This raises the hypothetical question, if we were to reconstruct ancient form A, would it be able to produce offspring if crossed with modern form B? In other words, having eliminated the temporal and any geographical barrier between the two forms (by placing them together in our ‘Jurassic Zoo’) would they produce offspring? Almost certainly not. And almost certainly the reason for this would be some disparity in their genes. If the time scale is great, there would be a transmission disparity. The genes responsible for this might affect mental states, anatomical and physiological features, or gametic characters. Even if not emotionally disparate, the two forms might differ anatomically or physiologically to such an extent that copulation would be impossible. And even if copulation were possible, gametic characters might have differed so preventing fusion to form the zygote (the fertilized egg).

But what if we were to reconstruct, not the most ancient form (A), but some intermediate form much closer to B (let’s call it “proto-B”)? This would correspond to a time-point on the shaft of the arrow (Figure 8.2a). As we moved progressively down the arrow we would arrive at a proto-B form that was so anatomically and/or functionally like B that the *transmission* barrier would be overcome. On the time scale separating proto-B from B, a transmission gene mutation would be improbable, but a developmental gene mutation would remain probable. The two forms would be able to copulate to generate a zygote but, even if able to complete some cell divisions, that zygote would not be able to develop into an embryo and thence into an adult. The transmission barrier would now have been replaced by a developmental barrier. There would be disparity between the parental genes required for development. The genes would be unable to cooperate. Their products would not work together.

Nevertheless, moving further down the arrow, we would eventually arrive at a form of proto-B that was so close to B that the developmental barrier would fall. Even though developmental differences were the most likely, in the time available the occurrence of such a mutation would be improbable. No longer would the two forms be reproductively isolated in the sense of not being able to produce a viable child. Would they still be reproductively isolated in the evolutionary sense – namely, would the seemingly healthy child, when an adult, be able to continue the line by reproducing with a B of the opposite sex? Almost certainly yes. And there would be a good chance that the offspring would be fertile. On this time-scale a defective gene producing hybrid sterility would seldom become manifest.

Similar considerations would apply if we could sample proto-B at various stages and back-cross to the original A? In the early stages (near to the origin of the arrow) the cross would be normal producing fertile offspring. But eventually a stage would be reached when the cross would fail due to the occurrence of a hybrid inviability mutation. Sampling Proto-B further down the arrow, we would find that a transmission barrier had superceded the inviability barrier.

Branching Evolution

The mode of evolution contemplated by Darwin depended on the classical triad – variation, natural selection and inheritance. However, in branching evolution (Fig. 8.2d), types of approximately equal fitness would initially be produced and there could be few phenotypic differences for natural selection to distinguish them by. To separate the ‘fit’ from the ‘fit’ there would need to be a barrier to prevent the blending of characters in offspring. *Another process would be required.* The triad would need to become a tetrad by the addition of some new form of reproductive isolation.

We first consider the linear stages that follow branching. At the bottom of the inverted Y in Figure 8.4 today’s Bs (say camels) and Cs (say pigs) are prezygotically isolated. Anatomical, physiological and psychological differences conspire to make this absolute. Yet, as we move back through time, up along the limbs of the inverted Y, one by one, these differences would diminish and disappear. Eventually a stage would be reached where copulation and fertilization to produce a zygote could occur between a proto-camel and a proto-pig. The transmission barrier would have fallen.

However, the developmental barrier would replace it. Hybrids would be inviable. Moving closer to the fork of the Y, the parental genes required for development would progressively lose their differences and their gene products would become increasingly compatible in the growing embryo. The hybrid inviability barrier would fall and development would proceed normally. We are now at the first barrier – the ‘spark’ that originated the speciation process.

Traditionally, the first barrier has been regarded as geographical and various plausible scenarios have been advanced. Indeed, a group within a species can easily

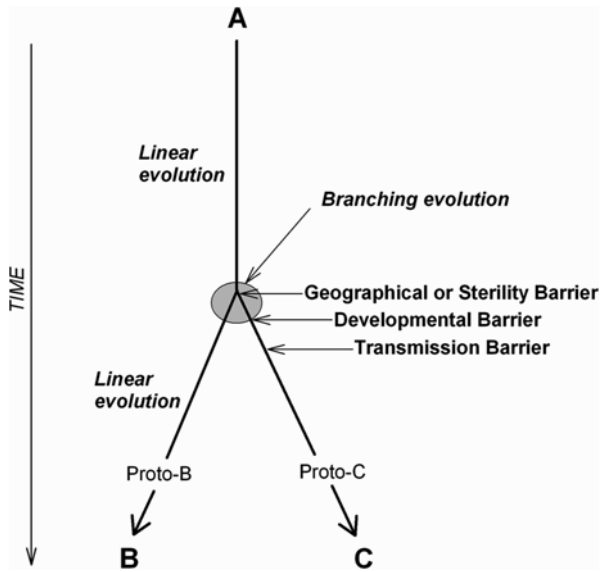


Fig. 8.4 Divergence into two species, which is preceded and followed by phases of linear evolution. Successive barriers to reproduction replace earlier ones. At the point of divergence the spark that initiates is some form of reproductive isolation, either extrinsic (geographic) or intrinsic (hybrid sterility). This may be followed by a developmental barrier (post-zygotic barrier producing hybrid inviability), and then by a gamete transfer barrier (prezygotic barrier)

become geographically isolated from the main species group. Under the protection of this *extrinsic* “allopatric” form of reproductive isolation, the groups can independently evolve to such an extent that, should the geographical barrier be removed, members of one group are no longer reproductively compatible with members of the other, due initially to hybrid inviability (developmental barrier), and eventually to prezygotic isolation (transmission barrier). More difficult to understand has been the *intrinsic* barrier of hybrid sterility (hybrid infertility). Protected by this barrier, a group within a species can become reproductively isolated *non-geographically* from the main species group.

By virtue of this “sympatric” reproductive isolation, the groups can independently evolve to such an extent that, after a sufficient interval, should the sterility barrier be removed, members of one group would no longer be reproductively compatible with members of the other, due to the developmental and transmission barriers that had arisen. Of course, mutations in genes affecting fertility being much rarer than mutations in genes affecting development, a developmental barrier would normally tend to preempt a *genic* sterility barrier. However, as we will see in Chapter 9, sterility does not necessarily have to reflect a genic malfunction. Numerous ‘collective’ mutations, most of them *non-genic*, can progressively cumulate below the ‘radar screen’ of natural selection to bring about sterility, and hence ‘spark’ divergent speciation.

Dobzhansky and Muller

Following the death of William Bateson in 1926, evolutionary ideas were contested in the 1930s by Goldschmidt, who offered a non-genic viewpoint, and by Theodor Dobzhansky and Herman Muller, both bedazzled by the seeming omnipotence of genes. They dismissed Goldschmidt's seemingly mystical wanderings and decreed that the *primary* change at the branch point (Fig. 8.4) was genic. When geographical (allopatric) isolation did not apply then a mutation in a "speciation gene" would itself suffice to initiate the sympatric isolation process. Their concern was not so much the stage in the generational cycle at which the speciation gene might act, but how a mutation in that gene could establish itself in the first place.

They imagined that a homozygous form with genes AA would not productively cross with a homozygous form aa because their hybrid (Aa) would be defective. By virtue of incompatibility between the products of A and a , the parental chromosomes might not cooperate to allow either normal development (hybrid inviability), or gametogenesis (hybrid sterility), or fertilization (transmission of gametes). The heterozygotes would be less fit than the parental homozygotes and hence natural selection would reproductively isolate the parents from each other. This condition could propitiate divergence into distinct lineages.

But to *begin* the process, Dobzhansky and Muller envisaged a *single* initial mutation in a member of a healthy homozygous population (AA). Since mutations generally occur stepwise, AA should first mutate to Aa , which would then mutate to aa . But, by their reasoning, the intermediate Aa would be both defective and rare. So it might not survive long enough either to further mutate to aa , or to interbreed with its own kind (another Aa) to produce offspring among which would be some aa types. So the allele a could not *begin* to become established in the population. The probability of aa homozygotes ever appearing was very low.

Their solution was to invoke multiple *non-allelic* incompatibilities, an example being two "speciation genes," a and b , that are incompatible. A healthy homozygous form might have two A alleles and two B alleles that are compatible (i.e. $AABB$). In one individual there might be a mutation to $AaBB$. In another individual there might be a mutation to $AABb$. Here the individual allelic heterozygous forms (Aa and Bb) were deemed viable. Among the offspring of crosses between $AaBB$ with $AaBB$ would be some $aaBB$ homozygotes (see Fig. 8.1). Likewise crossing of $AABb$ with $AABb$ would produce some $AAbb$ homozygotes. Thus the forms $AAbb$ and $aaBB$ could arise independently. However, hybrids between these forms would be incompatible due to the presence of both a and b , so branching into two lineages in sympatry could result.

There followed a search for speciation genes. Several decades later, few had been identified [17], and a promising early candidate was brought into question [18]. Currently, a meiosis-specific gene (PRDM9) that localizes to recombination 'hotspot' DNA sequences in a species-specific manner, is receiving much attention (see Chapter 17). But it is difficult to establish that a putative speciation gene is indeed primary and did not exert its effect secondarily after another form of reproductive isolation had operated.

Pattern Change

While Dobzhansky and Muller were elaborating their genic viewpoint, Goldschmidt, albeit with some hand-waving, was viewing chromosomes as a “reaction system” that, through a “repatting” involving “systemic mutations,” might change into another “reaction system” [19, 20]:

The classical theory of the gene and its mutations did not leave room for any other method of evolution. Certainly a pattern change within the serial structure of the chromosome, unaccompanied by gene mutation or loss, could have no effect whatsoever upon the hereditary type and therefore could have no significance for evolution. But now pattern changes are facts of such widespread and, as it seems, typical occurrence that we must take a definite stand regarding their significance. ... The pattern changes are in themselves effective in changing the genotype without any change of individual genes. ... Point mutations have never been known to change the point-to-point attractions between the homologous chromosomes in the heterozygote. ... A repatting of a chromosome may have exactly the same effect as an accumulation of mutations. ... The change from species to species is not a change involving more and more additional atomistic changes, but a complete change in primary pattern or reaction system into a new one, which afterwards may again produce intraspecific variation by micromutation.

Chromosomal “repatting” proceeded slowly and progressively, without necessarily producing any change in the structure or function of organisms, until a new species emerged that was reproductively isolated from the old one by virtue of the new pattern being “incompatible” with that of the old:

A systemic mutation (or series of such) ... consists of a change of intrachromosomal pattern... . Whatever genes or gene mutations might be, they do not enter this picture at all. Only the arrangement of the serial chemical constituents of the chromosomes into a new, spatially different order; i.e. a new chromosomal pattern, is involved. The new pattern seems to emerge slowly in a series of consecutive steps These steps may be without a visible effect until the repatting of the chromosome ... leads to a new stable pattern, that is, a new chemical system. This may have attained a threshold of action beyond which the physiological reaction system of development, controlled by the new genetic pattern, is so basically changed that a new phenotype emerges, the new species, separated from the old one by a bridgeless gap and an incompatible intrachromosomal pattern.

By “incompatible” Goldschmidt was here referring to differences between the chromosomes of two potential parents. These chromosomes would consequently not be able to cooperate functionally and/or to pair properly at meiosis within their child.

An unlimited number of patterns is available without a single qualitative chemical change in the chromosomal material, not to speak of a further unlimited number after qualitative changes (model: addition of a new amino acid into the pattern of a protein molecule). ... These pattern changes may be an accident, without any significance except for creating new conditions of genetic isolation by chromosomal incompatibility... .”

This may seem labored, but it should be recalled that in the 1930s it was known neither what genes were chemically, nor how that chemistry might be altered when mutations occurred. In striving to give some meaning to his concept of pattern, in 1940 Goldschmidt wrote in informational terms [20]:

Let us compare the chromosome with its serial order to a long printed sentence made up of hundreds of letters of which only twenty-five different ones exist. In reading the sentence a misprint of one letter here and there will not change the sense of the sentence; even the

misprint of a whole word (rose for sore) will hardly impress the reader. But the compositor must arrange the same set of type into a completely different sentence with a completely new meaning, and this in a great many different ways, depending upon the number of permutating letters and the complexity of the language (the latter acting as a 'selection'). To elevate such a model to the level of a biological theory we have, or course, to restate it in chemical terms.

Four years before Oswald Avery showed that DNA was the form in which hereditary information was transferred through the generations [21], and thirteen years before Watson and Crick presented their model for DNA, it was not unreasonable to think of chromosomal patterns in terms of amino acid, rather than nucleotide, sequences. Thus Goldschmidt wrote:

I do not think that an actual chemical model can yet be found. But we might indicate the type of such a model which fulfills at least some, though not all, of the requirements. It is not meant as a hypothesis of chemical chromosome structure, but only as a chemical model for visualizing the actual meaning of a repatterning process ... Let us compare the chromosome to a very long chain molecule of a protein. The linear pattern of the chromosome is then the typical pattern of the different amino acid residues.

The Goldschmidtian terminology was unwieldy and has not stood the test of time. But the underlying ideas resonate powerfully. The need for a change in reaction pattern for divergent evolution to occur finds an easy metaphor in human political systems. "Every boy and gal born into the world alive, is either a little Liberal, or else a little Conservative," went the Gilbert and Sullivan ditty as one Gladstone administration operating within a linear liberal reaction pattern was succeeded by a Disraeli administration operating within a linear conservative reaction pattern. Typically, dictatorships allow only linear progress. A change in reaction pattern may then require military intervention.

Summary

In principle there is no blending inheritance since characters correspond to discrete Mendelian units (genes) that are inherited quantally, in an either/or fashion, like the sex of an individual. However, the products of allelic genes (e.g. proteins) can sometimes interact to give the appearance of blending. Furthermore, many *quantitative traits*, such as height, depend on several non-allelic genes that can recombine in varying proportions resulting in a continuous, rather than discontinuous (quantal), distribution in a population. In this circumstance there is blending inheritance. Linear evolution proceeds cyclically through transmission of gametes, development of fertilized egg to adult, and gametogenesis. For a line of organisms to avoid blending inheritance, there must be branching into two lines (speciation) by interruption of the unitary generational cycle. Two independent cycles result. Since more genes are involved in development than in transmission or gametogenesis, if cycle interruption is of *genic* origin it is most likely to be first due to differences between parents in developmental genes – later trumped by differences in transmission

genes. On this basis, differences between parents in genes affecting gametogenesis in their child (hybrid) are unlikely to become manifest. However, Goldschmidt proposed cycle interruption due to failed gametogenesis because of a defective chromosome pairing of *non-genic* origin. Below the ‘radar screen’ of natural selection, “systemic mutations” would slowly accumulate to change chromosomal “patterning,” so that hybrids, although otherwise healthy, would be sterile. Nevertheless, with fresh compatible partners, the reproductively isolated parents would remain free to follow new evolutionary paths under the influence of natural selection. New species could then emerge.

References

1. Bennett JH (1983) *Natural Selection, Heredity and Eugenics. Including Selected Correspondence of R. A. Fisher with Leonard Darwin and Others*. Clarendon Press, Oxford, p 122
2. Butler S (1862) Darwin and the origin of species. Reproduced from *The Press* of Christchurch. In: Streatfeild RA (ed) *The First Year in Canterbury Settlement with Other Early Essays*. Fifield, London, (1914) pp 149–164
3. Fisher RL (1930) *The Genetical Theory of Natural Selection*. Oxford University Press, Oxford
4. Darwin C (1856) Letter to J. D. Hooker. In: Darwin F (ed) *Life and Letters of Charles Darwin*. Volume 1. Appleton, New York (1887) p 445
5. Bateson W (1909) Heredity and variation in modern lights. In: Seward AC (ed) *Darwin and Modern Science*. Cambridge University Press, Cambridge pp 85–101
6. Darwin C (1857) Letter to T. H. Huxley. In: Darwin F (ed) *More Letters of Charles Darwin*, Vol 1. Appleton, New York (1903) p 102
7. Hooker J (1860) On the origination and distribution of species. Introductory essay on the flora of Tasmania. *American Journal of Science & Arts* 29:1–25, 305–326
8. Mendel G (1865) Versuche uber Pflanzen Hybriden. *Verhandlung des naturforschenden Vereines in Brunn* 4:3–47 [The pea plant was a happy choice for Mendel. In this species the height character can be treated as unigenic. He was less fortunate in his studies with other plants.]
9. Romanes GJ (1894) Letter to Schafer, 18th May. Wellcome Museum of the History of Medicine, London [In the *Encyclopaedia Britannica* (1881), Romanes cited Mendel as one of those “who in more recent years have contributed to the literature on hybridism.”]
10. Darwin C (1866) Letter to A. R. Wallace. In: Marchant J (1916) *Alfred Russel Wallace. Letters and Reminiscences*. Harper, New York
11. Butler S (1878) *Life and Habit*, Trübner & Co., London, p. 168
12. Bateson W (1894) *Materials for the Study of Variation Treated with Especial Regard for Discontinuity in the Origin of Species*. Macmillan, London, pp 85, 573
13. Bateson W, Saunders ER (1902) Report 1. *Reports to the Evolution Committee of the Royal Society*. Harrison, London
14. Lind PA, Tobin C, Berg OG, Kurland CG, Andersson DI (2010) Compensatory gene amplification restores fitness after inter-species gene replacements. *Molecular Microbiology* 75:1078–1089
15. Forsdyke DR (2001) *The Origin of Species, Revisited. A Victorian Who Anticipated Modern Developments in Darwin's Theory*. McGill-Queen's University Press, Montreal
16. Crichton M (1990) *Jurassic Park*. Knopf, New York
17. Coyne JA, Orr HA (2004) *Speciation*. Sinauer, Sunderland, MA
18. Scharl M (2008) Evolution of *Xmrk*: an oncogene, but also a speciation gene? *BioEssays* 30:822–832

19. Richmond ML, Dietrich MR (2002) Richard Goldschmidt and the crossing-over controversy. *Genetics* 161:477–482
20. Goldschmidt R (1940) *The Material Basis of Evolution*. Yale University Press, New Haven, pp 205–6, 245–248 [The term “reaction system” was introduced to distinguish large genetic units, between which recombination was restricted (i.e. each was an individual “reaction system”), from individual genes that exhibited standard Mendelian behavior; see Goodspeed TH, Clausen RE (1917) *American Naturalist* 51:31–46, 92–101.]
21. Avery OT, Macloed CM, McCarty M (1944) Studies on the chemical transformation of pneumococcal types. *Journal of Experimental Medicine* 79:137–158

Chapter 9

The Weak Point

The introduction of slightly new elements into a method already established varies it beneficially; the new is soon fused with the old, and the monotony ceases to be oppressive. But if the new be too foreign, we cannot fuse the old and the new – nature seeming to hate equally, too wide a deviation from ordinary practice, and none at all. This fact reappears in heredity as the beneficial effects of occasional crossing on the one hand, and on the other, in the generally observed sterility of hybrids. If heredity be an affair of memory, how can an embryo, say of a mule, be expected to build up a mule on the strength of but two mule memories? Hybridism causes a fault in the chain of memory, and it is to this cause that the usual sterility of hybrids must be referred.

Samuel Butler 1890 [1]

On an evolutionary time scale, horse and donkey (ass) diverged from a common ancestor relatively recently. Their child, a “hinny” or “mule,” is a classic example of hybrid sterility. In Butler’s words, “reproduction is baffled” by being “pulled hither and thither by two conflicting memories” (Chapter 3). Darwin was also baffled. Thomas Huxley described Darwin’s proffered explanation as “the weak point” in an otherwise admirable theory. Huxley’s concerns were twofold. If offspring could be obtained from a cross between members of *different* species they were often sterile, but no breeder could produce offspring with this sterility from crosses between members of the *same* species. Furthermore, in proposing that evolution proceeded smoothly without jumps, punctuations or discontinuities, Darwin seemed to have imposed upon himself “an unnecessary difficulty” [2].

Romanes and Bateson were impressed by Huxley’s points. Yet, whereas the work of Mendel went unnoticed for a mere 35 years, their solutions, despite the work of Goldschmidt (Chapter 8), were disparaged or went unnoticed for over a century. This was due partly to the twentieth century’s obsession with genes, but also to the fact that they were struggling with novel concepts and the associated terminology. With terms such as “allele” and “homozygote” we have become familiar. But terms such as “collective variation” and “residue,” which would now appear highly relevant to speciation studies, have largely disappeared from view (Fig. 9.1).

Fig. 9.1 George John Romanes (1848-1894)



Cryptic Variability of the Reproductive System

In an address to the Linnaean Society in 1886, Romanes went beyond Darwin to point out that, apart from variations between individuals that we can easily detect (e.g. differences in eye color), there are likely to be variations that we cannot easily detect. Some variations in parental gonads might ultimately be detected by their conferral of the sterility phenotype on the otherwise healthy offspring of certain pairs. This would designate the parents as reproductively isolated from each other. The *partner-specific* reproductive-isolation-phenotype of the parents would be manifest as the sterility phenotype of their offspring. In Romanes' view, the establishment of this form of reproductive isolation – isolation of the *fit* from the *fit* that he called “physiological selection” – would in many cases have been critical for branching into two species, with or without the subsequent intervention of natural selection that would isolate (by elimination) the *unfit* from the *fit*.

Romanes' physiological selection was an abstraction for which there was then no cytological or biochemical basis. To give his Victorian audience an example of one form of physiological selection, Romanes pointed to a variation that might lead a plant to flower earlier than most members of its species. Early flowering forms would be reproductively isolated from late flowering forms by a *temporal barrier*. At an early stage, pollen kept from an early flowering form would still be capable of fertilizing flowers of late flowering forms. Over time, incompatibilities between early pollen and late ovules would randomly arise, and this mutual incompatibility (a further barrier) would eventually define the two forms as distinct species.

Today we attribute temporal (allochronic) and gamete-incompatibility (transmission) barriers to differences in genes. Likewise, a geographic (allopatric) barrier

provides an opportunity for genic differences to arise. But Romanes did not think these forms of variation would have operated in the general case to originate species. And he thought of variation affecting, not just single individuals, but a collection of them – “a section of a species:”

In many cases, no doubt, this ... variation, has been caused by the season of flowering or of pairing having been either advanced or retarded in a section of a species, or to sundry other influences of an extrinsic kind; but probably *in a still greater number of cases* it has been due to what I have called intrinsic causes, or to the ‘spontaneous’ variability of the reproductive system itself.

Romanes felt no obligation to further clarify intrinsic causes: “It is enough for the explanation which is furnished by Mr. Darwin’s theory of the evolution of adaptive structures by natural selection, that the variations in question take place; and similarly as to the present theory of the evolution of species by physiological selection”. On the other hand he was quite prepared to agree that this form of variation could proceed *continuously* as Darwin had proposed for the evolution of adaptations:

Starting from complete fertility within the limits of a single parent species, the infertility between derivative or divergent species, *at whatever stage in their evolution this began to occur*, must usually at first have been well-nigh imperceptible, and thenceforth have proceeded to increase stage by stage. ... Does it not therefore become ... in a high degree probable, that from the moment of its inception this isolating agency must have played the part of a segregating cause, in a degree proportional to that of its completeness as a physiological character? [Romanes’ italics]

Collective Variation of a Population Subset

Anticipating modern developments in population genetics, Romanes thought the slow increase of his postulated independent variation of the reproductive system might reflect its “collective” nature, in that it would affect a distinct lineage within a population. One of the misconceptions of his opponents was:

That I imagine physiological varieties always to arise ‘sporadically,’ or as merely individual ‘sports’ [mutations] of the reproductive system. On the contrary, I expressly stated that this is *not* the way in which I suppose the ‘physiological variation’ to arise, when giving origin to a new species; but that it arises, whenever it is effectual, as a ‘collective variation’ affecting a number of individuals simultaneously, and therefore characterizing ‘a whole race or strain’.

Free-intercrossing would have tended to swamp or blend the sporadically appearing variations upon which natural selection might act. But Romanes’ collective “independent variation” of the reproductive system – postulated to be a “general variability ... about a mean” (e.g. as human heights form a normal distribution) – was held to militate against this:

I am guilty of no inconsistency when thus arguing for a ‘collective variation’ on the part of the reproductive system, after having urged the difficulty against natural selection which arises from free intercrossing – i.e. the difficulty of supposing that a sufficient number of variations of the same kind should always be forthcoming simultaneously to enable natural selection to overcome the influence of free intercrossing. For, as previously explained, this

objection is only valid in the case of ‘accidental,’ ‘sporadic,’ or ‘spontaneous’ variations, which, *ex hypothesi*, are relatively very few in number. The objection does not apply to ‘collective variations,’ which, being due either to a common cause or to general variability of size, etc, about a mean, affect a number of individuals simultaneously.

There were many critics of physiological selection. Among them was Huxley who failed to recognize that his “weak point” criticism of Darwin had been met. Romanes had distinguished two types of variation: overt sporadic variations that might lead to the evolution by natural selection of better adapted forms (which the breeder might capture and selectively pair), and cryptic collective variations that might lead to the evolution of species by physiological selection (which the breeder would find hard to simulate).

Sporadic and Regular Sterilities

Romanes died in 1894, the year that Bateson published a major treatise – *Materials for the Study of Variation, Treated with Especial Regard to the Discontinuity in the Origin of Species*. At that time, in keeping with Huxley, Bateson saw the origin of species as a discontinuous process, similar to the sudden arising, in one generation, of a new variety of flower with one extra petal, or of a hand with an extra finger (Fig. 7.1). The suddenness of the emergence implied a discontinuous underlying mechanism uninfluenced by natural selection. By 1894 chromosomes had been named and their movements during cell division had been described. The division of somatic cells was named mitosis. The division of the germ cells in the gonad was seen to differ from mitosis and was later named meiosis (see below) (Fig. 9.2).



Fig. 9.2 William Bateson (1861–1926) with the biometrician Raphael Weldon (circa 1890). The two friends came to disagree strongly in later years.

Like various continental botanists, in the 1890s Bateson and Saunders began crossing various lines of plants and animals and scoring the inheritance of characters among the progeny. Following the discovery of Mendel's work in 1900, Bateson became its leading advocate in the English-speaking world. Given the intense opposition he faced from the mathematical biologists ("biometricians"), it is not surprising that his attention was directed more to verifying Mendel's laws than to pronouncing on their cellular basis. However, in October 1902 at a conference in New York he observed: "We have reason to believe that the chromosomes of the father plant and mother plant, side by side, represent blocks of parental characters".

In December 1901 Bateson and Saunders had reported to the Royal Society that the classical sterility of the hybrids resulting from a cross between species that were believed to have recently diverged from a common ancestor, did *not* follow a Mendelian pattern of character assortment among offspring. Although the offspring might have inherited some conventional character, such as color or height, in a Mendelian fashion (e.g. 3:1 ratio), usually the same offspring would *all* inherit the sterility phenotype (Fig. 9.3).

At that time they had to admit:

We know of no Mendelian case in which fertility is impaired. We may, perhaps, take this as an indication that the sterility of certain crosses is merely an indication that *they cannot divide up the characters among their gametes*. If the parental characters, however dissimilar, can be split up, the gametes can be formed. ... That the sterility of hybrids is generally connected in some way with inability to form germ-cells correctly ... is fairly clear, and there is in some cases actual evidence that this deformity of pollen grains in hybrids is due to irregularity or imperfection in the processes of division from which they result.

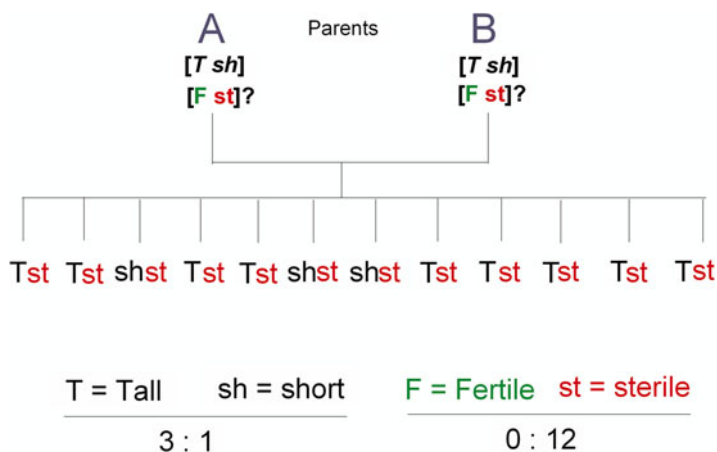


Fig. 9.3 The regular, predictable, sterility observed when members of *closely allied* species are crossed. Contrasted characters are tallness and shortness, and fertility and sterility. Parents, A and B, each have an allele for tallness (*T*) and an allele for shortness (*sh*). If tallness is dominant, the ratio of tall to short offspring (3 : 1) accords with Mendelian predictions. However, the *majority* of offspring are sterile (*st*), not fertile (*F*)

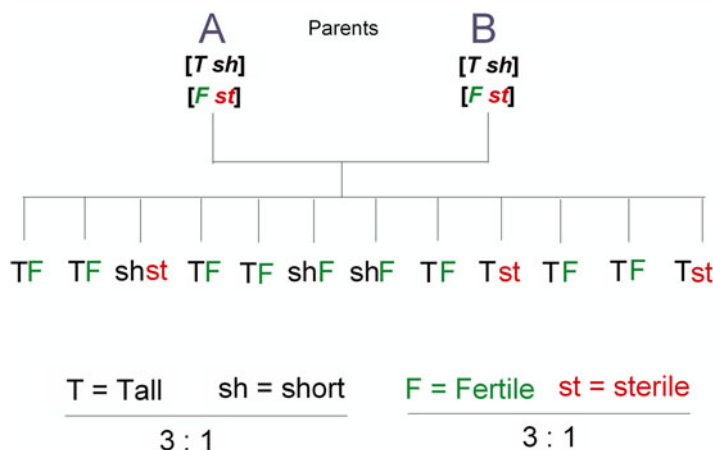


Fig. 9.4 The sporadic, unpredictable, sterility observed among the offspring of a *within-species* cross. Parents, A and B, are heterozygous for height as in Figure 9.3. They are also (unknowingly) heterozygous for a recessive sterility gene (*st*). The ratios of dominant to recessive among offspring accord with Mendelian predictions, thus indicating that sterility can be of genic origin

Thus, the hybrid sterility phenotype that resulted from a cross between allied species was something intrinsic to an organism and did not relate to extrinsic environmental elements. That they were thinking of an imperfection in the pairing of homologous parental chromosomes at meiosis is indicated by their citing the US biologist Michael Guyer who “in ignorance of Mendel’s work” had independently derived Mendel’s laws from the cytology of spermatogenesis in pigeon hybrids [3]. Furthermore, early in 1902 Bateson noted:

It is impossible to be presented with the fact that in Mendelian cases the cross-bred produces on an average *equal* numbers of gametes of each kind, that is to say, a symmetrical result, without suspecting that this fact must correspond with some symmetrical figure of distribution of those gametes in the cell-divisions by which they are produced.

But soon Bateson and his colleagues came across cases of infertility that followed Mendelian patterns of inheritance (Fig. 9.4). In a second report to the Royal Society in 1905 they drew a clear distinction between this unpredictable *sporadic* sterility seen among the offspring of crosses between *closely related* individuals, and the predictable sterility seen among the offspring of crosses between *distantly related* individuals (i.e. members of allied species):

Contabescent [withered] anthers were seen from time to time in many families ... This sporadic sterility has not been particularly studied. It is of interest to compare this ... with the familiar occurrence of sterility in cross-breds. Such a relationship has often been supposed to indicate remoteness of kinship, yet here a closely comparable effect occurs in ... a cross between two types which must be very nearly related. Mr. Gregory in a careful examination of the pollen-genesis, found that the divisions were normal up to the reduction division, when the chromosomes form shapeless knots and entanglements, failing to divide.

Today we would interpret the latter sterility as a malfunction due to mutation in a gene determining one of the many cell components required to produce pollen. Whatever the ultimate basis of the sterility, Bateson and his colleagues pointed out that this sporadic event arising from a *within-species* cross (unpredictable in that it had not been anticipated when the cross was first made), contrasted with the *regular* (predictable) sterility appearing when allied species in general were crossed. Only the within-species cross displayed a Mendelian (i.e. genic) pattern of inheritance.

In 1902 Bateson and Saunders recognized that Mendel's unit characters were "the sensible manifestations of physiological units of as yet unknown nature." Although Bateson did not begin to use the word "gene" until the second decade of the twentieth century, it is employed here when referring to the Mendelian "physiological units" that determine characters.

Chromosomes and the Residue or Irresoluble Base

Bateson was among the first to see a link between Mendelian "genes" and chromosomes. However, as evidence for a chromosomal location of genes became stronger, Bateson's doubts increased. It was not a question of Bateson not seeing the wood from the trees. His idea of what constituted the wood differed from that of others. They saw a collection of genes. He saw the genes plus something else that was in some way related to the question of the origin of species. What was the something else and where was it located? Was it on the chromosomes like the Mendelian factors? Or was it elsewhere, perhaps outside the nucleus in the cytoplasm? No clear evidence on this had emerged by 1926, the year of Bateson's death. Thus his pronouncements on chromosomes were always carefully hedged. That the story was somehow incomplete was recognized in 1902:

Has a given organism a fixed number of unit characters? Can we rightly conceive of the whole organism as composed of such unit characters, or is there some residue – a basis – upon which the unit characters are imposed?

This was expressed in slightly different form a few months later:

From the fact of the existence of interchangeable [allelic] characters we must, for purposes of treatment, and to complete the possibilities, necessarily form the conception of an *irresoluble base* [i.e. a base not resolvable by Mendelian analysis], though whether such a conception has any objective reality we have no means as yet of determining.

The report of Bateson and Saunders linked the concept to Darwin's great question, the nature of the origin of species – a "phenomenon" that did not "attach to" the elements (genes) that determined allelic characters:

We know, of course, that we cannot isolate this residue from the unit characters. We cannot conceive a pea, for example, that has no height, no colour, and so on; if all these were removed there would be no living organism left. But while we know that all these characters can be interchanged [through breeding experiments], we are bound to ask is there something not thus interchangeable? And if so, what is it? We are thus brought to face the further question

of the bearing of Mendelian facts on the nature of Species. The conception of species, however we may formulate it, can hardly be supposed to attach to [genes] ... We may be driven to conceive 'Species' as a phenomenon belonging to that 'residue' spoken of above, but on the other hand we get a clearer conception of the nature of sterility on crossing.

Two Factors must Complement to Produce Sterility

Bateson was considering the sterile offspring of crosses between allied species that were presumed to have derived from a common ancestral species. Since the offspring of crosses between members of that ancestral species would have been fertile (otherwise there would have been no descendents), then the hybrid sterility revealed the emergence of a reproductive barrier between the descendent species (i.e. an interruption of the unitary generational cycle; Fig. 8.3). This could have been an *originating* barrier leading to the state of reproductive isolation that defined them as distinct species. Bateson developed this theme abstractly for the Darwin centenary celebration in 1909 [4]. Here he concluded that "*Failure to divide* [in order to produce gametes] is ... the immediate 'cause' of the sterility," so that "we are justified in supposing that there are factors which can arrest or prevent cell division":

When two species, both perfectly fertile severally, produce on crossing a sterile progeny, there is a presumption that the sterility is due to the development in the hybrid of some substance which can only be formed by the meeting of two complementary factors.

Applying this line of reasoning to what would appear to be a single species, Bateson continued:

We see that the phenomenon could only be produced among the divergent offspring of one species by the acquisition of at least *two* new factors; for if the acquisition of a single factor caused sterility the line would then end. Moreover each factor must be separately acquired by distinct individuals, for if both were present together, the possessors would by hypothesis be sterile. And in order to imitate the case of species, each of these factors must be acquired by distinct breeds. The factors need not, and probably would not, produce any other perceptible effects; they might ... make no difference to the form of other characters. Not until the cross was actually made between the complementary individuals would either factor come into play, and the effects even then might be unobserved until an attempt was made to breed from the cross-bred.

Thus, the factors (say Z') which distinguished certain members of a species from their fellows (bearing say Z) would be latent (cryptic). The factors would become patent, revealing a sterility-of-offspring phenotype, only if, by chance, there were a cross between certain individuals that had come to differ in the factors (i.e. Z' and Z ; not Z' and Z' , or Z and Z). At least one of the individuals would be rare, since the normal abundant types were generally fertile when intercrossed (i.e. no role was ascribed to Z in the presence of Z , or Z' in the presence of Z'). How often could this happen?

If the factors responsible for sterility were acquired, they would in all probability be peculiar to certain individuals and would not readily be distributed to the whole breed. Any member of the breed into which *both* the factors were introduced would drop out of the pedigree by virtue of its sterility. Hence the evidence that various domesticated breeds, say

of dogs or fowls, can when mated together produce fertile offspring, is beside the mark. The real question is, Do they ever produce sterile offspring? I think the evidence is clearly that sometimes they do, oftener perhaps than is commonly supposed.

Sterility-of-offspring being a *parental* phenotype, the fact that a completely sterile offspring could not be subjected to further breeding analysis would not matter. The parents of the offspring complemented sufficiently to produce the phenotype, but the complementary factor of one parent (Z') might be more prone to produce the phenotype than the complementary factor of the other (Z). Through segregation (further cross-breeding of each parent with other individuals) these complementary factors might be investigated.

In 1909 Bateson considered two sterility factors working together (complementing) to *impede* fertility (Fig. 9.5a). However, he later noted that organisms could be

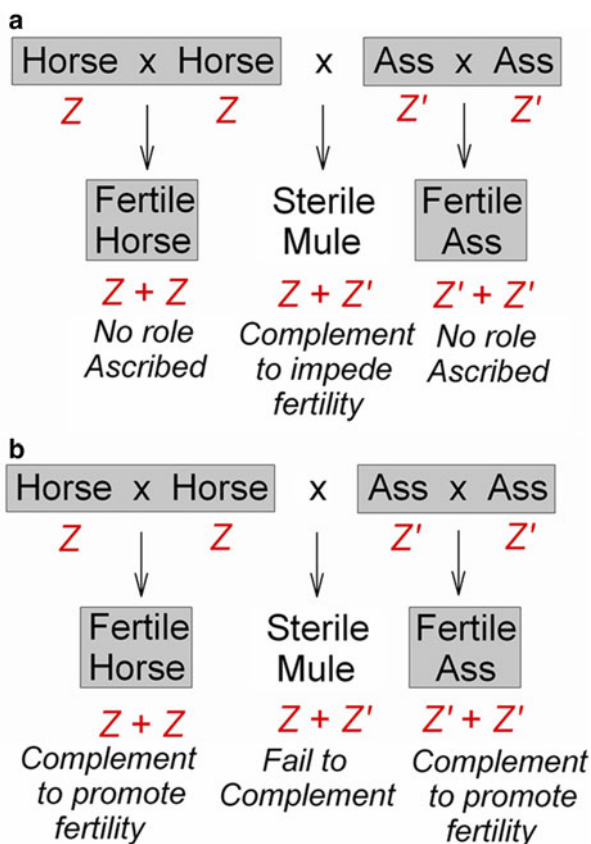


Fig. 9.5 Hybrid sterility in crosses between allied species (e.g. horse, ass) as due to hypothetical factors, Z and Z' , which either (a) *complement* to bring about a negative effect, or (b) *fail to complement* to bring about a positive effect. In both cases the result is negative (i.e. sterility). Since parents display fertility with their own kind (intraspecific cross), Bateson first deduced (a) that in an interspecific cross Z and Z' complement to bring about a negative effect. Later, he entertained the converse (b), that Z and Z' fail to complement to bring about a positive effect

reproductively incompatible because the reproductive system of “each is lacking in one of two complementary elements” that *promote* fertility (Fig. 9.5b). This view, reminiscent of Romanes, was later adopted by Goldschmidt, who proposed that normally parents contribute complementary factors (e.g. Z and Z) making parental chromosomes *compatible* at meiosis in their hybrids, which hence are fertile (i.e. the parental factors work together to produce a *positive* effect). When the factors are not sufficiently complementary (e.g. Z and Z') the parental chromosomes are *incompatible* in their hybrids, which hence are sterile.

Complementation like Sword and Scabbard

Bateson persisted with his Huxleyan idea that the sterility of a hybrid resulting from a cross between allied species was “a distinction in kind, of a nature other than those we perceive among our varieties.” And he was still attached to the Huxleyan idea of discontinuous variation. In 1921 he addressed the Toronto meeting of the American Association for the Advancement of Science:

But that particular and essential bit of the theory of evolution which is concerned with the origin and nature of *species* [Bateson's italics] remains utterly mysterious. We no longer feel ... that the process of variation, now contemporaneously occurring, is the beginning of a work which needs merely the element of time for its completion; for even time cannot complete that which has not yet begun. The conclusion in which we were brought up, that species are a product of a *summation of variations* ignored the chief attribute of species first pointed out by John Ray that the product of their crosses is frequently sterile in greater or less degree. Huxley, very early in the debate pointed out this grave defect in the evidence, but before breeding researches had been made on a large scale no one felt the objection to be serious.

A subsequent interchange with Plymouth physician, C. R. Crowther, is of particular interest [5]. Questioning Bateson's attitude to the mechanism of meiotic chromosome pairing (Fig. 9.6), Crowther began by noting that, while parental chromosomes had to cooperate for *development* of the zygote from embryo to adult, a far *higher* degree of cooperation would be needed when the chromosomes paired (conjugated) in the gonad of that adult:

Homologous chromosomes ... have to cooperate to produce the somatic cell of the hybrid, and their co-operation [for embryo development] might be expected to require a certain resemblance; but for the production of sexual cells [gametogenesis] they must do more, they must conjugate [pair]; and for conjugation it is surely reasonable to suppose that a much more intimate resemblance would be needed. We might, therefore, expect, on purely theoretical grounds, that as species and genera gradually diverged, it would be increasingly difficult to breed a hybrid between them; but that, even while a hybrid could still be produced, a fertile hybrid would be difficult or impossible, since the cells of the germ-track would fail to surmount the meiotic reduction stage when the homologous chromosomes conjugate. This is exactly what happens: the cells go to pieces in the meiotic phase.

Bateson's disparagement of the idea that species might be “a product of a summation of variations” left Crowther “frankly puzzled,” for “the proposition is certainly not self-evident.” Surely, if the sterility of an offspring were due to a failure

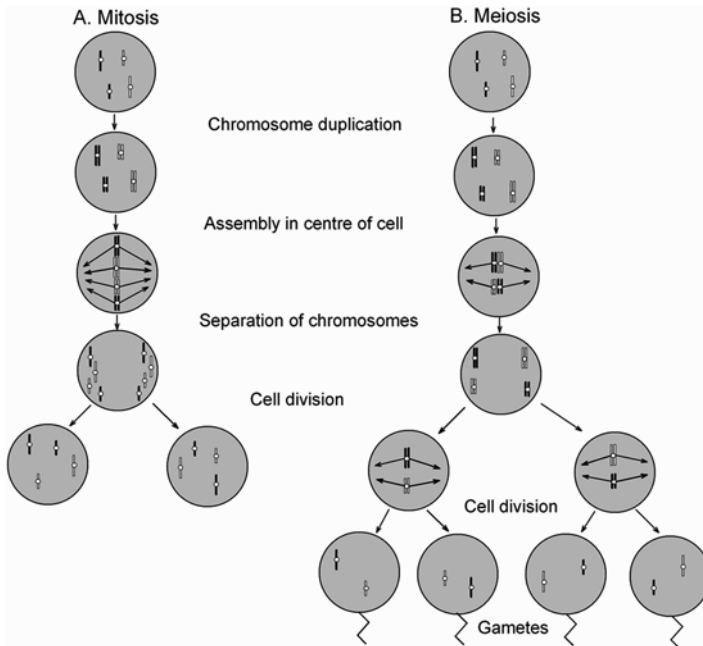


Fig. 9.6 Chromosome movements in (a) somatic cells during mitosis, and (b) germ-line cells during meiosis. Maternal and paternal chromosomes are white and black, respectively. Pairing (“conjugation”) of homologous chromosomes occurs in meiosis

within that offspring of homologous chromosomes to pair, it mattered little whether the lack of complementarity responsible for that failure was produced by one large variation, or by the summation of many smaller variations. That Crowther was thinking of primary variations occurring at the chromosomal level, rather than anatomical variations of the sterile individual, was explicit:

If a sword and scabbard are bent in different directions, it will happen sooner or later that the sword cannot be inserted, and the result will be the same whether the bending be effected by a single blow, or whether it be, in Dr. Bateson’s words, ‘a product of a summation of variations.’ Is this illustration apt? The sword and the scabbard are the homologous chromosomes. ... It seems easier to imagine sterility arising from a gradual modification, spread over a length of time, and involving many chromosomes.

Bateson conceded that discontinuity of variation was not critical:

It is ... not difficult to ‘imagine’ interspecific sterility produced by a gradual (or sudden) modification. That sterility might quite reasonably be supposed to be due to the inability of certain chromosomes to conjugate, and Mr. Crowther’s simile of the sword and the scabbard may serve to depict the sort of thing we might expect to happen. But the difficulty is that we have never seen it happen to swords and scabbards which we know to have belonged originally to each other. On the contrary, they seem always to fit each other, whatever diversities they may have acquired.

Modern Chromosomal Views

In 1909 Bateson did not spell out the relation between his postulated complementary sterility factors and the above “residue” or “irresoluble base.” In Bateson’s mind the “residue” was not only distinct from genes – the agencies determining the “transferable characters” that were the subject of Mendelian analysis – but it could also *limit* the extent to which organisms could exchange genes. This limitation (i.e. absence of gene flow) would be complete when the organisms belonged to independent species:

These transferable characters are attached to or implanted upon some basal organization, and the attributes of powers which collectively form that residue may perhaps be distinguished from the transferable qualities. The detection of the limits thus set upon the interchangeability of characters would be a discovery of high importance and would have a most direct bearing on the problem of the ultimate nature of Species.

In 1922 Bateson agreed with Crowther that a fundamental form of reproductive isolation, manifest as the hybrid sterility seen when members of allied species were crossed, was due to an incompatibility that could be characterized cytologically as impaired pairing of paternal and maternal chromosomes at meiosis. It was inferred that if we can understand what makes chromosomes incompatible, then we can understand hybrid sterility. And if we can understand hybrid sterility, we can understand an origin of species.

Indeed, in Denmark Öjvind Winge had suggested in 1917 that the incompatibility hypothesis predicted that the sterility would be ‘cured’ if the chromosome number (normally diploid) were doubled (to become tetraploid). Then each paternal chromosome could pair with its paternal homolog, and each maternal chromosome could pair with its maternal homolog [6]. There is now abundant evidence supporting this (see Chapter 10). But how do chromosomes that are homologous (i.e. are *alike*) pair with each other? Do they pair by virtue of this likeness (like-with-like), of by virtue of some key-in-lock (sword-in-scabbard) complementarity, which implies that they are not really alike? One must be the sword and the other the scabbard.

This paradox was resolved when it was appreciated that hereditary information was stored and transmitted as duplex DNA with two strands – a ‘Watson’ strand and a ‘Crick’ strand – that paired with each other by virtue of base complementarity (Chapter 2). So, in Crowther’s terminology, potentially the sword strand of one chromosome can pair with the scabbard strand of the homologous chromosome (and vice versa). For this swords have to be unsheathed from their own scabbards and then each inserted into the scabbards of the other. Thus the Watson strand of one chromosome can pair with the Crick strand of the other, and vice versa. This requires that the Watson strand be displaced from pairing with the Crick strand of its own chromosome. Likewise, the Crick strand of the homologous chromosome must be displaced from pairing with the Watson strand of its own chromosome. Then cross-pairing can occur.

The pairing requires complementarity of DNA base sequences. A sporadically appearing change in an *individual* base could, if dominant, introduce a new phenotype (e.g. an extra finger), but would not greatly affect the overall complementarity between parental chromosomes. However, over time, base changes – including some affecting genes, but *many* not affecting genes – could accumulate.

Romanes' "collective variation" that would build up "in a whole race or strain" (i.e. "a section of a species:"), can now be interpreted as the variation of the base composition of DNA about some characteristic value for the species (i.e. a normal distribution). When, through such 'micromutations,' differences between chromosomal homologs reached a critical value, meiotic pairing would be impaired and gametogenesis would abruptly fail (see Chapter 10). This early gametogenesis barrier would eventually yield to the developmental and transmission barriers – both of genic origins (Chapter 8).

Impaired gametogenesis can also follow variations in chromosome structure and number ('macromutations'), and the question arises as to which chromosomal variation is primary and which secondary, in the general case. Today's mules differ in numerous chromosome segments from their parental lines, and also have 63 chromosomes whereas parents have 62 and 64 [7]. This does not exclude the possibility that base composition changes were primary and the segmental and numerical changes were secondary. For yeast, in 2009 David Greig excluded segmental chromosomal changes as primary and favored "simple sequence divergence," a hypothesis that is gaining increasing support in other organisms [8–12].

Nevertheless, doubts remain. In the general case, over 4 billion years of bioevolution have non-genic differences sparked species originations more often than genic changes? Are non-genic differences involving multiple individual base changes more likely to have been primary, with gross chromosomal changes coming later?

Darwin's Difficulties

Darwin thought that natural selection would itself suffice to bring about branching evolution. Indeed, it *appears* to do so in certain circumstances, as when segments of a species become geographically isolated from each other (Fig. 8.4). However, here the branching *agency* is whatever caused the geographical isolation. A gene-encoded tendency to migrate may contribute. But the primary cause of geographical isolation is usually *not* natural selection. Speciation requires isolation in some shape or form. The problem of *the* origin of species is that of determining what form isolation takes *in the general case*.

In his faith in natural selection, Darwin was like the early chemists who were satisfied with atoms as the ultimate basis of matter. But for some chemists, phenomena such as swinging compass needles (magnetism), falling apples (gravity), and (later) radioactivity, were manifestations of something more fundamental in chemistry than atoms. Likewise, for some biologists the phenomenon of hybrid sterility

seemed to manifest something more fundamental in biology than natural selection. Since he held there was a necessary relationship between natural selection and hybrid sterility, Darwin was baffled [13]:

On the case of natural selection ... much of the sterility of hybrids could not possibly be of any advantage to them, and therefore could not have been acquired by the continued preservation of successive profitable degrees of sterility.

He here wrote about a hybrid that, by virtue of its sterility, would be dead in an evolutionary sense since it could no more continue the line than a slow deer, which had been eaten by a tiger. But what really mattered were the *parents* of the sterile hybrid, which itself was merely the expression of *their* incompatibility. Darwin seemed not to see that at least one of the parents might have been on a ‘journey’ (by accumulating mutations) towards a new species, thus renouncing membership of the old one. As long as it was traveling in the right direction without adverse phenotypic changes due to genic mutations, the further that parent ‘traveled’ from the old species the closer it *must* have got to the new one. Sterility of the offspring produced by crossing with a member of the old species would be considered disadvantageous to a parent when it was still perceived as being close to most members of that old species (i.e. when it had not ‘traveled’ far by mutation). But when perceived as a member of a potential new species, increasing degrees of sterility of the offspring produced by crossing with members of the old species (as the member of the potential new species continued ‘traveling’ and accumulating further mutations) could be considered extremely “profitable,” since it could eventually lead to ‘species arrival.’ There would then be complete reproductive *incompatibility* with members of the old species, but complete reproductive *compatibility* with fellow members of the new species.

This point was made to Darwin in February 1868 by Wallace who looked at the issue from the perspective of the line of organisms (“form”) rather than of the sterile individual itself [14]:

I do not see your objection to sterility between allied species having been aided by natural selection. It appears to me that, given a differentiation of species into two forms, each of which was adapted to a special sphere of existence, every slight degree of sterility would be a positive advantage, not to the individuals who were sterile, but to each form.

In other words, Wallace was postulating that selection of the species could override selection of the individual.

Species Selection

Pushing the metaphors further, the old species can be viewed as constituting a ‘reproductive environment’ that *molds* the genome phenotype (“reprotype”) by negatively selecting (by denying reproductive success to) variant organisms that attempt (by mating and producing healthy, fertile, offspring) to *recross* the emerging interspecies boundary. Thus, the old species positively selects *itself* by

negatively selecting variants. Should these variants find compatible mates, then they might accumulate as a new species that, in turn, would positively select itself by negatively selecting further variants. This is “species selection,” a form of group selection that some biologists have found hard to imagine.

As we shall consider further (Chapter 10), the old species is the established main species, members of which imperil themselves only marginally by mating with (denying reproductive success to) members of a small potentially incipient species. Thus, in reproductive interactions between a main and an incipient species, survivability of the main species is coupled *negatively* to the sum of the survivabilities of individual members of the incipient species (i.e. it survives when they do not survive), just as much as it is coupled *positively* to the sum of the survivabilities of its own individual members (i.e. it survives when they survive).

Of course, by individual survivabilities is meant, not just mere survival, but survival permitting unimpeded production of fertile offspring. Survival of members of an incipient species occurs, not only when classical Darwinian phenotypic interactions are favorable (e.g. escape from a tiger), but also when reprotypic interactions are favorable (e.g. no attempted reproduction with members of the main species). Tigers are a *phenotypic* threat. Members of the main species are a *reprotypic* threat.

Individual members of a main species that are involved (when there is attempted intercrossing) in the denial of reproductive success to individual members of an incipient species, are like individual stones in the walls of a species fortress against which the reproductive arrows of an incipient species become blunted and fall to the ground. Alternatively, the main species can be viewed as a Gulliver who barely notices the individual Lilliputian incipients brushed off or trampled in his evolutionary path. Just as individual cells acting in collective phenotypic harmony constitute a Gulliver, so individual members of a species acting in collective reprotypic harmony constitute a species. That harmony is threatened, not by its own members, but by deviants that, by definition, are no longer members of the main species (since a species is defined as consisting of individuals between which there is no reproductive isolation). These deviants constitute a potential incipient species that might one day pose a phenotypic threat to the main species (i.e. through their increase they may become an environmental hazard to the latter).

As we shall see (Chapter 10), members of the main species are most likely to encounter *other members of the main species* – hence, there will be fertile offspring. Members of an incipient species, being a minority, are also most likely to encounter members of the main species – hence, there will be infertile (sterile) offspring. Much more rarely, a member of an incipient species will encounter a fellow incipient species member with which it can successfully reproduce – an essential precondition for species divergence. Once branching is *initiated*, natural selection should help the branches sprout (extend in length) by allowing the survival of organisms with advantageous genic variations, and disallowing the survival of organisms with disadvantageous genic variations. The advantageous variations could include a type that the main species would not organically countenance. Thus the *range* of variation could expand. New environments could be explored.

Still baffled, Darwin turned the discussion over to his mathematical son, George, whose attempts to ‘enlighten’ Wallace tended to cloud rather than clarify. Sticking to his guns, in March 1868 Wallace wrote to Darwin [15]:

1. A species varies occasionally in two directions, but owing to their free intercrossing they the varieties never increase.
2. A change of conditions occurs which threatens the existence of the species, - but the *two varieties* are adapted to the changing conditions and if accumulated will form *two new species adapted to the new conditions*.
3. Free crossing however renders this impossible, and so the species is in danger of extinction.
4. If *sterility* could be induced, then the pure races would increase more rapidly and replace the old species.
5. It is admitted that *partial sterility* between *varieties* does occasionally occur. It is admitted that the *degree* of this *sterility varies*; is it not probable that Nat. Select. can accumulate these variations and thus save the species? If Nat. Select. can *not* do this how do species ever arise, except when a variety is isolated [geographically]? ...

The regularity with which Nat. species are sterile together, even when *very much alike*, I think is an argument in favour of the sterility having been generally produced by Nat. Select. For the good of the species. [all italics are Wallace’s]

This approximates to Romanes’ “collective variations” case. But sadly, Wallace and Romanes could not see that they were in agreement, and there was some sour correspondence [16].

Summary

Huxley identified as Darwin’s “weak point” the failure of breeders when crossing varieties *within* a species to simulate the sterility of hybrids derived from crosses between species that had recently diverged from a common ancestor (allied species). Due to the sterility, the parents of a hybrid are, in an evolutionary sense, reproductively isolated from each other. In his theory of “physiological selection,” Romanes postulated germ-line “collective variations” that accumulate in certain members of a species, who are “physiological complements” producing fertile offspring when mutually crossed, but sterile offspring when crossed with others. Unlike Darwin’s natural selection that secured reproductive isolation of the fit by elimination of the unfit, physiological selection postulated variations in the reproductive system that were not targets of natural selection; these sympatrically isolated the fit from the fit, leaving two species where there was initially one. Bateson approved of physiological selection. He postulated a non-genic “residue” that is “attached to or implanted upon” the chromosomes where the genes reside. Here mutations that might correspond to the “collective variations” of Romanes can accumulate, eventually exceeding a critical value in a sector of the population. In this event a major *discontinuity* – an origin of species – can result. Modern studies in a variety of species support the Romanes-Bateson viewpoint. While genic differences *differentiate* species, the spark that *originates* can be non-genic.

References

1. Butler S (1890) The deadlock in Darwinism. *Universal Review*. Reprinted in: *Essays on Life, Art and Science*, Streatfeild R (ed) Grant Richards, London, 1904, pp 234–340
2. Forsdyke DR (2010) George Romanes, William Bateson, and Darwin's "weak point." *Notes & Records of the Royal Society* 64:139–154
3. Bungener P, Buscaglia M (2003) Early connection between cytology and Mendelism: Micheal F. Guyer's contribution. *History and Philosophy in the Life Sciences* 25:27–50. [Michael Guyer's thesis (1900) may be viewed at: <http://post.queensu.ca/~forsdyke/guyer01.htm>]
4. Bateson W (1909) Heredity and variation in modern lights. In: Seward AC (ed), *Darwin and Modern Science*. Cambridge University Press, Cambridge, pp 85–101
5. Crowther CR (1922) Evolutionary faith and modern doubts. *Nature* 109, 777 [Crowther served with the rank of Captain in WW1 (25th Field Ambulance, RAMC).]
6. Winge Ö (1917) The chromosomes. Their numbers and general importance. *Comptes Rendus des Travaux du Laboratoire Carlsberg* 13:131–275. [Based on Winge's doctoral thesis, selections from this important paper are displayed at: <http://post.queensu.ca/~forsdyke/chromos1.htm>]
7. Chandley AC, Jones RC, Dott HM, Allen WR, Short RV (1974) Meiosis in interspecific equine hybrids. 1. The male mule (*Equus asinus* X *E. caballus*) and hinny (*E. caballus* X *E. asinus*). *Cytogenetics and Cell Genetics* 13: 330–341
8. Greig D (2009) Reproductive isolation in *Saccharomyces*, *Heredity* 102, 39–44
9. Kao KC, Schwartz K, Sherlock G (2010) A genome-wide analysis reveals no nuclear Dobzhansky-Muller pairs of determinants of speciation between *S. cerevisiae* and *S. paradoxus*, but suggests more complex incompatibilities. *PLOS Genetics* 6, e1001038
10. Nevo E (2012) Speciation: chromosomal mechanisms. *Encyclopedia of Life Sciences*. John Wiley, Chichester
11. Nei M, Nosawa M (2011) Roles of mutation and selection in speciation: from Hugo de Vries to the modern genomic era. *Genome Biology & Evolution* 3:813–829
12. Dion-Côté A-M, Symonová R, Ráb P, Bernatchez L (2015) Reproductive isolation in a nascent species pair is associated with aneuploidy in hybrid offspring. *Proceedings of the Royal Society B* 282: article 20142862. [Stressing the importance of evaluating genome changes early in the speciation process, the authors studied lines of whitefish that had diverged from a common ancestor 20,000 generations earlier. Noting "no detectable chromosomal rearrangements between parental forms," it was concluded that "meiotic breakdown" arose because "genetic incompatibilities may accumulate early during speciation and limit hybridization between nascent species."]
13. Darwin C (1872) *On the Origin of Species by Means of Natural Selection*. 6th Edition. John Murray, London
14. Vorzimmer PJ (1970) *Charles Darwin: The Years of Controversy*. Temple University Press, Philadelphia, pp. 206–207
15. Wallace AR (1868) Letter to Darwin, 24th March. Available at <https://www.darwinproject.ac.uk/entry-6045>
16. Forsdyke DR (2001) *The Origin of Species, Revisited*. McGill-Queen's University Press, Montreal, pp. 207–209

Chapter 10

Chargaff's GC rule

DNA is in its composition characteristic of the species from which it is derived. This can ... be demonstrated by determining the ratios in which the individual purines and pyrimidines occur There appear to exist two main groups of DNA, namely the 'AT type,' in which adenine and thymine predominate, and the 'GC type,' in which guanine and cytosine are the major constituents."

Erwin Chargaff (1951) [1]

Evolutionary selective pressures sometimes act to preserve nucleic acid features at the expense of encoded proteins. That this can occur in the case of nucleic acid secondary structure was noted in Chapter 5. That this might also apply to the species-dependent component of the base composition, (G+C)%, was shown by Sueoka in 1961 [2]. The amino acid composition of the proteins of bacteria is influenced, not only by the demands of the environment on the proteins, but also by the (G+C)% of the genome encoding those proteins.

From the genetic code (Table 7.1) it can be inferred that **AT**-rich genomes tend to have more codons for the amino acids **Phe**, **Tyr**, **Met**, **Ile**, **Asn**, and **Lys**. These amino acids have single-letter abbreviations that collectively spell "FYMINK." **GC**-rich genomes tend to have more codons for the amino acids **Gly**, **Ala**, **Arg** and **Pro**. These amino acids have single-letter abbreviations that collectively spell "GARP." If you know the FYMINK/GARP ratio of an organism's proteins, then you can make a good guess as to its (G+C)%. Sueoka's observation has since been shown to apply to a wide variety of animal and plant species. Could an organism's (G+C)% have arisen randomly and yet be so powerful that it could force a gene to encode an amino acid 'against its will' (i.e. to encode an amino acid that might not best serve the needs of the encoded protein, and hence, of the organism)? Can a selective force (pressure) be recognized?

Uniformity of (G+C)%

Chargaff's "GC rule" is that the ratio of (G+C) to the total bases (A+G+C+T) tends to be constant in a particular species, but varies between species. Sueoka further pointed out that for individual "strains" of *Tetrahymena* (ciliated unicellular organisms) the (G+C)% (referred to as "GC") tends to be uniform *throughout* the genome:

If one compares the distribution of DNA molecules of *Tetrahymena* strains of different mean GC contents, it is clear that the difference in mean values is due to a rather uniform difference of GC content in individual molecules. In other words, assuming that strains of *Tetrahymena* have a common phylogenetic origin, when the GC content of DNA of a particular strain changes, all the molecules undergo increases or decreases of GC pairs in similar amounts. This result is consistent with the idea that the base composition is rather uniform not only among DNA molecules of an organism, but also with respect to different parts of a given molecule.

There was also a link between (G+C)% and reproductive isolation:

DNA base composition is a reflection of phylogenetic relationship. Furthermore, it is evident that those strains which mate with one another (i.e. strains within the same 'variety') have similar base compositions. Thus strains of variety 1 ..., which are freely intercrossed, have similar mean GC content.

Whether Sueoka's "strains" or "varieties" were in fact "species" is uncertain. But relative uniformity of (G+C)% is now known to apply to a wide variety of species. Each species can be considered to have a specific value, but there is often a distribution about this value that reflects genome fine-sectoring into regions (homostability regions or isochores) of lower or higher (G+C)% (see Chapter 11).

The Holy Grail

It seems that, in identifying (G+C)% as the component of the base composition that varies between species, Chargaff may unknowingly have uncovered the 'holy grail' of speciation postulated by Romanes and elaborated by Bateson (Chapter 9) and by Goldschmidt (Chapter 8). Romanes had drawn attention to what we would now call non-genic variations (germ-line mutations that usually do not affect gene products). Manifest as the sterility of hybrid offspring, these cryptic variations would tend to isolate a parent reproductively from most members of the species to which its close ancestors had belonged, but not from members that had undergone the same non-genic variation; with these it would be reproductively compatible producing fertile offspring.

Romanes held that, in the general case, this isolation was an essential *precondition* for the preservation of characters (gene-encoded) that were distinctive of a new species. His holy grail (speciating factor) was an "intrinsic peculiarity" of the reproductive system. Bateson's was an abstract "residue." Goldschmidt's was a chromosomal "pattern" caused by "systemic mutations." Variations in (G+C)% would satisfy these postulates. Indeed, in bacteria, which when so inclined can transfer DNA in a sexual fashion [3], (G+C)% differences appear early in speciation [4], as with Sueoka's ciliates.

Language

Spoken and written languages provide a metaphor for information in DNA (Chapter 3). A common language brings people together, so is conducive to sexual relations. This is more so when accents are compatible. When accents are incompatible (i.e. there is a linguistic barrier), communication, and hence reproductive potential, is impaired. The distribution of some quantitative measure of accent (linguistic as in the cockney accent, or (G+C)% as in DNA) is shown in Figure 10.1.

The bell-curve shows that most individuals are of accent type B. They are the majority type and therefore are most likely to encounter each other. So B-with-B reproductive interactions to produce fertile B-type offspring are the norm. Much rarer are individuals of type A on one limb of the curve. It is most likely that A-types will meet only the abundant majority type – type B – with which, by virtue of the differences in (G+C)% accent, they are likely to be reproductively incompatible (i.e. if offspring are produced they are sterile). Very rarely, an A-type will meet an A-type of the opposite sex and reproduction to produce fertile A-type offspring can occur (Fig. 10.2).

This will usually not recur because of the rarity of type A. If however there is a concomitant genic mutation in an A type, leading to an advantageous phenotypic adaptation, that adaptation will be protected from blending effects because of reproductive isolation from the main population. A can then become more abundant (i.e. it will be positively selected) and so the probability of meeting more A types, to perpetuate the new line, increases. An ‘incipient’ species can arise (Fig. 10.3).

The main-line of organisms (long gray column) proceeds in linear, unbranching fashion. On rare occasions a potential new line can briefly maintain itself by reproduction, and hence expand in numbers (short gray column). However, under usual conditions the maintenance of A-types in the population depends largely on their continuous formation through new mutations rather than because, by chance, they

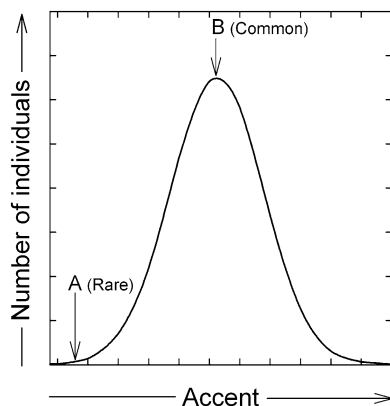


Fig. 10.1 Distribution of an entity – ‘accent’ – that is subject to variation within a population. In spoken language this could be the extent to which H’s are dropped (omitted). In DNA language this could be (G+C)%. A is a rare type. B indicates the population norm. We consider here just these two types (subsets), but actually there are many intermediates (i.e. variation is continuous)

Fig. 10.2 When rare types A reproduce with their like, in isolation from B, then their frequency in the population can increase. A discontinuity arises and a separate bell curve appears

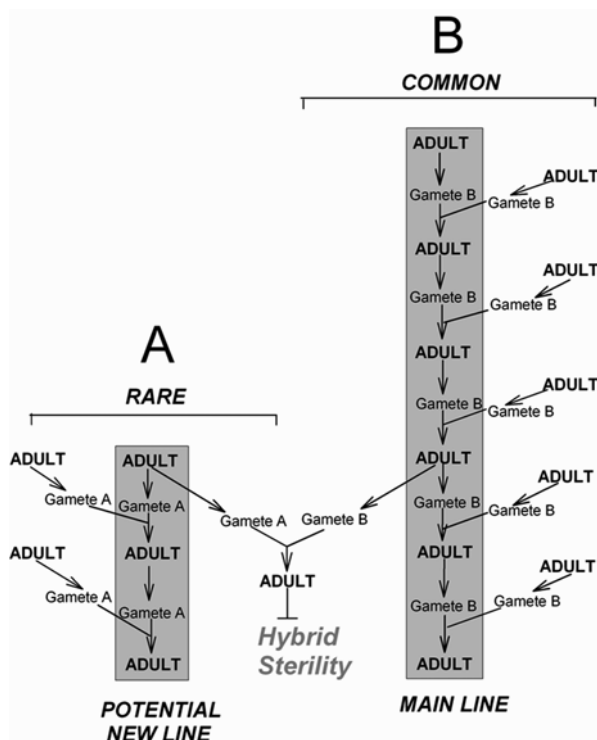
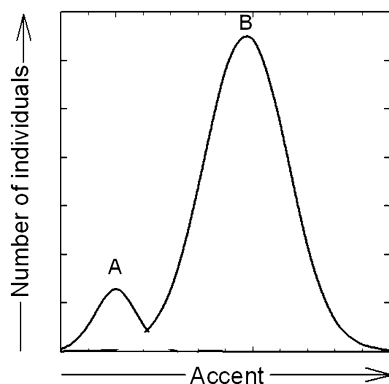


Fig. 10.3 Reproductive isolation (due to hybrid sterility) between the population norm (type B) and a rare type (A). During the course of linear unbranching evolution type B meets type B and produces B type gametes which convey the B accent through the generations (tall gray column). On rare occasions common type B meets rare type A. The resulting child grows into an adult, but the adult is sterile, so the line ends. Parental types A and B are reproductively isolated. On rare occasions, there is linear unbranching evolution when rare type A meets another rare type A (short gray column). Propitious mutations now escape blending with type B and population A increases

succeed in reproducing themselves. Reproduction of the A-types will best occur when there are concomitant mutations in genes that confer characters that are favored by natural selection. These provide selective footholds that, by virtue of the *pre-existing* reproductive isolation, cannot be thwarted by blending with B-types. Thus, *reproductive isolation empowers natural selection*.

Crick's Unpairing Postulate

At the molecular level, we see similar forces acting at the level of meiosis – the dance of the chromosomes. In the gonad paternal and maternal chromosomes (homologs) align (Fig. 9.6). The early microscopists referred to this as “conjugation.” If there is sufficient sequence identity (i.e. the DNA ‘accents’ match), then the band plays on. The chromosomes continue their minuet, progressing through various check-points and gametes are formed. If there is insufficient identity (i.e. the DNA ‘accents’ do not match) then the music stops. Meiosis fails, gametes are not formed, and the child is sterile – a ‘mule.’ Thus, the parents of the child (their “hybrid”) are reproductively isolated *from each other* (i.e. unable to generate a line of descendents due to hybrid sterility), but not necessarily from other members of their species. At least one of the parents has the potential to be a founding member of a new species, *provided* it can find a mate with the *same* DNA ‘accent.’ Differences in (G+C)% have the potential to initiate the speciation process creating first “incipient species” with partial reproductive isolation, and then “species” that, by definition, are fully reproductively isolated. To see how this might work, we consider the chemistry of meiotic chromosome alignment.

In 1922 Muller suggested that the pairing of genes as parts of chromosomes undergoing meiotic synapsis in the gonad might provide clues to gene structure and replication [5]:

It is evident that the very same forces which cause the genes to grow [duplicate in parallel] should also cause like genes to attract each other [pair] If the two phenomena [duplication and pairing] are thus dependent on a common principle in the make-up of the gene, progress made in the study of one of them should help in the solution of the other.

In 1954 he set his students an essay “How does the Watson-Crick model account for synapsis?” [6]. The starting point was a model that had the two DNA strands ‘inward-looking’ (i.e. the bases on one strand were paired with the bases on the other strand; Fig. 2.1). Crick took up the challenge in 1971 with his “unpairing postulate” by which the two strands of a DNA duplex would *unpair* to expose free bases in single-stranded bubble-like regions [7]. This would allow a search for sequence similarity (homology) between two chromosomes (i.e. between two independent parallel duplexes). This was later referred to as “structure-mediated homology recognition” [8]. In principle, no other factors (e.g. proteins) would be needed. Indeed, there is some, albeit indirect, evidence consistent with the recognition occurring within cells without participation of gene products [9, 10]. This is not to say that other factors did not get into the act at some point after the fundamental pairing process had evolved.

Structure-Mediated Homology Recognition

Homologous recombination occurs between DNA sequences which are identical, or very closely so. It is easy to imagine – and many textbook diagrams support the idea – that one strand of a DNA duplex is cut (“cut first model”) and the liberated single-strand seeks a complementary strand in another DNA duplex, the strands of which obligingly unpair. However, certainly in fruit fly and nematode worms [11], and in mice [12], homology recognition and synapses do not require initial strand breakage. Indeed, be they natural or artificial, homology recognition can occur between intact DNA duplexes in a simple salt solution in the absence of proteins [13, 14], which hence might not be critical within intact cells [15]. Structural studies suggest that “the DNA duplex is labile on a millisecond time scale, allowing local bubble formation” (i.e. strand separation without breakage). Thus, the “alignment of DNA molecules of identical sequence and length could be stabilized by such perturbations occurring *simultaneously (and transiently)* at identical positions in the sequence” [16].

But even on short time scales the ‘bubbles’ would be unlikely to remain single stranded. The crowded cytosol (see Chapter 15) favors higher ordered nucleic acid structures. If not prevented by single-strand binding proteins or unusual sequence characteristics, the individual extruded single-stranded DNA bubbles should quickly adopt folded stem-loop configurations. In studies of pairing between single-stranded RNA molecules (Chapter 6), it was found that complementary molecules must first recognize each other, prior to hybridization, by exploratory ‘kissing’ interactions (base pairing) between the loops of such stem-loop structures. In 1993 Kleckner and Weiner [17] proposed that this might also apply to the meiotic recognition of palindromic DNA sequences that would be extruded as stem-loops from DNA duplexes (Fig. 5.2). Initial loop-loop interactions would progress to the annealing of complementary strands (“pair first model”), thus setting the stage for subsequent strand breakage and recombination.

From this perspective, some long-standing recombination models, classified as “paranemic” (no initial strand breakage), can now be seen as having much to commend them [18–21]. The rate-limiting step in meiotic recombination may be, not the actual *pairing* of homologous DNA strands, but the *extrusion* of appropriate stem-loop structures from duplex DNA. Indeed, an adaptive advantage of stem-loops is that they stabilize partially unfolded DNA, thus greatly accelerating the extrusion process [22].

An example of how structure-mediated homology recognition might work is shown in Figure 10.4. The extrusion from an individual duplex is normally symmetrical, affecting both strands equally and, as we shall see below, is critically dependent on base composition. For meiotic alignment, maternal and paternal chromosomal homologs should mutually explore each other and *test* for ‘self’ DNA complementarity by the ‘kissing’ mechanism noted in Chapter 6. The sequences do not *commit* themselves, by incurring strand-breakage, until a degree of complementarity has been recognized.

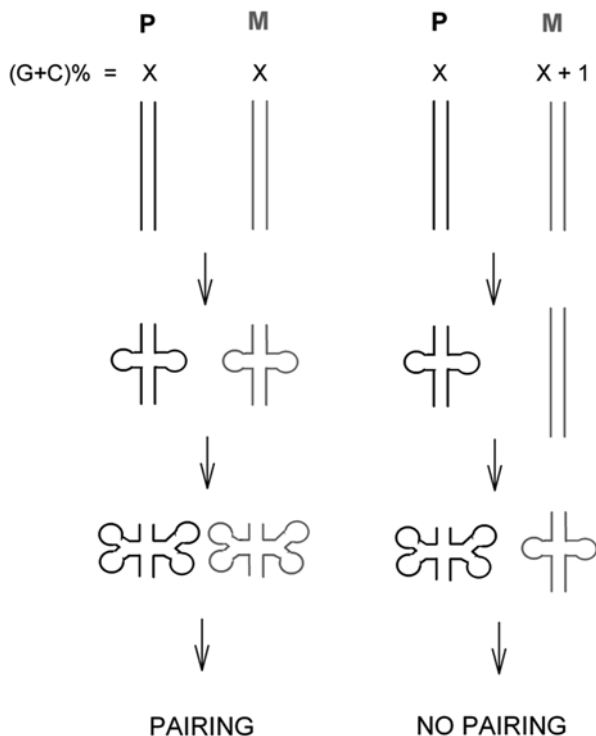


Fig. 10.4 The exquisite sensitivity of stem-loop extrusion from duplex DNA to differences in base composition can prevent the initiation of pairing between nearly homologous DNA sequences. At the left, paternal (P) and maternal (M) duplexes have the same $(G+C)\%$ value (X). As negative supercoiling (torsion) progressively increases, the strands of each duplex synchronously open to allow formation of equivalent stem-loop secondary structures so that ‘kissing’ interactions between loops can progress to pairing. At the right, paternal and maternal duplexes differ slightly in $(G+C)\%$ (X , and $X + 1$). The maternal duplex of higher $(G+C)\%$ opens less readily as torsion increases, so strand opening is not synchronous, ‘kissing’ interactions fail, and there is no progress to pairing. In this model, chromosome pairing occurs *before* the strand breakage that accompanies recombination (not shown). Even if strand breakage were to occur first (as required by some models), unless inhibited by single-stranded DNA-binding proteins, the free single strands so exposed would rapidly adopt stem-loop conformations. So the homology search could still involve ‘kissing’ interactions between the tips of loops

The mechanism is essentially the same as that by which tRNA anticodon loops recognize codons in mRNAs (Fig. 5.1), except that the stem-loop structures first have to be extruded from DNA molecules that would normally be in classical duplex form [23]. In all DNA molecules examined, base-order supports the formation of such secondary structures (Chapter 5). If sufficient complementarity is found between the sequences of paternal and maternal chromosome homologs (i.e. the genomes are “reprototypically” compatible), then crossing over and recombination can occur (i.e. the ‘kissing’ can be ‘consummated’).

The main adaptive values of this would be the proper assortment of chromosomes among gametes, and the correction of errors in chromosome sequences (see Chapter 17). A genome which has diverged from others in its species may no longer be a reliable template for error-correction. As such it must be excluded from recombination with other members of the species, but when so excluded the deviant genome then becomes a candidate for an incipient speciation event. 'Kissing' turns out to be a powerful metaphor, since it implies an exploratory interaction that may have positive or negative reproductive consequences.

The model predicts that, for preventing recombination (i.e. creating reproductive isolation), a non-complementarity between the sequences of potentially pairing strands, in itself, is *less* important than a non-complementarity associated with sequence differences that change the pattern of stem-loops [24]. As we shall see next, this implies differences in the *quantities* of members of the Watson-Crick base pairs in single strands (i.e. a parity difference). This is because parity between these bases would be needed for optimum stem formation (Chapter 4). Parity differences should correlate with differences in stem formation, and hence, with different stem-loop patterns.

(G+C)% Controls Pairing

What role could the (G+C)% 'accent' play in meiotic pairing? From calculated DNA secondary structures, it has been inferred that small fluctuations in (G+C)% have great potential to affect the extrusion of stem-loops from duplex DNA molecules and, hence, to affect the pattern of loops which would then appear (Fig. 5.2). A small difference in (G+C)% (reprotypic difference) would mark as 'not-self' a DNA molecule that was attempting to pair meiotically with another DNA ('self'). This would impair the kissing interaction with the DNA, and so would disrupt meiosis and allow divergence between the two parental lines, thus possibly initiating a speciation event.

The total stem-loop potential in a sequence window can be analyzed quantitatively in terms of the relative contributions of base composition and base order, of which base *composition* plays a major role (Chapter 5). Of the various factors likely to contribute to the base composition-dependent component of the folding energy of an extruded single stranded DNA sequence, the four simplest are the quantities of the four bases.

Two more complex factors are the individual bases from each potential Watson-Crick base pair that are present in lowest amounts. For example, if the quantities of A, G, C and T in a 200 nucleotide sequence window are 60, 70, 30 and 40, respectively, then what may be referred to as " $\text{AT}_{\min}$ " would be 30, and the corresponding " $\text{GC}_{\min}$ " would be 30. These numbers would reflect the upper limit on the number of base pairs that could form stems; the quantity of the Watson-Crick pairing partner that was *least* would place a limit on the possible number of base pairs. This value might be expected to correlate *positively* with folding stability.

Conversely, the excess of bases without a potential pairing partner (in the above example $A \cdot T = 20$ and $G \cdot C = 40$) might provide an indication of the maximum number of bases available to form loops. Since loops tend to destabilize stem-loop structures, these “Chargaff difference” values might be expected to correlate *negatively* with folding stability.

Although the bases are held in linear order, a vibrating single-stranded DNA molecule has the potential to adopt many structural conformations with Watson-Crick interactions occurring between widely separated bases. Accordingly, pairing can also be viewed as if the result of random interactions between free bases in solution. This suggests that the two products of the quantities of pairing bases could be important (60×40 , and 70×30 , in the above example). The products would be maximal when pairing bases were in equal proportions in accordance with Chargaff’s PR2. Thus, $[50 \times 50] > [60 \times 40] > [70 \times 30]$). The corresponding product values ($2500 > 2400 > 2100$) provide an index both of the absolute quantities of the members of a Watson-Crick base pair, and of their relative proportions.

In an attempt to derive formulae permitting prediction of folding energy values directly from the proportions of the four bases, Jih-H. Chen and his colleagues [25] examined the relative importance of eight of the above ten factors in determining the base composition-dependent component of the folding energy (FORS-M; Chapter 5). These factors were A , G , C , T , AT_{min} , CG_{min} , $A \times T$, and $G \times C$ (where A , C , G , and T refer to the quantities of each particular base in a sequence window). The products of the quantities of the Watson-Crick pairing bases ($A \times T$, and $G \times C$) were found to be of major importance, with the coefficients of $G \times C$ (the strongly interacting S bases), greatly exceeding those of $A \times T$ (the weakly interacting W bases). Less important were AT_{min} and CG_{min} , and the quantities of the four bases themselves. All ten parameters were examined in an independent study [26], which confirmed the major role of the product of the quantities of the S bases in a segment (Table 10.1) .

Table 10.1 Relative abilities of ten base composition-derived parameters to predict the base composition-dependent component of the folding energy of single-stranded DNA. Best predictors are marked as plus. When only one predictor is allowed, the product of G and C has a correlation of 0.92, as assessed by the r^2 value (see Appendix A). Adding further predictors (rows 2, 3 and 4) does not greatly improve on this value

Number of predictors	Individual bases				Base products		Minimum bases		Chargaff differences		r^2
	A	T	C	G	$A \times T$	$G \times C$	AT_{min}	GC_{min}	$A - T$	$G - C$	
1	-	-	-	-	-	+	-	-	-	-	0.92
2	-	-	+	-	-	+	-	-	-	-	0.96
3	-	-	+	+	-	+	-	-	-	-	0.98
4	-	-	+	+	-	+	-	-	+	-	0.98

Of particular importance is that it is not the absolute quantities of the **S** bases, but the *product* of the multiplication of these absolute quantities. This should *amplify* very small fluctuations in **(G+C)%**, and so should have a major impact on the folding energy of a segment and, hence, on the pattern of stem-loops extruded from the duplex DNA in a chromosome engaged in a 'kissing' homology search for a homologous chromosome segment.

If stem-loops are of critical importance for the initiation of pairing between segments of nucleic acids at meiosis, then differences in **(G+C)%** could strongly influence the establishment of meiotic barriers, so leading to speciation. But barriers may be transient. Having served its purpose, an initial barrier may be superseded later in the course of evolution by a more substantial barrier (Figure 8.4). In this circumstance, evidence for the early transient barrier may be difficult to find. However, in the case of different, but related, virus species (allied species) that have the potential to coinfect a *common* host cell, the original **(G+C)%** barrier may have been retained.

Mutational Meltdown

Modern retroviruses, such as those causing AIDS (HIV-1) and human T cell leukemia (HTLV-1), probably evolved by divergence from a common ancestral retrovirus. Branching phylogenetic trees linking the sequences of modern retroviruses to such a primitive retroviral 'Eve' are readily constructed, using either differences between entire sequences, or just **(G+C)%** differences [27]. The fewer the differences, the closer are two species on such trees.

Unlike most other virus groups, retroviruses are diploid. As indicated in Chapter 2, diploidy entails a considerable redundancy of information, a luxury that most viruses cannot afford. They need compact genomes that can be rapidly replicated, packaged and dispersed to new hosts. However, different virus groups have evolved different evolutionary strategies.

The strategy of retroviruses is literally to mutate themselves to the threshold of oblivion ('mutational meltdown'), so constituting a constantly moving target that the immune system of the host cannot adapt to. To generate mutants, retroviruses replicate their nucleic acids with self-encoded enzymes (polymerases) that do not have the error-correcting ('proof-reading') function that is found in the corresponding enzymes of their hosts. Indeed, this is the basis of AIDS therapy with AZT (azidothymidine), which is an analogue of one of the nucleotide building blocks that are joined together (polymerized) to form linear nucleic acid molecules ("polymers;" Chapter 2). AZT is recognized as foreign by host polymerases, which eject it. But retroviral polymerases cannot discriminate, and levels of mutation (in this case termination of the nucleic acid sequence) attain values above the oblivion threshold ('hypermutation') from which it is impossible to recover ('error catastrophe'). Below the threshold, there is a most effective mechanism to counter mutational damage.

The retroviral counter-mutation strategy *demand*s that *two* complete single-strand retroviral RNA genomes be packaged in each virus particle (i.e. diploidy). Each of these genomes may be severely mutated but, since mutations occur randomly, there is a chance that each genome will have mutations at different sites. Thus, in the next host cell there is the possibility of recombination (cutting and splicing) between the two genomes to generate a new genome with many less, or zero, mutations [28].

The copackaging of the two genomes requires a process analogous to meiotic pairing. On each genome a “dimer linkage” nucleotide sequence folds into a stem-loop structure. “Kissing” interactions between the loops precede the formation of a short length of duplex RNA (Fig. 6.8), so that the two genomes form a dimer. This allows their collective packaging and, in the next host, recombination can occur.

Coinfecting Viruses

What if two diploid viruses both infected the *same* host cell, thus releasing four genomes into an environment conducive to recombination? In many cases this would be a most favorable circumstance, since there would now be four damaged genomes from which to regenerate, by repeated acts of recombination, an ideal genome. Thus, it would seem maladaptive for a virus with this particularly strategy to evolve mechanisms to prevent entry of another virus (‘superinfection’) into a cell that it was occupying, at least in the early stages of infection [29].

This presupposes that a co-infecting virus will be of the *same* species as the virus which first gained entry. However, HIV-1 and HTLV-1 belong to allied, but distinct, retroviral species. They have a common host (humans) and common host cell (known as the CD4 T-lymphocyte). When in the course of evolution these two virus species first began to diverge from a common ancestral retroviral species, a barrier to recombination had to develop *as a condition of successful divergence*. Yet, these two virus types needed to retain a common host cell in which they had to perform similar tasks. This meant that they had to retain *similar genes*. Many similar gene-encoded functions are indeed found. But similar genes implies similar sequences, and similar sequences implies the possibility of recombination between the two genomes. Thus, co-existence in the same host cell could result in the viruses destroying each other, as distinct species members, by mutually recombining (shuffling their genomes together) to produce blended genomes.

Without a recombination barrier *each virus was part of the selective environment of the other*. This should have provided a pressure for genomic changes that, while not interfering with conventional phenotypic functions, would protect each virus against recombination with the other type. If (G+C)% differences could create such a recombination barrier (while maintaining, through choice of appropriate codons, the abilities to encode similar amino acid sequences), then such differences would be selected for. When we examine the (G+C)% values of each of these species there is a remarkable difference. HIV-1 is one of the lowest (G+C)% species known (i.e. it is AT-rich).

HTLV-1 is one of the highest (G+C)% species known (i.e. it is GC-rich). This might be regarded as just a remarkable coincidence save for the fact that, in some *other* situations where two viruses from different but allied species occupy a common host cell, there are also wide differences in (G+C)% [30, 31]. As set out above, these (G+C)% differences – differences in reprototype rather than in phenotype – should alone suffice to prevent recombination. Indeed, they can provide an *early warning of a newly emerging species of potentially pathogenic virus* (Chapter 7).

So we see that different species of intracellular pathogens that can occupy a common host cell can come to display *wide* (G+C)% differences. Likewise, whereas an ancestral human malarial parasite probably had a low (G+C)% value, today we find *Plasmodium falciparum* (23.8%) and *Plasmodium vivax* (46.2%). It is possible that the need for a recombination barrier drove a (G+C)% increase in the latter [32].

Between these intracellular pathogens, there is the equivalent of the hybrid sterility barrier to reproduction, and there has been little progression to processes analogous to prezygotic isolation and hybrid inviability (Chapter 8). Although rare, this absence of supervening barriers can sometimes be found between normal species. Indeed, two species of fern that diverged from a common ancestor 60 million years ago are still capable of forming a hybrid, albeit a sterile one, that can only propagate vegetatively [33]. At the time of this writing, their genomic (G+C)% values are not available. However, data from certain polyploids are more informative.

Polyploidy Cures Hybrid Sterility

The plant which gives us tobacco, *Nicotiniana tabacum*, is a tetraploid which emerged some six million years ago when the two diploid genomes of *Nicotiniana sylvestris* and *Nicotiniana tomentosiformis* appeared to fuse. *Nicotiniana tabacum* is designated an *allotetraploid* (rather than an *autotetraploid*) since the two genomes were from different source species (Greek: *allos* = other; *autos* = same). The two species are estimated to have diverged from a common ancestral species 75 million years ago. As allied species, they should have retained some sequence similarities; so within a common nucleus in the tetraploid there should have been ample opportunity for recombination between the two genomes. *Yet, the genomes have retained their separate identities*. This can be shown by backcrossing to the parental types. Half the chromosomes of the tetraploid pair at meiosis with chromosomes of one parent type; the other half with those of the other parent. Thus, recombination of the half the chromosomes of the tetraploid with those of one parent is in some way prohibited. In 1940 Goldschmidt noted [34]:

Clausen ... has come to the conclusion that *N. tabacum* is an allotetraploid hybrid, one of the genomes being derived from the species *sylvestris*, the other from *tomentosa*. By continuous backcrossing to *sylvestris* the chromosomes derived from *sylvestris* can be tested because they form tetrads with the *sylvestris* chromosomes. They have been found to be completely different genetically [from *tomentosa*]. The idea of reaction system [pattern] thus becomes less generalized [less nebulous] and actually applies to the whole architecture of individual chromosomes.

As noted in Chapter 9, the separateness of the genomes in allotetraploids is reflected in meiosis where, although there are four potentially pairing copies of each chromosome (four homologous chromosomes), each parental type only pairs with its *own* type. The two sets of chromosomes in the tobacco plant differ in a key sequence characteristic, their **(G+C)%** [35].

In this case, the distinctiveness of the two genomes is apparent. But analysis of many genomes has shown that whole genome duplications of ancient origin are widespread. Many plants and animals are covert polyploids, a feature disguised by the subsequent shuffling of DNA segments and loss of many duplicated genes. It is hard to imagine how polyploids would have survived to perform these molecular gymnastics if they had derived from an *intraspecies* duplication of a diploid to produce a tetraploid with genomes of very close **(G+C)%** values (e.g. autopolyploidy). Each set of four homologous chromosomes would have become irremediably tangled at meiosis.

On the other hand, diploids of different species can produce an allotetraploid not only differing in appearance and function from the parental lines, but also *immediately* reproductively isolated from those lines. A new species may then seem to ‘jump’ into existence – a saltation. This was reported for a plant (evening primrose) early in the twentieth century by Hugo de Vries (Chapter 1). He thought, incorrectly, that he had discovered a general path to new species. In 1917, “the father of yeast genetics.” Öjvind Winge [36], made a suggestion that is now supported by studies of yeast polyploids. Two yeast species (haploid) can cross to produce a diploid that is unable to generate gametes. Because of this hybrid sterility the line cannot continue. However, the sterility is ‘cured’ if the chromosome number (diploid) is doubled (to become tetraploid). Then each paternal chromosome can pair with its paternal homolog, and each maternal chromosome can pair with its maternal homolog. There is no tangling of chromosomes. The ‘cure’ demonstrates *a selective advantage for the tetraploid state*. The whole genome duplication, albeit subject to some remodeling, will have an opportunity to persist [37].

Protein Pressure Reinforces

Under reprotypic selection pressure the overall **(G+C)%** in an organism will begin to change and this, in turn, may provoke further adaptations. When, without changing the amino acid, a **GC**-rich codon adapts to become **AT**-rich, then a different tRNA with an anticodon that matches the synonymous codon may be employed more frequently. If the efficiency or accuracy of mRNA translation is not rate-limiting then this may not matter. Indeed, for some mRNAs the rate of *initiation* of translation is rate-limiting and the speed of ribosomal movement along the mRNA, as influenced by the availability of appropriate tRNAs, is largely irrelevant [38]. But for other mRNAs, such as those encoding ‘emergency’ heat-shock proteins, speed of protein synthesis should be critical. In this case there will be a pressure for regulatory mutations that will change tRNA populations to better match the new

synonymous codons appearing in the first mRNAs whose genes have succumbed to reprotypic **GC**-pressure. Over time, other genes, for which speed of protein synthesis is less critical, will then 'come into line' by accepting mutations that create codons that match the changed tRNA populations. Thus selection on codons "appears to be consistently acting in the same direction as the nucleotide substitution bias of genomes," so that the "identity of favored codons tracks the **GC** content of the genomes" [39]. In other words, reprotypic pressure on **(G+C)%** will be *reinforced* by pressures operating at the translational level.

Higher K-mer Differences as Ultimate Drivers

A high **(G+C)%** DNA segment is likely to contain higher order k-mers enriched in these bases – e.g. GCG, CGG, etc., rather than ATA, TAA, etc.. These k-mers are likely to have determined the **(G+C)%** values, rather than the converse (Chapter 4). Thus, while it has been convenient here to deal solely in terms of mononucleotide frequencies (base compositions), the actual driving force for recombination-related phenomena may be at the oligonucleotide level. Consistent with this, Gladyshev and Kleckner [23] postulate 3-mers as the “minimal effective homology recognition unit” required for DNA “break-independent pairing” between DNA segments in bread mould (*Neurospora crassa*). Furthermore, Lee and coworkers find periodically arranged 3-mers to drive recombination in bacteria and yeast [40]. Likewise pairing of ‘RNA antibodies’ with complementary duplex DNAs (Chapter 15) involves periodically arranged 5-mers [41].

So back to Sueoka [2]. We can now identify, not **(G+C)%**, but higher k-mers, as *making* a protein accept an amino acid ‘*against its will*.’ Indeed, shifts in genome (nucleotide) and proteome (amino acid) compositions are coupled, with DNA base compositions showing much greater variation between species than amino acid compositions. In 1959, prior to the discovery of mRNA, this puzzled Crick [42]:

This large variation of DNA composition is very unexpected. The abundance of the various amino acids does not, as far as we know, vary much from organism to organism; leucine is always common, methionine usually rather rare. ... the large variation reported for DNA needs some special explanation.

That “special explanation” is now at hand. Using statistical approaches analogous to that used for Table 10-1, Fran Supek and his colleagues found that species oligomer (higher k-mer) frequencies could predict species amino acid compositions with higher accuracy than species monomer (1-mer) frequencies (base compositions); this left only 9% of the variation between species unexplained (see Appendix A). They concluded that “evolutionary shifts in overall amino acid composition appear to occur almost exclusively through factors shaping the global oligonucleotide content of the genome.” Thus, it is oligonucleotide composition (frequency of certain oligonucleotides), acting as an *internal* driving force, distinct from conventional, externally-imposed, natural selection, that leads the way [43]. As we shall see (Chapters 11, 12), this contrasts with the conventional wisdom

that it is amino acid, rather than nucleotide, compositions, that mostly reflect environmental pressures (e.g. high temperatures). It is by varying the *order*, rather than the compositions, of their amino acids, that proteins seem best to adapt to serve the organisms that contain them.

Saltum?

Darwin held that biological evolution reflected the accumulation of many very small variations, rather than few intermittent large variations. That Nature did not work by means of large jumps was encapsulated in the Latin phrase “*Natura non facit saltum*.” However, Huxley (Chapter 9), while supporting many of Darwin’s teachings, considered it more likely that evolution had proceeded in jumps (“*Natura facit saltum*”). According to the arguments of this chapter, both are correct. Within some members of a species small variations in the genome phenotype (i.e. variations in (G+C)% and the related k-mers) accumulate below the ‘radar screen’ of natural selection. Thus, these members become progressively more reproductively isolated from most other members of the species, initially *without* major changes in the conventional phenotype. As it accrues, reproductive isolation increasingly favors rapid change in the conventional phenotype, often under the influence of natural selection. So, when their appearance is viewed on a *geological* time scale, new species can seem to ‘jump’ into existence.

The rate increase reflects better preservation of frequent phenotypic micromutations rather than of infrequent phenotypic macromutations (i.e. of “hopeful monsters,” to use an unfortunate term of Goldschmidt). While there is *continuity* of variation at the genotype level, as far as speciation is concerned variants (mutant forms) seem to emerge *discontinuously* at the phenotype level. Organisms with phenotypic macromutations are not the stuff of evolution. Being infrequent, they are unlikely to find a member of the opposite sex with the same change. Thus, they cannot continue the line.

Summary

Single strands extruded from duplex DNA have the potential to form stem-loop structures which, in turn, support further extrusion. Through exploratory loop-loop ‘kissing’ interactions, these structures are involved in the homology search that precedes recombination. The total stem-loop potential in a sequence window can be analyzed quantitatively in terms of the relative contributions of base composition and base order, of which base composition, and particularly the *product* of the two S bases ($G \times C$), plays an important role. Very small differences in parental (G+C)% values should impair meiotic pairing in their hybrid offspring, so impeding gamete production. Because their hybrid is sterile, the parents are, in an evolutionary sense,

reproductively isolated from each other – a state that favors branching speciation. (G+C)% differences (1-mer differences) are an expression of more fundamental oligonucleotide (k-mer) differences, which can be so powerful as to overrule the dictates of natural selection in determining the amino acid compositions of proteins. Thus, Chargaff's species-dependent component of base composition, (G+C)%, may be the 'holy grail' responsible for reproductive isolation (non-genic), as postulated by Romanes, Bateson and Goldschmidt. Many of the old disputes among evolutionists (e.g. *saltum* versus *non-saltum*) now appear reconcilable.

References

1. Chargaff E (1951) Structure and function of nucleic acids as cell constituents. *Federation Proceedings* 10:654–659
2. Sueoka N (1961) Compositional correlations between deoxyribonucleic acid and protein. *Cold Spring Harbor Symposium on Quantitative Biology* 26:35–43
3. Gratia JP, Thiry M (2003) Spontaneous zygogenesis in *Escherichia coli*, a form of true sexuality in prokaryotes. *Microbiology* 149:2571–2584
4. Bellgard M, Schibeci D, Trifonov E, Gojobori T (2001) Early detection of G + C differences in bacterial species inferred from the comparative analysis of the two completely sequenced *Helicobacter pylori* strains. *Journal of Molecular Evolution* 53:465–468
5. Muller HJ (1922) Variation due to change in the individual gene. *American Naturalist* 56:32–50
6. Carlson EA (1981) *Genes, Radiation and Society. The Life and Work of H. J. Muller*, Cornell University Press, Ithaca, NY, p 390
7. Crick F (1971) General model for chromosomes of higher organisms. *Nature* 234:25–27 [Robert Olby (2009) in *Francis Crick. Hunter of Life's Secrets* (Cold Spring Harbor Laboratory Press, New York, p. 359) considers this paper “now only of historical interest,” perhaps because key points were made in figure legends. In his submission letter to the Editor (Sept. 3 1971), Crick noted that “the legends to the figures are as long as the main body of the paper, but I am convinced this is the appropriate way to present it. ... You usually put the legends in small print, but Sydney [Brenner] has suggested to me that in this case you break your rule and print them the size used in the text to make for easier reading, since so many important points are contained in them.”]
8. Baldwin GS, et al. (2008) DNA double helices recognize mutual sequence homology in a protein free environment. *Journal of Physical Chemistry B Letters* 112:1060–1064
9. Bateman JR, Wu CT (2008) A genome wide survey argues that every zygotical gene product is dispensable for the initiation of somatic homolog pairing in *Drosophila*. *Genetics* 180:1329–1342
10. Blumenstiel JP, Fu R, Theurkauf WE, Hawley RS (2008) Components of the RNAi machinery that mediate long-distance chromosomal associations are dispensable for meiotic and early somatic homolog pairing in *Drosophila melanogaster*. *Genetics* 180:1355–1365
11. Moore G, Shaw P (2009) Improving the chances of finding the right partner. *Current Opinion in Genetics & Development* 19:99–104
12. Boateng KA, Bellani MA, Gregoretti IV, Pratto F, Camerini-Otero RD (2013) Homologous pairing preceding SPO11-mediated double-strand breaks in mice. *Developmental Cell* 24:196–205
13. Kornyshev AA, Wynveen A (2009) The homology recognition well as an innate property of DNA structure. *Proceedings of the National Academy of Sciences USA* 106:4742–4746
14. Danilowicz C, et al. (2009) Single molecule detection of direct, homologous, DNA/DNA pairing. *Proceedings of the National Academy of Sciences USA* 106:19824–19829

15. Wang X, Zhang X, Mao C, Seeman NC (2010) Double-stranded DNA homology produces a physical signature. *Proceedings of the National Academy of Sciences USA* 107:12547–12552
16. Inoue S, Sugiyama S, Travers AA, Ohya T (2007) Self-assembly of double-stranded DNA molecules at nanomolar concentrations. *Biochemistry* 46:164–171
17. Kleckner N, Weiner BM (1993) Potential advantages of unstable interactions for pairing of chromosomes in meiotic, somatic and premeiotic cells. *Cold Spring Harbor Symposium in Quantitative Biology* 58:553–565
18. Sobell HM (1972) Molecular mechanism for genetic recombination. *Proceedings of the National Academy of Sciences USA* 69:2483–2487
19. Wagner RE, Radman M (1975) A mechanism for initiation of genetic recombination. *Proceedings of the National Academy of Sciences USA* 72:3619–3622
20. Doyle GG (1978) A general theory of chromosome pairing based on the palindromic DNA model of Sobell with modifications and amplification. *Journal of Theoretical Biology* 70:171–184
21. Wilson JH (1979) Nick-free formation of reciprocal heteroduplexes: a simple solution to the topological problem. *Proceedings of the National Academy of Sciences USA* 76:3641–3645
22. Schreck JS, Ouldrige TE, Romano F, Šulc P, Shaw L, Louis AA, Doye JPK (2015) DNA hairpins primarily promote duplex melting rather than inhibiting hybridization. *Nucleic Acids Research* 43:6181–6190
23. Gladyshev E, Kleckner N (2014) Direct recognition of homology between double helices of DNA in *Neurospora crassa*. *Nature Communications* 5:3509
24. Forsdyke DR (2007) Molecular sex: the importance of base composition rather than homology when nucleic acids hybridize. *Journal of Theoretical Biology* 249:325–330
25. Chen J-H, Le S-Y, Shapiro B, Currey KM, Maizel JV (1990) A computational procedure for assessing the significance of RNA secondary structure. *CABIOS* 6:7–18
26. Forsdyke DR (1998) An alternative way of thinking about stem-loops in DNA. A case study of the human *GOS2* gene. *Journal of Theoretical Biology* 192:489–504
27. Bronson EC, Anderson JN (1994) Nucleotide composition as a driving force in the evolution of retroviruses. *Journal of Molecular Evolution* 38:506–532
28. Levy DN, Aldrovandi GM, Kutsch O, Shaw GM (2004) Dynamics of HIV-1 recombination in its natural target cells. *Proceedings of the National Academy of Sciences USA* 101:4204–4209. [HIV is an extreme example of the ability of viruses to escape *innate* immune defenses, in that there are few external viral features that remain constant from generation to generation. Thus, it is the *adaptive* immune system that is primarily challenged (see Part V). In contrast, many species of bacterium have characteristic surface features (e.g. sugars), and potential hosts have recognition molecules (e.g. lectins) the information for which is in the germ-line (*innate* immunity). At an early stage lectins ‘batter’ bacteria to facilitate their ingestion by phagocytic cells (see “Almroth Wright, opsonins, innate immunity and the lectin pathway of complement activation: a historical perspective.” Forsdyke 2016; *Microbes & Infection*, in press.)]
29. Michel N, Allespach I, Venzke S, Fackler OT, Keppler OT (2005) The nef protein of human immunodeficiency virus established superinfection immunity by a dual strategy to downregulate cell-surface CCR5 and CD4. *Current Biology* 15:714–723 [The authors state: “Primate lentiviruses appear to have evolved time windows during which the permission or prevention of superinfection is regulated by gene expression. According to this model, after infection and prior to early HIV gene expression, superinfection can readily occur in order to permit recombination. As the most abundant early viral gene product, Nef defines the start point of a successful productive infection and functions as a master switch for the establishment of superinfection resistance by downregulating the entry receptor complex.”]
30. Wyatt GR (1952) The nucleic acids of some insect viruses. *Journal of General Physiology* 36:201–205
31. Forsdyke DR (1996) Different biological species ‘broadcast’ their DNAs at different (G+C)% ‘wavelengths.’ *Journal of Theoretical Biology* 178:405–417
32. Nikbakht H, Xia X, Hickey DA (2014) The evolution of genomic GC content undergoes a rapid reversal within the genus *Plasmodium*. *Genome* 57:507–511

33. Rothfels CJ et al. (2015) Natural hybridization between genera that diverged from each other approximately 60 million years ago. *American Naturalist* 185:433–442 [Since the hybrid fern is sterile, chromosomal incompatibilities *alone* (albeit not yet fully documented) seem to have been adequate to maintain reproductive isolation between the parental species over this vast timespan.]
34. Goldschmidt R (1940) *The Material Basis of Evolution*. Yale University Press, New Haven, p 220
35. Matassi G, Melis R, Macaya G, Bernardi G (1991) Compositional bimodality of the nuclear genome of tobacco. *Nucleic Acids Research* 19:5561–5567
36. Szybalski W (2001) My road to Öjvind Winge, the father of yeast genetics. *Genetics* 158:1–6
37. Marcet-Houben M, Gabaldón T (2015) Beyond the whole-genome duplication: phylogenetic evidence for an ancient interspecies hybridization in the baker's yeast lineage. *PLoS Biology* 13:e1002220
38. Kudla G, Murray AW, Tollervey D, Plotkin JB (2009) Coding-sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–258
39. Hershberg R, Petrov DA (2009) General rules for optimal codon choice. *PLOS Genetics* 5:e1000556
40. Lee JY et al. (2015) Base triplet stepping by Rad51/RecA family of recombinases. *Science* 349:977–981
41. Zhang Y, Sontheimer EJ (2014) Cascading into focus. *Science* 345:1452–1453
42. Crick F (1959) The present position of the coding problem. *The Structure and Function of Genetic Elements. Brookhaven Symposia in Biology* 12:35–39
43. Brbić M, Warnecke T, Kriško A, Supek F (2015) Global shifts in genome and proteome composition are very tightly coupled. *Genome Biology & Evolution* 7:1519–1532 [The conflict between the needs of DNA and of the proteins it encodes is strongly in favor of DNA.]

Chapter 11

Homostability

Genetic information is ‘written’ by a variation in sequence on the one hand, and the physical stability of the double-stranded structure is determined by the base composition on the other hand. ... DNA is found to consist of a number of homostability regions which come from homogenous base sequences consisting of 500 base pairs or more. ... Biologically, it is hard ... to believe that such regional homostability originates in a fundamental characteristic of the genetic code itself. It is quite plausible ... that the homostability region plays an important part somewhere in the biological process within which the DNA is closely related. If so, then the evolutionary selective force can be considered to have fixed such regions of DNA. From the size of the homostability region, recombination might be one possible process which is aided by it. In any case, the wobble bases [in codon third positions] must give the necessary redundancy to make a homostability region without spoiling the biological meaning of the genetic code.

Akiyoshi Wada et al. (1975) [1]

Your new dog must be kept from running off. First you tether it with a leash. Then you build a fence. So the leash now becomes available for other roles. Similarly, following establishment of a developmental or transmission barrier between species, a first barrier – (G+C)% – can become available for other roles. It need serve no longer as an *intergenomic* barrier. It is no longer a barrier to recombination between diverged paternal and maternal genomes at meiosis within their hybrid (Chapter 10). Here we consider its role in the prevention of recombination *within* a genome (*intragenomic* recombination). This can involve the differentiation of regions of relatively uniform (G+C)%, that Japanese physicists Akiyoshi Wada and Akira Suyama referred to as having a “homostabilizing propensity” and Giorgio Bernardi and his coworkers named “isochores” (Greek: *iso* = same; *choros* = group). These have the potential to recombinationally isolate different parts of a genome.

Gene Duplication and Isochores

From time-to-time genes duplicate to the advantage or disadvantage of their host organism. The survival of a duplicate copy of a gene depends on a variety of factors, including (i) natural selection favoring organisms where a function encoded by the gene is either increased or changed (i.e. there is either *concerted* or *divergent* gene evolution), (ii) a recombination-dependent process known as gene conversion, and (iii) a recombination-dependent process that can lead to copy-loss (Fig. 11.1). These *intragenomic* recombinations can occur when there is a successful search for

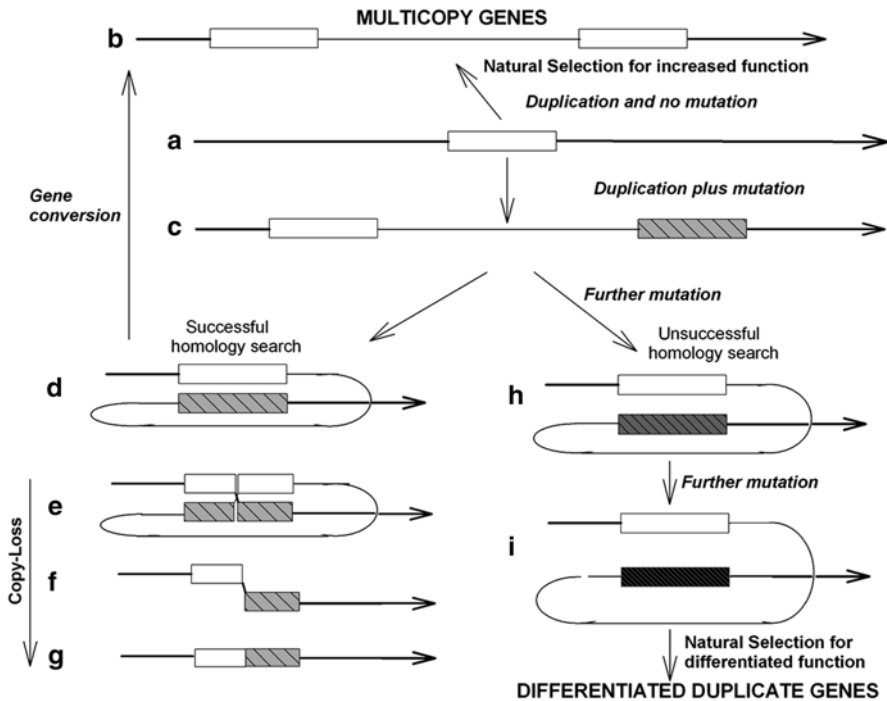


Fig. 11.1 Model for possible outcomes of a gene duplication. The duplication from (a) can result in identical multicopy genes (b) that confer an ability to produce more of the gene product. If this is advantageous, then the multicopy state will tend to be favored by natural selection. If unmutated (white box in (b)) or only slightly mutated (light gray striped box in (c)), there are not sufficient differences between the duplicates to prevent a successful homology search (d). This allows the mutation (c) to be reversed to (b) by the process known as “gene conversion.” This maintains identical copies, so allowing concerted evolution (conservation) of the multicopy genes. However, the recombination necessary for gene conversion can also result in removal of a circular intermediate (e, f), and restoration of the single copy state (g). The risk of copy-loss due to recombination (d-g) can be decreased by further mutation (dark gray striped box in (h)). This will decrease the probability of a successful homology search. Being protected against recombination (i.e. *preserved*), the duplicate is then free to differentiate further by mutation (black box in (i)). If the product of the new gene confers an advantage, then the duplicate will be further preserved by natural selection (divergent gene evolution). In the general case, mutation facilitating recombinational isolation (h) *precedes* mutation facilitating functional differentiation (i) under positive Darwinian selection

similarity between DNA strands. This is likely to be greatly influenced by the (G+C)% environment of the original gene and the (G+C)% environment where the duplicate copy locates.

For example, we have multiple genes encoding globin, which carries a pigment (haem) that makes our blood cells red. The attempted duplication of an ancestral globin gene to generate the α -globin and β -globin genes of modern primates might have failed, since sequence similarity would favor recombination between the two genes and incipient differences (early sequence divergence) could have been eliminated ("gene conversion;" Figs. 11.1, 11.2). However, the duplication appears to have involved *relocation* to a region (isochore) with a different (G+C)%, so the two genes became recombinationally isolated to the extent that initially the sequences flanking the genes differed in (G+C)%. Later the new gene would have increased its recombinational isolation by mutating to acquire the (G+C)% of its host isochore. As a consequence of the differences in (G+C)%, the corresponding mRNAs today utilize different codons for corresponding amino acids, even though both mRNAs are translated in the *same* cell using the *same* ribosomes and *same* tRNA populations. So it is most unlikely that the primary pressure to differentiate codons arose at the translational level.

Each isochore would have arisen as a random fluctuation in the base composition of a genomic region such that a copy of a duplicated gene that had transposed to that region was able to survive without recombination with the original gene for a sufficient number of generations to allow differentiation between the copy and its original to occur. This would have provided not only greater recombinational isolation, but also an opportunity for functional differentiation. If the latter differentiation were advantageous, organisms with the copy would be favored by natural selection. The regional base compositional fluctuation could then have 'hitch-hiked' through the generations by virtue of its linkage to the successful duplicate (i.e. the copy would have been positively selected). Thus, by preserving the duplicate copy from recombination with the original copy, the isochore would, in turn, have *itself* been preserved by virtue of its linkage to the duplicate copy [2, 3].

When functional differentiation of a duplicate is necessary for it to be selected (divergent evolution), there is the danger that, *before* natural selection can operate, either recombination-mediated gene conversion will reverse any incipient differentiation, or intragenic recombination between the copies (paralogs) will result in copy-loss. In the case of duplicate eukaryotic genes that have diverged in sequence, Koichi Matsuo and his colleagues noted that divergence was greatest at third codon positions, usually involving a change in (G+C)% [4–7]. Thus, there was a codon bias in favor of the positions of *least* importance for the functional differentiation that would be necessary for the operation of natural selection. Where amino acids had not changed, different gene copies used *different* synonymous codons. It was proposed that the (G+C)% change was an important "line of defence" against homologous recombination between the duplicates.

Thus, recombinational isolation of the duplicate (largely involving third codon position differences in (G+C)%) would *protect* (preserve) the duplicate so allowing time for functional differentiation (largely involving first and second codon position differences). Only then would natural selection come into play. In the general case,

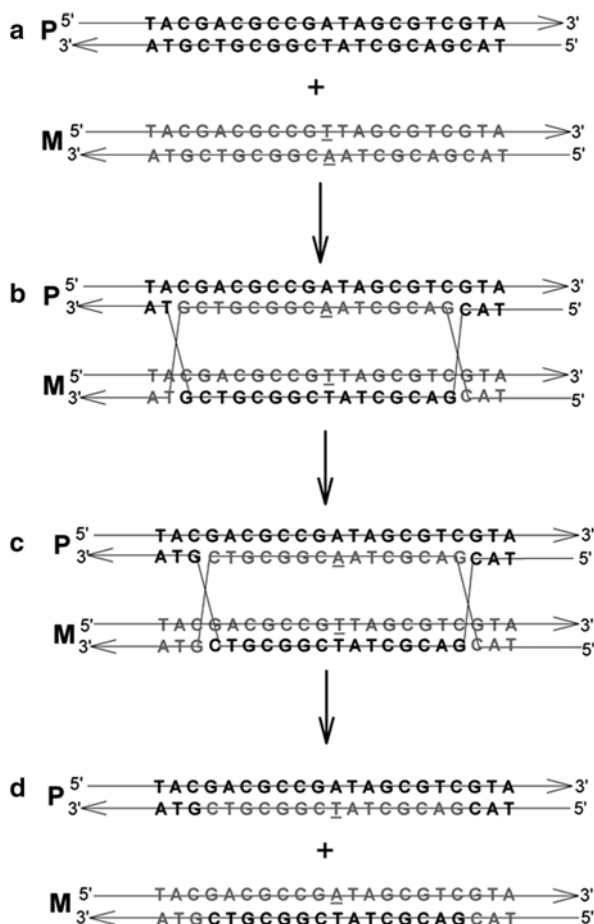


Fig. 11.2 Recombination repair and gene conversion. (a) Sequences 3.1 and 3.3 of Chapter 3 are taken as paternal (P) and maternal (M) versions (alleles) of a particular gene, with a single base-pair difference (underlined in the gray maternal version). (b) Following a successful homology search, exchange of single-strands results in a P segment becoming part of the bottom duplex, and an M segment becoming part of the top duplex, without strand breakage ("paranemic joint"). Because of the single base-pair difference, the recombination complex has an A mismatched with an A (top duplex), and a T mismatched with a T (bottom duplex). (c) Cross-over points ("Holliday junctions") can migrate without any permanent strand breakage (shown at left). (d) Strand breakage and resealing (ligation) occur, so that single-strand segments are exchanged between the duplexes (*recombination*). Without guidance, the 'repair' of each mismatch has a 50% chance of restoring the *status quo* as in (a) (i.e. replacing the A-A non-Watson-Crick base pair with an A-T Watson-Crick base-pair in the top duplex, and replacing the T-T non-Watson-Crick base-pair with a T-A Watson-Crick base pair in the bottom duplex). In the alternative shown here, the *status quo* is restored to the top duplex (an A is mutated to T), but in the bottom duplex the T-T non-Watson-Crick base-pair is replaced with an A-T Watson-Crick base-pair (i.e. a T is mutated to an A). Thus, there has been *conversion* of the sequence of the original M allele to that of the P allele. There has been a loss of heterozygosity (as in (a)) and a gain of homozygosity (as in (d)).

isolation would *precede* functional differentiation, not the converse. (G+C)% differentiation, largely involving third codon positions, would precede functional differentiation, largely involving first and second codon positions under positive Darwinian selection.

From all this it would be predicted that, if a gene from one isochore were transposed to an isochore of different (G+C)%, where it retained or increased its selective advantage, then the gene would preferentially accept mutations converting its (G+C)% to that of the new host isochore (i.e. organisms with those mutations would be genetically fitter and thus likely to leave more fertile offspring than organisms without the mutations). Indeed, there is evidence supporting this. The sex chromosomes (X and Y) tend not to recombine at meiosis except in a small region (the “pseudoautosomal” region; see Chapter 17). Transfer of a gene from a non-recombining part of a sex chromosome to the pseudoautosomal region, forces the gene rapidly to change its (G+C)% value [8].

For various reasons (e.g. large demand for the gene product), certain genes are present in multiple identical copies. But, in the absence of some restraint, copies that are initially identical will inevitably diverge in sequence. So how can multicopy genes (e.g. rRNA genes) preserve their similarity to each other? To prevent divergence through the generations (i.e. to allow concerted evolution), they should mutually correct each other to eliminate deviant copies. This is likely to occur by a recombination-dependent process – gene conversion (Figs. 11.1 and 11.2). Thus, multicopy genes should all be in, either the same isochore, or isochores of very close (G+C)%, so that recombination is favored.

Gene conversion can involve either homologous genes (non-allelic “paralogs”) *in series* on the same chromosome (Fig. 11.1) [9], or *parallel* copies of genes (alleles) on homologous chromosomes (Fig. 11.2). Here, for simplicity, we consider a single base-pair difference (gray underlined in the maternal version). But the difference might involve many consecutive base pairs or even an entire gene if a rare duplication event had occurred in one genome as shown in Figure 11.1. In this case the mismatched maternal region would be displaced as a large loop. Repair with strand breakage would result either in a copying of the loop sequence into the genome of paternal origin, or excision of the loop sequence from the genome of maternal origin.

It is important to note that crossover points (junctions) are not fixed and this can assist gene conversion when conditions are less than propitious. Sequence 3.1 (P in Fig. 11.2) can form a stem-loop with the central bases being located in the loop (see sequence 4.4 in Chapter 4). Since the single base-pair difference between P and M versions is in this loop, then the M version has the potential to form a similar stem-loop. But, because the loops differ slightly, during the initial homology search loop-loop ‘kissing’ interactions might fail and prohibit subsequent steps. However, crossover points (junctions) can migrate over both long and short distances (e.g. (b) to (c) in Fig. 11.2). So that if crossover is prohibited in one region there is the possibility of migration from a crossover in a neighboring region that would reveal mismatches through malpairing. Thus, multiple distributed incompatibilities (base differences) should be needed to inhibit the pairing of homologous chromosomes.

Isochores Early

Before DNA sequencing was available, “isochores” were described as DNA segments that could be identified on the basis of their distinct densities in samples of duplex DNA obtained from organisms whose cells had nuclei (eukaryotes). The method involved physically disrupting (sheering) DNA in solution by vigorous stirring. The resulting fragments were then centrifuged in a salt density gradient. It was observed that, when the sheering was carefully controlled to break DNA down to duplex segments of about 300 kilobases, bands of distinct densities were formed. The densities could be related to the average $(G+C)\%$ values of the segments – the greater these values, the greater the densities. This way of assessing the $(G+C)\%$ of a duplex DNA segment distinguished one large segment from another, and largeness became a defining property of isochores.

Isochores, as so defined, were not identified in bacteria, which do not have distinct nuclear membranes (prokaryotes; see Chapter 13). Since prokaryotes (e.g. bacteria) and eukaryotes (e.g. primates) are considered to have evolved from a common ancestor, does this mean that the ancestor had isochores that were subsequently lost by prokaryotes during or after their divergence from the eukaryote lineage (isochores-early)? Or did the ancestor not have isochores, which were therefore freshly acquired by the eukaryotic lineage after its divergence from the prokaryotic lineage (isochores-late)? If prokaryotes could be shown to have isochores, then this would favor the isochores-early hypothesis.

Indeed, prior to modern sequencing technologies, physical methods demonstrated small segments of distinct $(G+C)\%$ in the genomes of prokaryotes and their viruses. The 48 kb duplex genome of phage lambda (Chapter 5) was extensively sheered to break it down to subgenome-sized fragments. These were resolved by the above density method into six distinct segments differing in $(G+C)\%$ [10]. Another, more sensitive, method (thermal denaturation spectrophotometry), demonstrated thirty four “gene sized” segments [11].

With the advent of sequencing, in 1984 Mervyn Bibb and his colleagues were able to plot the average $(G+C)\%$ values of every third base for small windows in the sequences of various bacteria (Fig. 11.3) [12]. Three plots were generated, the first beginning with the first base of the sequence (i.e. bases in frame 1, 4, 7, etc.), the second beginning with the second base of the sequence (i.e. bases in frame 2, 5, 8, etc.), and the third beginning with the third base of the sequence (i.e. bases in frame 3, 6, 9, etc.). In certain small regions $(G+C)\%$ values were relatively constant within each frame. *These regions of constant $(G+C)\%$ corresponded to genes.* Thus, individual genes were shown to have a relatively uniform $(G+C)\%$ values *to which each codon position made a distinctive contribution.*

This observation is not confined to bacteria. Wada and Suyama noted that, whether prokaryotic or eukaryotic, “every base in a codon seems to work cooperatively towards realizing the gene’s characteristic value of $(G+C)$ content.” This was a “homostabilizing propensity” allowing a gene to maintain a distinct $(G+C)\%$, relatively uniform along its length, *which would differentiate it from other genes in the same genome* [13]. Thus, each gene constitutes a homostabilizing region in DNA, and $(G+C)\%$ values change, sometimes dramatically, at genic boundaries [14].

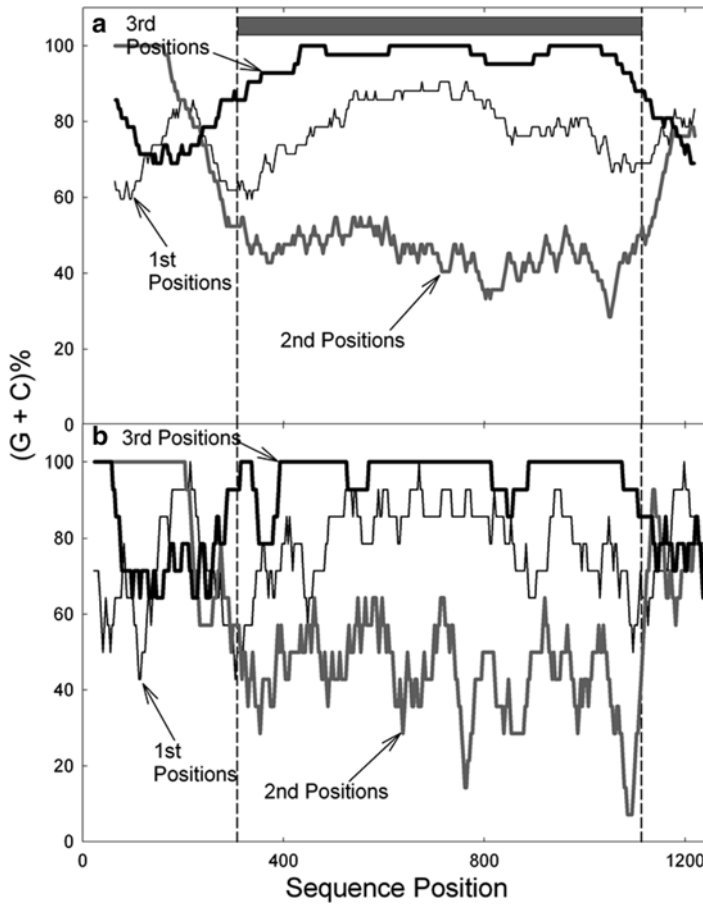


Fig. 11.3 In the region of a gene the $(G+C)\%$ of every third base tends to be both distinct from neighboring bases, and constant. Bases were counted in windows of 126 bases (a) or 42 bases (b), and in one of three frames, either beginning with base 1 (frame 1, 4, 7, etc.), or with base 2 (frame 2, 5, 8, etc.), or with base 3 (frame 3, 6, 9, etc.). Thus, for calculating average $(G+C)\%$ values, each frame takes into account either 42 bases (a) or 14 bases (b). The frames are here named according to the positions of bases in the codons of the gene (encoding an enzyme in the GC-rich bacterium *Streptomyces fradiae*). Vertical dashed lines and the gray rectangle indicate the limits of the rightward-transcribed gene. Note that the relative constancy of $(G+C)\%$ is most for the third codon position (least dependent on the encoded amino acids), and least for the second codon position (most dependent on the encoded amino acids). The fluctuation in values at the second codon position is more apparent when a window size equivalent to 14 codons is used (b) than when a window size equivalent to 42 codons is used (a). This figure was redrawn from [12]

Stated another way, if large size is excluded as a defining property, many bacteria have isochores. When isochores are defined as DNA segments of relatively uniform $(G+C)\%$ that are coinherited with specific sequences of bases, then bacteria have isochores. To contrast with the classical isochores of Bernardi, these are termed “microisochores,” and their length is that of a gene, or small group of genes

(see Chapter 12). Thus, classical eukaryotic isochores (“macroisochores”) can be viewed as constellations of microisochores of a particular $(G+C)\%$. The proposed antirecombination role of differences in $(G+C)\%$ (Chapter 10) would require that, unless they represent multicopy genes, microisochores sharing a common macroisochore (i.e. they have a common $(G+C)\%$) have other sequence differences that are sufficient to prevent recombination between themselves [15]. Within an organism, genes with similar $(G+C)\%$ values may sometimes be expressed in similar tissues, so that there is a tissue-specific codon usage tendency [5].

Since both prokaryotic (e.g. bacterial) and eukaryotic (e.g. primate) lineages have some form of isochore, this appears most consistent with the isochores-early hypothesis. While not endorsing a particular role for $(G+C)\%$, this underlines the fundamental importance of $(G+C)\%$ differences in biology. Let metaphors multiply! A given segment of DNA is coinherited with a ‘coat’ of a particular $(G+C)\%$ ‘color.’ A given segment of DNA ‘speaks’ with a particular $(G+C)\%$ ‘accent’ (and hence has a distinct potential vibrational frequency; Fig. 5.2). A fundamental duality of information levels is again manifest!

The Gene as a Unit of Recombination

As will be seen in Chapter 12, it is likely that differences in $(G+C)\%$ serve to isolate recombinationally, both genes within a genome, and genomes within a group of species (a taxonomic group). The power to recombine is fundamental to all life forms because, for a variety of reasons, it is advantageous (see Chapter 17). However, the same power threatens to homogenize (blend) genes within a genome, and to homogenize (blend) the genomes of members of allied species within a taxonomic group (i.e. genus). This would countermand evolution both within a species and between species. *Thus, functional differentiation, be it between genes in a genome, or between genomes in a taxonomic group (speciation), must, in the general case, be preceded (or closely accompanied) by the establishment of recombinational barriers.* Species have long been defined in terms of recombinational barriers (Chapter 8). In some contexts, genes can be defined similarly. A species can be defined as a unit of recombination (or rather, of antirecombination with respect to other species). So can a gene.

Most definitions of the ‘gene’ contain a loose or explicit reference to function [16]. Biologists talk of a gene encoding information for tallness in peas. Biochemists talk of a gene encoding information for growth hormone (a protein), and relate this to a DNA segment. However, before it can function, information must be *preserved*. Classical Darwinian orthodoxy has it that function, through natural selection, is *itself* the preserving agent. Thus, function and preservation go hand-in-hand, but function is more fundamental than preservation. In 1966 US biologist George Williams, an originator with Hans Kalmus of what became known as the ‘selfish gene’ concept, seemed to argue the converse when arriving at a new definition [17–19].

The function of any multipart entity, which needs more than one part for this function, is usually dependent on its parts not being separated. Preservation can be more fundamental than function. Williams proposed that a gene should be defined entirely by its property of remaining intact as it passes from generation to generation. He identified recombination as a major threat to that intactness. Thus, for Williams, “gene” meant any DNA segment that had the potential to persist for enough generations to serve as a unit for natural selection; this required that it not be easily disruptable by recombination. The gene is a unit of recombination (or rather, of antirecombination with respect to other genes):

Socrates’ genes may be with us yet, but not his genotype, because meiosis and recombination destroy genotypes as surely as death. It is only the meiotically dissociated fragments of the genotype that are transmitted in sexual reproduction, and these fragments are further fragmented by meiosis in the next generation. If there is an *ultimate indivisible fragment* it is, by definition, ‘the gene’ that is treated in the abstract discussions of population genetics. Various kinds of suppression of recombination may cause a major chromosomal segment or even a whole chromosome to be transmitted entire for many generations in certain lines of descent. In such cases the segment, or chromosome, behaves in a way that approximates the population genetics of a single gene. ... I use the term *gene* to mean “that which segregates and recombines with appreciable frequency”. ... A gene is one of a multitude of meiotically dissociable units that make up the genotypic message.

Despite this, Williams did not invoke any special chromosomal characteristic that might act to facilitate preservation. Pointing to “the now discredited theories of the nineteenth century,” and lamenting an opposition that “arises ... not from what reason dictates, but from the limits of what the imagination can accept,” his text, *Adaptation and Natural Selection*, made what seemed a compelling case for “natural selection as the primary or exclusive creative force.” No other agency was required. This tendency, which can be infectious, to bolster the scientific with the *ad hominem* in otherwise rational discourse, will be considered in the Epilogue. In contrast, we have here considered intergenomic and intragenomic differences in (G+C)% as agencies, essentially independent of natural selection, that preserve the integrity of species and genes, respectively.

Within a species individual genes differ in their (G+C)%. But *relative* positions of genes on the (G+C)% scale are usually preserved through speciation events. If, in an ancestral species, gene *A* was of higher (G+C)% than gene *B*, then this relationship is sustained in the modern species that resulted from divergences from that ancestral species. Accordingly, when the (G+C)% values of the genes of one of the modern species are plotted against the corresponding (G+C)% values of similar (orthologous) genes in another modern species, the points usually display a linear relationship (c.f. Fig. 2.5).

Of course, species with *intragenomic* isochore differentiation can themselves further differentiate into new species. In this case, a further layer of *intergenomic* (G+C)% differentiation would be imposed upon the previous intragenomic differentiation. Again, when a sufficient degree of reproductive isolation had been achieved this initial barrier between species would usually be replaced by other barriers, thus leaving (G+C)% free to continue differentiating in response to intragenomic demands.

The Boundary Problem

Dawkins agreed with William's definition of the gene [20]:

Sexual reproduction has the effect of mixing and shuffling genes. This means that any one individual body is just a temporary vehicle for a short-lived combination of genes. The *combination* of genes which is any one individual may be short-lived, but the genes themselves are potentially very long-lived. Their paths constantly cross and re-cross down the generations. One gene may be regarded as a unit which survives through a large number of successive individual bodies. (Dawkins' italics)

Yet Dawkins recognized that William's "ultimate indivisible fragment" might need some qualification. Although certain parts of genomes were more prone to recombination than others ("hotspots;" see Chapter 17), it was not thought that recombination would respect genic boundaries:

Even a cistron [gene] can be cut in two by crossing over. The gene is defined as a piece of a chromosome which is sufficiently short for it to last, potentially, for *long enough* for it to function as a significant unit of natural selection. (Dawkins' italics)

Thus, to get round the boundary problem, Dawkins advanced a statistical conception of the gene. Genes were small parts of genomes, so the probability was high that a small chromosome segment containing a gene would escape meiotic fragmentation for many generations. Eventually the gene would succumb to the recombination mill, but this would be a rare event. A better solution would be for branch migration to move an intragenic crossover junction to the genic boundary where strand-breakage could occur (Fig. 11.4).

Thus, there should be a mechanism for guiding paranemic cross-overs from the interior of a gene to its boundaries, thus preserving the gene intact. Then the Williams-Dawkins definition of the gene would approximate to the definition of biochemists [16]. Branch migration from within a gene would proceed along a region of relatively uniform (G+C)% (corresponding to the gene) until it approached a region of different (G+C)% (intergenic DNA, a neighboring gene or an intron). The abrupt change in (G+C)% would halt the migration. This would be sensed by the recombinogenic enzymes that initiate strand-breakage. (For multi-exon genes we should read exon instead of gene.)

While not excluding intragenic recombination, this process would greatly decrease its probability. Indeed, branch migration is known to be sequence dependent [21]. It proceeds when the Holliday junction unfolds, and stalls when the junction folds – a process that is sensitive to differences in (G+C)% [22].

Thermophiles

(G+C)% is never entirely free. It can be constrained by the demands on gene function (i.e. natural selection) that primarily affect first and second codon positions. This is particularly evident in certain environments where natural selection can make direct demands on base composition.

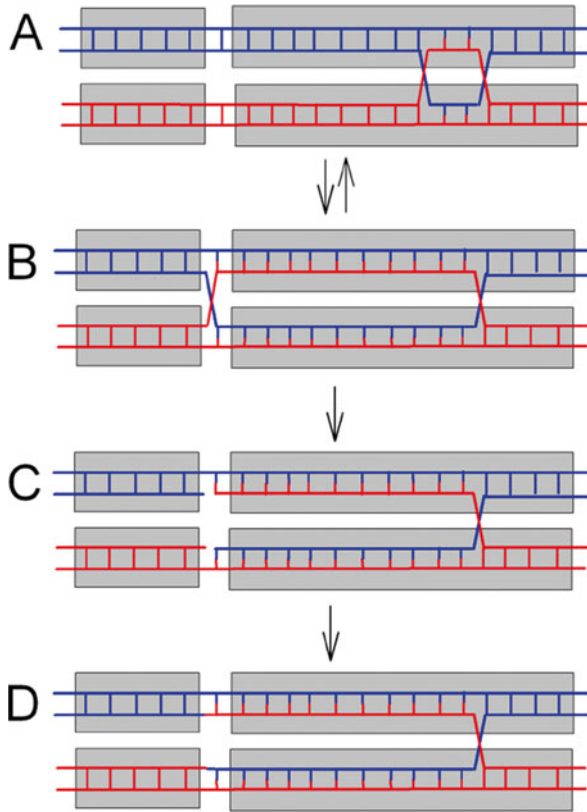


Fig 11.4 Migration of a Holliday junction to the genic boundary where the change in base composition triggers strand breakage. A. The cross-over intermediate. B. Branch migration leftward to the genic boundary without strand breakage. C. Strand breakage. D. Ligation so that a maternal strand (red) is joined to a paternal strand (blue). The other Holliday junction can migrate to the right.

Hot springs, oceanic thermal vents, and the radioactive discharges of nuclear reactors are among the few environments on this planet where living organisms are not expected. Yet all contain living organisms (“extremophiles”). These are usually species of bacteria or archaea. Fortunately, since heat and radiation are convenient ways of achieving sterilization in hospitals, none of these organisms has been found (or genetically engineered to become) pathogenic (so far). Thermophiles are so-called because they thrive at high temperatures. Proteins purified from thermophiles may show high stability at normal temperatures, a feature that has attracted commercial interest (i.e. they have a long ‘shelf life’). Hence, the full genomic sequences of many prokaryotic thermophiles are now available.

Some thermophiles normally live at the temperature of boiling water. Pure nucleic acids in solution at this temperature soon degrade. So how do nucleic acids survive in thermophiles? The secondary structure of nucleic acids with a high (G+C)% is more stable than that of nucleic acids with a low (G+C)%. This is consistent with Watson-Crick G-C bonds being strong, and A-T (or A-U) bonds being

weak (see Table 2.1). Do thermophiles have high (G+C)% DNA? It depends on the gene. In the case of genes corresponding to RNAs whose *structure* is vital for RNA function – rRNAs and tRNAs – the answer is affirmative. Free of coding constraints (they are not mRNAs), yet required to form part of the precise structure of ribosomes where protein synthesis occurs, genes corresponding to rRNAs appear to have had the flexibility to accept mutations that increase G+C. That is, organisms that did not accept such mutations perished by natural selection, which presumably acted because such organisms had less efficient protein synthesis at high temperatures.

The G+C content of rRNAs is directly proportional to the normal growth temperature, so that rRNAs of thermophilic prokaryotes are highly enriched in G and C [23]. Yet, although optimum growth temperature correlates positively with the G+C content of rRNA (and hence of rRNA genes), optimum growth temperature does *not* correlate positively with the overall G+C content of genomic DNA, and hence with that of the numerous mRNA populations transcribed from the genes in that DNA (Fig. 11.5a). Instead, optimum growth temperature correlates positively with A+G content (Fig. 11.5b) [24].

The stability of duplex DNA at high temperatures can be achieved in ways other than by an increase in G+C content. These include association with small basic peptides (polyamines) and relaxation of torsional strain (supercoiling) [25, 26]. Thus, there is every reason to believe that, whatever their (G+C)%, thermophiles are able, both to maintain their DNAs in classical duplex structures with Watson-Crick hydrogen-bonding between opposite strands, and to adopt any necessary extruded secondary structures involving intrastrand hydrogen-bonding (i.e. stem-loops).

All this returns us to the topics of purine-loading (Chapter 6) and potential conflicts among various adaptations at the nucleic acid level. These can also impact adaptations at the protein level (see Chapters 10, 12).

Summary

Once a speciation process has initiated, other factors may replace (G+C)% as a barrier to reproduction (preventing *intergenomic* recombination *between* species). This leaves (G+C)% free to assume other roles, such as preventing *intragenomic* (e.g. *intergenic*) recombination *within* a species. Many organisms have “macroisochores,” defined as long segments of relatively uniform (G+C)% that are coinherited with specific sequences of bases. Isochores may facilitate gene duplication. Indeed, each gene has a “homostabilizing propensity” to maintain itself as a “microisochore” of relatively uniform (G+C)%. The protection, afforded by differences in (G+C)%, against inadvertent recombination, facilitates the duplication both of genes and of genomes (speciation). George Williams’ definition of a gene as a unit of recombination, rather than of function, is now seen to have a chemical basis. For organisms living in extreme environments (e.g. thermophiles), natural selection seems to influence values for (G+C)% less than for (A+G)%.

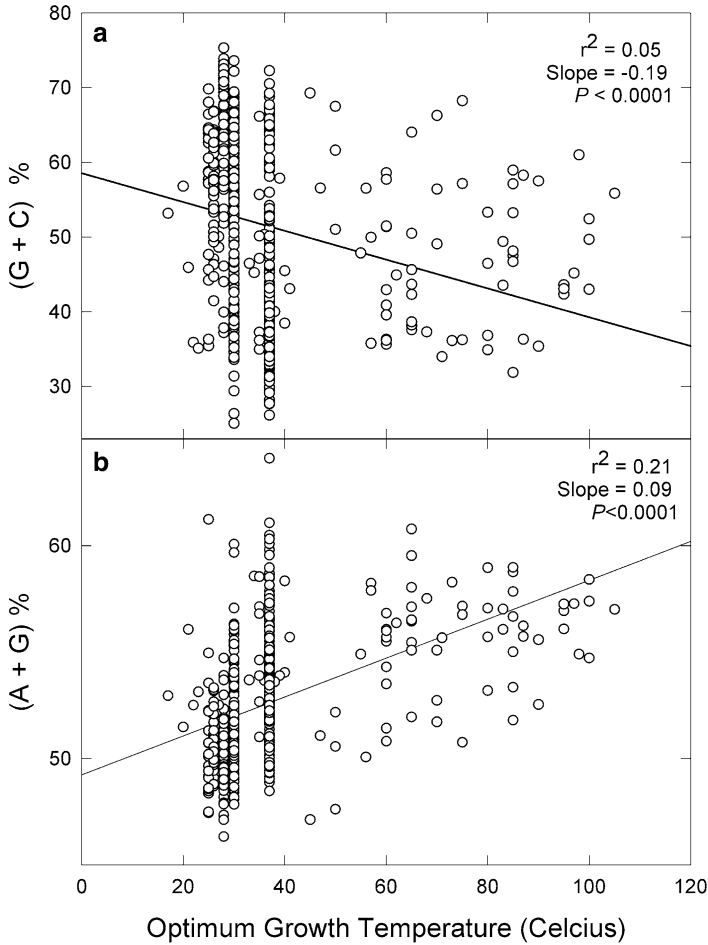


Fig. 11.5 Variation of average (a) (G+C)%, and (b) (A+G)% of the genes of prokaryotic species (bacteria and archaea), with the temperature required for optimal growth of each species. The large number of data points corresponding to ambient temperatures (20-30°C) and human blood temperature (37°C) reflect the fact that relatively few thermophiles had been sequenced circa 2000 (when the graph was prepared). Note that, whereas in (a) only 5% of the variation between points can be explained by growth temperature ($r^2 = 0.05$), in (b) 21% can be explained on this basis ($r^2 = 0.21$; see Appendix A)

References

1. Wada A, Tachibana H, Gotoh O, Takanami M (1975) Long range homogeneity of physical stability in double-stranded DNA. *Nature* 263:439–440 [Some third codon position bases are deemed to “wobble” to better fit the codon to its complementary anticodon sequence.]
2. Forsdyke DR, Mortimer JR (2000) Chargaff’s legacy. *Gene* 261:127–137

3. Rodin SN, Parkhomchuk DV (2004) Position-associated GC asymmetry of gene duplicates. *Journal of Molecular Evolution* 59:372–384
4. Matsuo K, Clay O, Kunzler P, Georgiev O, Urbanek P, Schaffner W (1994) Short introns interrupting the Oct-2 POU domain may prevent recombination between POU family genes without interfering with potential POU domain ‘shuffling’ in evolution. *Biological Chemistry Hoppe-Seyler* 375:675–683
5. Newgard CB, Nakano K, Hwang PK, Fletterick RJ (1986) Sequence analysis of the cDNA encoding human liver glycogen phosphorylase reveals tissue-specific codon usage. *Proceedings of the National Academy of Sciences USA* 83:8132–8136
6. Moore R C, Purugganan MD (2003) The early stages of duplicate gene evolution. *Proceedings of the National Academy of Sciences USA* 100:15682–15687
7. Zhang Z, Kishino H (2004) Genomic background drives the divergence of duplicated *Amylase* genes at synonymous sites in *Drosophila*. *Molecular Biology & Evolution* 21:222–27
8. Montoya-Burgos JI, Boursot P, Galtier N (2003) Recombination explains isochores in mammalian genomes. *Trends in Genetics* 19: 128–130
9. Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC (2003) Abundant gene conversion between arms of palindromes in human and ape chromosomes. *Nature* 423:873–876
10. Skalka A, Burgi E, Hershey AD (1968) Segmental distribution of nucleotides in the DNA of bacteriophage lambda. *Journal of Molecular Biology* 34:1–16
11. Vizard DL, Ansevin AT (1976) High resolution thermal denaturation of DNA: thermalites of bacteriophage DNA. *Biochemistry* 15:741–750
12. Bibb MJ, Findlay PR, Johnson MW (1984) The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* 30:157–166
13. Wada A, Suyama A (1985) Third letters in codons counterbalance the (G+C) content of their first and second letters. *Federation of European Biochemical Societies Letters* 188:291–294
14. Zhang L, Kasif S, Cantor CR, Broude NE (2004) GC/AT-content spikes as genomic punctuation marks. *Proceedings of the National Academy of Sciences USA* 101:16855–16860
15. Kudla G, Helwak A, Lipinski L (2004) Gene conversion and GC-content evolution in mammalian Hsp70. *Molecular Biology & Evolution* 21:1438–1444
16. Forsdyke DR (2009) Scherrer and Josts’ symposium. The gene concept in 2008. *Theory in Biosciences* 128:157–161
17. Forsdyke DR (2011) The selfish gene revisited: reconciliation of the Williams–Dawkins and conventional definitions. *Biological Theory* 5:246–255
18. Kalmus H (1950) A cybernetical aspect of genetics. *Journal of Heredity* 41: 19–22
19. Williams GC (1966) *Adaptation and Natural Selection*. Princeton University Press, Princeton, pp 24–25. [WD Hamilton was also an important contributor to the selfish gene concept.]
20. Dawkins R (1976) *The Selfish Gene*. Oxford University Press, New York, pp 25–38
21. Sun W, Mao C, Liu F, Seeman NC (1998) Sequence dependence of branch migratory minima. *Journal of Molecular Biology* 282:59–70
22. Karymov MA, Boganov A, Lyubchenko YL (2008) Single molecule fluorescence analysis of branch migration of Holliday junctions: effect of DNA sequence. *Biophysical Journal* 95:1239–1247
23. Galtier N, Lobry JR (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution* 44:632–636 [This paper is much cited, but there are serious misinterpretations of the data.]
24. Lambros R, Mortimer JR, Forsdyke DR (2003) Optimum growth temperature and the base composition of open reading frames in prokaryotes. *Extremophiles* 7:443–450
25. Oshima T, Hamasaki N, Uzawa T, Friedman SM (1990) Biochemical functions of unusual polyamines found in the cells of extreme thermophiles. In: Goldembeg SH, Algranati ID (eds) *The Biology and Chemistry of Polyamines*. Oxford University Press, New York, pp 1–10
26. Friedman SM, Malik M, Drlica K (1995) DNA supercoiling in a thermotolerant mutant of *Escherichia coli*. *Molecular & General Genetics* 248:417–422

Part IV
Conflict Within Genomes

Chapter 12

Conflict Resolution

There would appear to be no reason why the recombinator should not be a sequence which codes in the normal way for amino acids, and that it would therefore be within the genes rather than between them.

Robin Holliday (1968) [1]

The tasks of evolutionary bioinformatics (EB) are to identify the forms of information that genomes convey and show how potential conflicts between different forms are reconciled. Apparent redundancies (e.g. diploidy; Chapter 2), and beliefs in the existence of ‘neutral’ mutations (Chapter 7), and of ‘junk’ DNA (see Chapter 15), tended to support the view that there is much vacant genome space. Hence there is ‘room for all’ in the journey of the genes through the generations. Suggestions that there might be conflicts between different forms of information were not taken too seriously. However, when genomic information was thought of in the same way as the other forms of information with which we are familiar (Chapters 2, 3 and 4), it became evident that apparent redundancies might play important roles – error-detection and correction, and much more. The possibility of conflict could no longer be evaded. The essential argument of this book is that many puzzling features of genomes can best be understood in such terms, as will be emphasized in this and subsequent chapters.

The idea of competition at different levels can be traced back to Galton [2]. In 1876 he noted that “the limitation of space in the stirp [germ-line] must compel a limitation ... in the number of varieties of each species of germ.” Furthermore, there was the possibility of interactions:

Each germ [gene] has many neighbours: a sphere surrounded by other spheres of equal sizes, like a cannon ball in the middle of a heap of them, when they are piled in the most compact form, is in immediate contact with at least twelve others. We may therefore feel assured, that the germs may be affected by numerous forces on all sides, varying with their change of place, and that they must fall into many positions of temporary and transient equilibrium, and undergo a long period of restless unsettlement, before they severally attain the positions for which they are finally best suited.

Likewise a century later Gregory Bateson noted [3]:

Two or more information sources come together to give information of a sort different from what was in either source separately. At present there is no existing science whose special interest is the combining of pieces of information. But I shall argue that the evolutionary process must depend upon such double increments of information. Every evolutionary step

is an addition of information to an already existing system. Because this is so, the combinations, harmonies, and discords between successive pieces and layers of information will present many problems of survival and determine many directions of change.

Two Levels of Information

DNA molecules engage in various protein-mediated transactions (transcription, replication, recombination, repair), which require that specific DNA sequences be recognized by specific proteins. Included among these recognition sequences in DNA might be sequences that themselves encode proteins. Thus some sequences could *both* encode, and be recognized by, proteins. This was noted by British geneticist Robin Holliday in 1968 in the context of a specific sequence recognized by proteins mediating recombination that he referred to as a “recombinator” sequence [1].

In this light, in 1971 Israeli geneticist Tamar Schaap distinguished the potential for two kinds of information in a DNA sequence, one “active” and one “passive” [4]:

If indeed recognition sites are located intragenically, the DNA must contain two kinds of information: the information transcribed into mRNA and translated into polypeptides, henceforth named “active,” and the information serving to distinguish particular regions of the DNA molecule, henceforth called “passive.”

Recognizing that the coding needs of the two kinds might differ, Schaap continued:

It seems therefore logical to assume that the two kinds of information are carried by different codes using the same letters: the four nucleotides, which are read in triplets in the active information code, form ‘words’ of a different length and/or different structure in the passive information code. A word in the passive code may, for example, consist of a nucleotide sequence in which only every second or third nucleotide is essential to the message.

Thus, an individual base mutation should have the potential to affect either one or both coding functions. Although Schaap used the term “passive,” an underlying function could be affected by mutations in the passive information code. A mutation in Holliday’s recombinator sequence, for example, might impair DNA recognition by recombination proteins, and hence impair recombination-dependent functions. So recombination could be impaired by mutations both in the DNA *encoding* the recombination proteins themselves, and in the DNA *targets* (substrates) of those recombination proteins. The former mutations would tend to be localized to the regions of the corresponding genes. The latter mutations would tend to be general.

Codon Bias

Schaap noted that the needs of “passive” information might dictate that certain codons be used preferentially (i.e. there would be codon bias):

Were homologous [synonymous] codons fully equivalent, the establishment of any particular one would have been most easily understood as the result of a random process

However, non-equivalence of homologous codons may stem from their participation in sequences forming recognition sites. Consequently, the choice of a particular codon rather than its homologs may be a function of the selective value of recombination, or initiation of replication or transcription in its vicinity. Thus, the establishment of specific codons at particular sites may well be the result of natural selection rather than a random process.

A consequence of this was that a mutation, even though not changing an amino acid, was deemed unlikely to be neutral in its effect on the organism:

The dual information carried by DNA can also explain the so-called neutral mutations The difference between the selective values of an active and an inactive recognition site may be greater than the one between two polypeptides differing in one amino acid. In such cases, the passive rather than the active information will be the criterion for selection, causing the establishment of mutations with no apparent selective advantage.

Hence, selection could be on the genome phenotype (Chapter 7), rather than on the conventional (classical) phenotype:

The establishment of homologous [synonymous] and neutral mutations, usually attributed to random processes, can be accordingly understood as the result of natural selection acting at the level of the DNA rather than that of the polypeptide.

In the 1970s codon bias emerged as a general phenomenon with species specificity. There appeared to be codon redundancy (Table 12.1). Yet there was a tendency for all the coding regions of the genes of a particular species to use the *same* sub-set of the range of codons. This could require a ‘choice’ from among as many as six codons for a particular amino acid.

The issue of which codon was employed in a particular circumstance was considered in France by one of the founders of EB, the USA-expatriate Richard Grantham. He observed in 1972 [5] that codon usage was not random in microorganisms, “suggesting a mechanism against [base] composition drift.” Since “little latitude appears left for ‘neutral’ or synonymous mutations”, he was led to his “genome hypothesis,” which specified that undefined adaptive genomic pressure(s) caused changes in base composition and hence in codon usage [6]:

Each ... species has a ‘system’ or coding strategy for choosing among synonymous codons. This system or dialect [accent] is repeated in each gene of a genome and hence is a characteristic of the genome.

Grantham sensed that the coding strategy was of deep relevance to an organism’s biology:

What is the fundamental explanation for interspecific variation in coding strategy? Are we faced with a situation of continuous variation within and between species, thus embracing a Darwinian perspective of gradual separation of populations to form new species ...? This is the heart of the problem of molecular evolution.

We needed to determine “how much independence exists between the two levels of evolution” (that of the genome phenotype and of the classical phenotype), for “it is too easy just to say most mutations are neutral.”

Grantham posed good questions, but left them unanswered. “Easy” non-adaptive “neutralist” explanations won general support. Mathematicians were happy to have a ‘null’ frame-of-reference. Surely, codon biases must reflect biases in the generation

Table 12.1 Codon sets (sextets, quartets, duets). Codons within a set are not uniformly employed. Species, and to a lesser extent genes within a species, differ in the 'choice' of their usual codon to encode a particular amino acid (codon bias). The 61 codons for amino acids are grouped into 20 sets of between one and six synonymous members. The code includes eight quartets and twelve duets, the isoleucine trio, the single codons of methionine and tryptophan, plus the three codons that signal termination of a protein sequence. Changes in first codon positions always change the nature of the encoded amino acids, with leeway only in the sextet sets. Changes in second codon positions always change the nature of the encoded amino acids, with leeway only in the case of the serine sextet. Changes in third codon positions fail to change the nature of the encoded amino acid, with leeway only in the duets (e.g. CAA to CAU), and odd numbered codon sets (e.g. AUU to AUG)

Sextet = A Quartet plus a Duet

Arginine	Leucine	Serine
CGU	CUU	UCU
CGC	CUC	UCC
CGA	CUA	UCA
CGG	CUG	UCG
AGA	UUA	AGU
AGG	UUG	AGC

Quartets

Threonine	Proline	Alanine	Glycine	Valine
ACU	CCU	GCU	GGU	GUU
ACC	CCC	GCC	GGC	GUC
ACA	CCA	GCA	GGA	GUA
ACG	CCG	GCG	GGG	GUG

Duets

Lysine	Asparagine	Glutamine	Histidine	Glutamic
AAA	AAU	CAA	CAU	GAA
AAG	AAC	CAG	CAC	GAG
Aspartic	Tyrosine	Cysteine	Phenylalanine	
GAU	UAU	UGU	UUU	
GAC	UAC	UGC	UUC	

Odd Numbers

Isoleucine	Methionine	Tryptophan	Terminators
AUU	AUG	UGG	UAA
AUC			UAG
AUA			UGA

or repair of mutations. However, it was conceded that mRNAs of highly expressed genes might need to be efficiently translated [7]. Adaptive factors such as the need to translate an abundant mRNA efficiently – so freeing ribosomes for use by other mRNAs – had the potential to influence codon usage. But this did not seem to explain the generality of codon bias, since only in certain species did translation appear to be rate-limiting in the process of maximizing the number of descendents (i.e. maximizing biological fitness). Furthermore, while mutational biases (i.e. a tendency to mutate unidirectionally from one base to another) were known [8] (see Chapter 15), this explanation was a ‘default’ (fall-back) explanation that thrived mainly because other explanations for codon-bias had not emerged.

This book argues that genomic factors, such as **GC**-pressure, purine-loading pressure and fold pressure, play an important, and sometimes *dominant*, role in determining which codons are preferred to encode a particular amino acid in a particular species (and even in a particular gene within that species). In general, codon-bias is not a primary phenomenon, but is *secondary* to other underlying phenomena.

If, for reasons set out in Chapter 10, a particular species has a high genomic (**G+C**)% (which affects both coding regions and non-coding regions), there will be a bias in favor of **GC**-rich, rather than **AT**-rich, codons. Thus, instead of the **RNY** periodicity (Chapter 7), coding sequences will tend to adopt a **GNC** periodicity (e.g. ...**GNC,GNC,GNC,GNC**, ...). Similarly, coding sequences in **AT**-rich genomes will tend to adopt an **ANT** periodicity (e.g. ...**ANT,ANT,ANT,ANT**, ...). If accommodation to high **GC** is insufficient, then an encoded amino acid may have to change to one of the “GARP” amino acids. If accommodation to low **GC** is insufficient, then an encoded amino acid may have to change to one of the “FYMINK” amino acids.

GC-Pressure Versus Protein-Pressure

Since it is often independent of the encoded amino acid, the third codon position should be freer than the other codon positions to respond to pressures other than those related directly to protein synthesis. As such, it serves as a ‘control’ against which to compared protein pressures on codons (i.e. the need to encode a particular amino acid at a particular position). A contrast can then be drawn between the conventional phenotype and the genome phenotype.

A way of showing this, elegantly exploited by Akira Muto and Syozo Osawa in Japan, is to plot separately the *average* base composition of *each* codon position of the set of *all* the sequenced genes of a species, against the average base composition of the genome of that species [9]. Figure 12.1 shows the three plots that result in the case of the (**G+C**)% of 1046 bacterial species [10, 11]. Each point here refers to a codon position in an individual bacterial species.

The first thing to note is the wide difference in (**G+C**)% among species, which ranges from around 20% to 80% in bacteria (see X-axis). We shall see in Chapter 14 that it is convenient to regard values around 50% as the norm, so that values below 50% indicate that there has been “downward **GC**-pressure,” and values above 50%

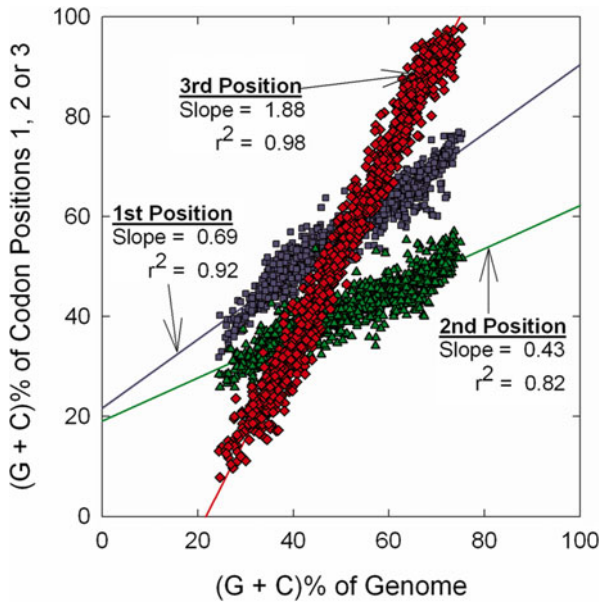


Fig. 12.1 Genomic codon position plots showing relative responses of different codon positions to GC-pressure in the 1046 species of bacteria represented in GenBank in the year 2000 by four or more genes. For each species there are 3 values corresponding to the average (G+C)% of first codon positions (blue squares), the average (G+C)% of second codon positions (green triangles), and the average (G+C)% of third codon positions (red diamonds). These three values are plotted against an estimate of average genomic (G+C)% derived from the coding regions of all sequenced genes of the species. Note that the slope values (of the regression lines) increase from 0.43 (second codon positions) to 1.88 (third codon positions), and sum to 3.00. It might appear that the points for third codon positions fit their line better than the points for first and second codon positions fit their respective lines. However, the standard errors of the mean (see Appendix A) are 2.54 (first codon position plot), 2.54 (second codon position plot) and 3.74 (third codon position plot)

indicate that there has been “upward GC-pressure.” However, it is sometimes convenient to think of bacterial species as responding to a pressure progressively to increase their overall (G+C)%. In this respect, all three codon positions participate *positively* in the species ‘response’ to GC-pressure (i.e. the slopes in Figure 12.1 are all positive). However, third codon positions are freer to respond than first and second codon positions (i.e. the plot for third codon positions has the steepest slope). Values for (G+C)% for third codon positions cover a wide range, with some species having values below 10%, and others having values approaching 100% (see Y-axis).

Thus, for genomes that are at the extremes of (G+C)% (low or high), the third codon position appears to *amplify* either a low average genomic (G+C)% by exceeding the first and second codon positions in its AT-richness, or a high average genomic (G+C)% by exceeding the first and second codon positions in its GC-richness. In some bacterial species, *every* third base in every codon is almost exclusively A or T. In others it is almost exclusively G or C. First and second codon positions are *constrained* by the pressure to encode proteins (Table 7.2), some of which may be species-specific proteins with distinctive amino acid compositions; here the range of

(G+C)% values (see Y-axis) is much narrower than in the case of third codon positions. It is a ‘zero sum game’ in that if one slope value falls, another rises.

The variations in a *large* collection of diverse proteins among 1046 species are best reflected in the gentle upward slope for second codon positions in codon position plots against genomic (G+C)%. Yet in the year 2000 the sequencing of many of the 1046 bacterial genomes (Fig. 12.1) was incomplete. In some cases only 4 gene sequences – hardly representative of a whole genome – were available. But the clarity of the result suggests that sometimes it is better to have the sequences of a few genes from many genomes than the sequences of many genes from a few genomes. Indeed, Grantham conjured up breath-taking generalizations from a mere 160 partial sequences [5]. This consideration does not apply for small mitochondrial genomes where there have long been an abundance of complete sequences.

Species Barcode in Mitochondria

The codon position plots for animal mitochondria show the same features as those of bacteria – third position plots are the most steep and second position plots are the least steep, due to the constraint of protein pressure at this position (Fig. 12.2). The second position *slope* summates inputs from genes corresponding to *all* the proteins encoded by all mitochondrial species. Some proteins differ across species more than others, and a second codon position plot that was restricted to the codons of an *individual* protein should show a slope that was *distinctive for that protein*. Proteins that could, by changing their amino acid compositions, accommodate flexibly to GC-pressure, would correspond to most steep second position slopes. Proteins that were less flexible should correspond to less steep slopes. We observe the *average* slope in Figure 12.2.

So, from gene to gene, there is variable input from proteins to the overall G+C content of mitochondrial DNA. That content value, summing inputs from all three codon positions, always seems to have at least some second position input. Yet, if there existed a protein that remained *invariant* from species to species, the second position contribution to the overall slope *from that protein* would be zero. The slopes of the first and, especially, the third positions for that gene, should then provide a relatively *unimpeded index* of the species specific component of (G+C)% – the genome phenotype. Amazingly, Paul Hebert and his colleagues in Canada found the ideal sequence in the form of a mitochondrial gene whose product – cytochrome c oxidase subunit one – needs a high degree of structural precision for the functioning of its several membrane-spanning domains. Amino acid changes are not tolerated [12]. Organisms with amino acid mutations in this sequence have been promptly counter-selected.

A study of 849 species by Xiang Jia Min and Donal Hickey [13] showed that, whereas for *all* mitochondrial proteins the second position slope is, on average, 0.44 (Fig. 12.2), in the case of a 216 codon segment from the cytochrome c oxidase subunit the slope is near zero (Fig. 12.3). Thus there has been *elimination* of the gene specificity component while *retaining* the species specificity component. This can lead with facility from an individual gene to the identification of its host species. This sequence from the gene in a given species can act as a ‘barcode’

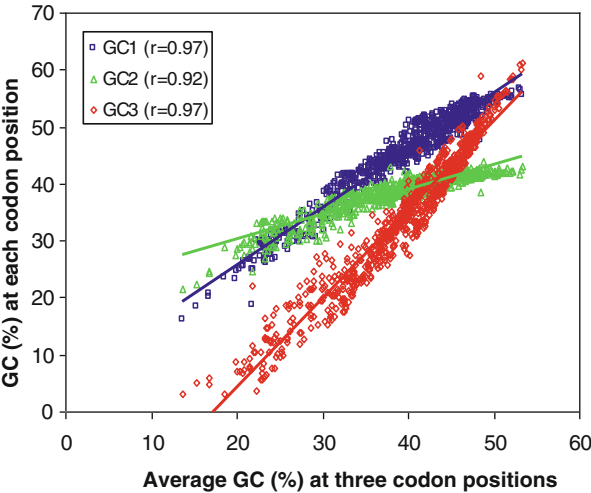


Fig. 12.2 Genomic codon position plots for all codons in 849 animal mitochondria. The slopes of first, second and third codon positions are 1.01, 0.44 and 1.55, respectively (summing to 3.00). Each point corresponds to a codon position in an individual species. For details see Figure 12.1. Reproduced with permission from reference [13]

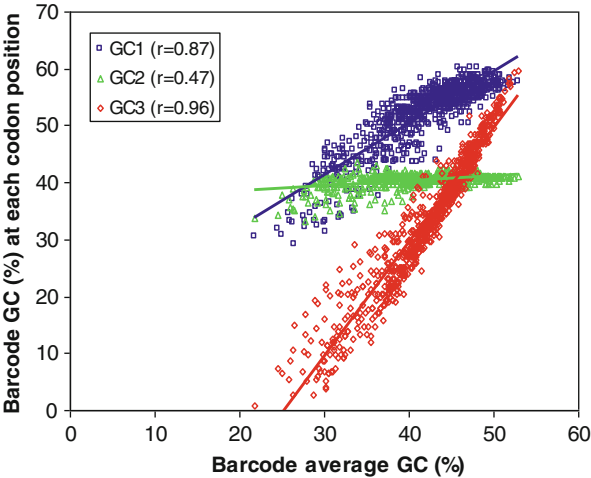


Fig. 12.3 Codon position plots for segments of 216 codons from the cytochrome c oxidase sub-unit gene for each of the 849 animal mitochondrial genomes shown in Figure 12.2. The slopes of first, second and third codon positions are 0.90, 0.09 and 2.00, respectively (summing to 3.00). Each point corresponds to a codon position in the gene of an individual species. Reproduced with permission from reference [13]

permitting rapid identification of that species. Mitochondrial DNA evolves more rapidly than nuclear DNA (Chapter 6), so DNA differences between organisms from closely related species are likely to be found first in their mitochondria.

The 216 codon sequence from the above “sentinel gene” is proving a powerful tool for classifying organisms, and also helps identify questionable foodstuffs. Apart from these practical uses, studies of differences between species of birds, which vary greatly in population sizes, indicate that mitochondria are subject to *extreme* purifying (natural) selection that affects *all* codon positions; this challenges the “neutral” hypothesis much favored by population geneticists [14].

Species Versus Genes

Whereas Figures 12.1 to 12.3 show the variation of base composition for different codon positions among entire *genomes* (i.e. among species), Figure 12.4 shows this for different codon positions among the sequenced *genes* of two bacterial species, one of very low, and the other of very high, genomic (G+C)%. In Figures 12.1 and 12.2 each point represents an average value for all sequenced genes of a species, so each point represents a species. Likewise, in Figure 12.3 each point represents a value for one gene in a species so, again, each point represents a species.

However, each point in Figure 12.4 represents an average value for an individual gene *within* a species, so each point represents a *gene*. Being at the extremes, the (G+C)% values for all codon positions tend to be low in genes of the low (G+C)% species (so all points tend to the bottom left corner of the figure), and high in genes

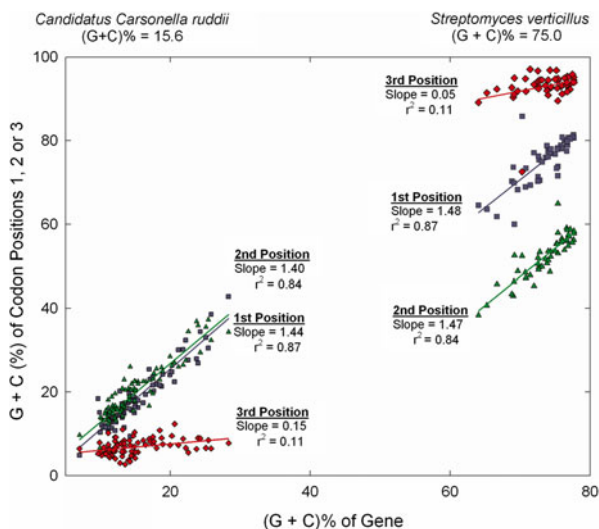


Fig. 12.4 Genic codon position plots showing responses to GC-pressure of different codon positions within individual genes for a low genomic (G+C)% species (15.6% average G+C), and for a high genomic (G+C)% species (75.0% average G+C). Each point represents the average (G+C)% value for a particular codon position in all the codons of a particular gene (i.e. there are three points for each gene – a square, a triangle and a diamond)

of the high (G+C)% species (so all points tend to the top right corner). Yet there is still much differentiation among genes. Of note is that, whereas slopes for genic first and second codon positions are steep, slopes for genic *third* codon positions are almost *flat*. Thus, within each of these two genomes, genes differ from each other in (G+C)% mainly in their first and second codon positions. This would largely reflect the needs of the encoded proteins. In contrast, within a genome *all* genes have approximately the *same* (G+C)% in their third codon positions.

The (G+C)% values of third codon positions would seem to be *dedicated* to the roles of keeping genomic (G+C)% low in the low (G+C)% species, and high in the high (G+C)% species. This emphasizes the independence of third codon positions, which appear to serve exclusively the needs of GC-pressure (i.e. *the needs of the species*), not the needs of the genes that contain them. Whatever the gene, its third codon position (G+C)% value is essentially the *same* as that of other genes of the same species. Thus, for species at the extremes of genomic (G+C)% (either low or high), the distinctiveness of the (G+C)% of a particular gene is largely determined by the (G+C)% of first and second codon positions. Yet, despite the tendency towards uniformity by virtue of constant third codon position (G+C)% values, there is some differentiation of genes (into distinct “microisochores;” Chapter 11).

The range of this differentiation is limited. In the low (G+C)% species the range varies from 5-30%. In the high (G+C)% species the range varies from 60-80% (see X-axis of Fig. 12.4). Nevertheless, having decreased the species-specificity component (near-zero third codon position slopes), the (G+C)% value of a small genomic segment (a *genic* ‘bar code’) would allow rapid identification of a gene in a genome (if this were so desired).

But most species have intermediate genomic (G+C)% values. One might predict that plots for genic third codon positions would again be almost flat, so that the needs of the species would again dominate. This is not so. In fact, *the very opposite is observed*. Slopes for genic third codon positions are *steeper* than slopes for genic first and second codon positions. The gut bacterium *E. coli*, provides a good example. Figure 12.5, which shows average codon position (G+C)% values for *each* of the genes of *E. coli*, appears like Figure 12.1, which shows average values for the *set* of all the sequenced genes of each of a wide range of bacterial species. Furthermore, the genes differ from each other over a *wide* range of (G+C)% values (25% to 70%). Thus, the dispersion of (G+C)% values among *genes within a genome* (Fig. 12.5) matches the dispersion of (G+C)% values among *genomes within a taxonomic group* (Fig. 12.1). Again, the ‘zero sum’ rule (limit 3.0) applies.

Guided by their “homostability” principle (Chapter 11), Wada and his coworkers in 1991 drew attention to this similarity between codon position plots for genes, and codon position plots for genomes [15]. The plots are essentially context-independent (i.e. the same both for genes and for genomes). They concluded that the plots “might be universal ones and the constraint parameters might have general biological meanings in relation to the DNA/RNA and protein functions.” Thus, they were invoking some scale-free common principle – homostability – that was likely to be of fundamental significance. *Whatever was causing the dispersion among genes might also be causing the dispersion among genomes.*

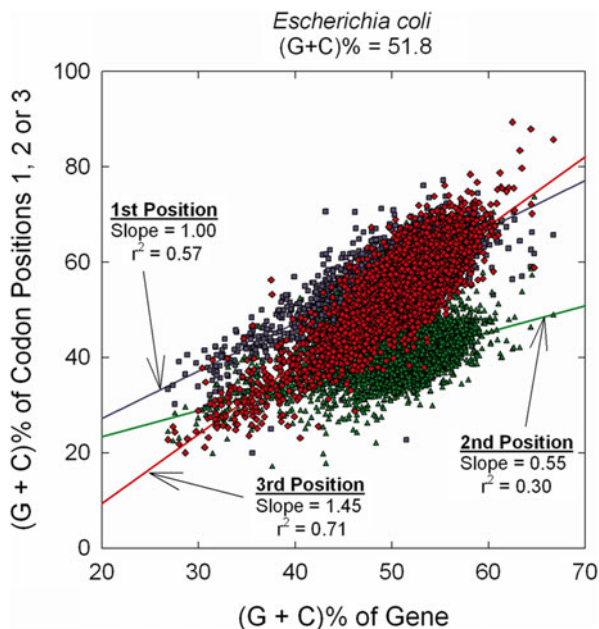


Fig. 12.5 Genic codon position plots showing responses to GC-pressure of different codon positions for each of the 4290 genes of an intermediate genomic (G+C)% species (*E. coli*). Each point represents the average (G+C)% value for a particular codon position in all the codons of a particular gene (i.e. there are three points for each gene). The standard errors of the mean (see Appendix A) are 4.20 (first codon position plot), 4.07 (second codon position plot) and 4.53 (third codon position plot). As usual, slope values sum to 3.0

Wada's ideas seemed to conflict with the “neutral” hypothesis of his compatriot, Kimura (Chapter 5), and he had little success in getting an automated DNA sequencing project launched in Japan [16]. Subsequently, “the genome project” emerged in the USA, culminating (2003) in what were close to the entire sequences of our haploid twenty three human chromosomes. The DNAs used were from unique individuals. Now there are more to compare.

Species Win at (G+C)% extremes

In species with *extreme* genomic (G+C)% values, the slope values for genic first and second codon positions exceed those for third positions (Fig. 12.4). The opposite holds for species with *intermediate* genomic (G+C)% values (Fig. 12.5). Thus, at extreme (G+C)% values, in addition to encoding amino acids, first and second positions are assigned the extra role of differentiating genic (G+C)%. But at intermediate (G+C)% values, third positions assume much of this burden and first and second positions are relatively unhampered in their ‘desires’ to encode amino acids. Accordingly, at some (G+C)% values between extreme and intermediate, *the*

patterns must switch. This switch-over is shown when the three slope values taken from genic codon position plots for numerous individual bacterial species (of which Figs. 12.4 and 12.5 are examples) are plotted against the genomic (G+C)% values for those species. The switches occur at 37% and 68% (Fig. 12.6).

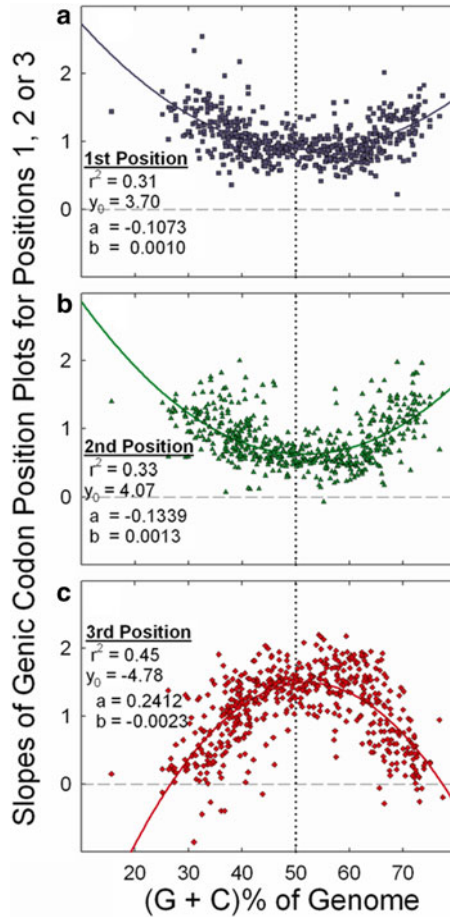


Fig. 12.6 Slopes of genic codon position plots for first and second codon positions (a, b) are low in species of intermediate genomic (G+C)%, and high in species with extreme genomic (G+C)%. Slopes for third codon positions (c) show the opposite trend. Slope values are from individual genic codon position plots for each of 546 bacterial species (each with at least 20 sequenced genes). Slope values were fitted to second order regression curves (see Appendix A). At intermediate (G+C)%, differentiation of genes for protein-encoding (dependent on first and second positions) is reflected in slope values of 0.9 (or less) for these positions (a, b). It is unlikely that, in extreme (G+C)% species, genic differentiation for the protein-encoding function would be greater than in intermediate (G+C)% species. Yet, in extreme (G+C)% species, slope values are around 1.5 in plots for first and second positions (a, b). The extra 0.6 slope units would reflect the need of genes in extreme (G+C)% species to differentiate their (G+C)% values (to prevent intergenic recombination). This differentiation, as reflected in first and second position values, would be less in species of intermediate (G+C)%, since here third codon positions (c) play the major role in achieving genic (G+C)% differentiation. If superimposed, the red curve crosses the others at 37% and 68%

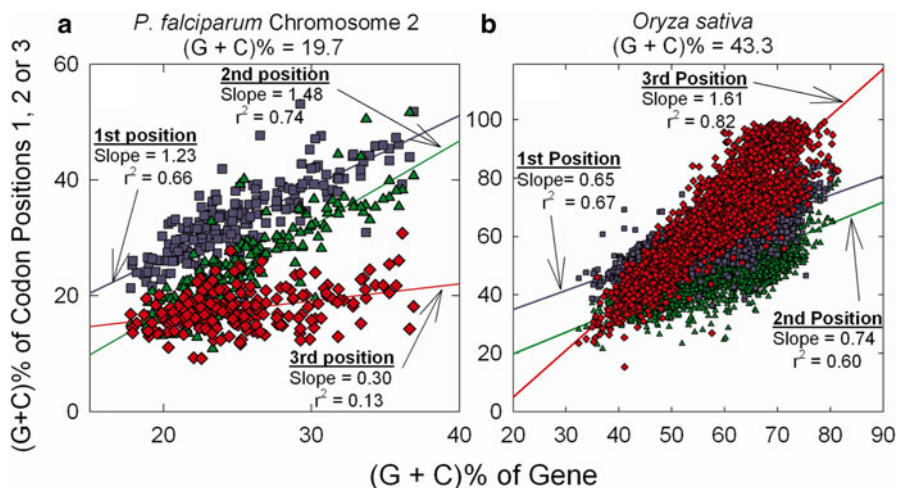


Fig. 12.7 Genic codon position plots of eukaryotes resemble those of prokaryotes of the same average genomic (G+C)%. Examples are (a) 205 genes from the second chromosome of the malaria parasite (*P. falciparum*; 19.7% average G+C), and (b) 3111 genes from the genome of the rice plant (*Oryza sativa*; 43.3% average G+C)

Because they cover a wide range of genomic (G+C)% values, bacterial species are ideal for studies of this nature. However, similar results are found in ‘higher’ organisms. Figure 12.7a shows that the low (G+C)% genome of the malaria parasite *P. falciparum* gives essentially the same result as the low (G+C)% genome of a bacterium (Fig. 12.4). Figure 12.7b shows that the intermediate (G+C)% genome of rice (*Oryza sativa*), gives essentially the same result as the intermediate (G+C)% genome of a bacterium (Fig. 12.5).

Hence, it is generally true that species with extreme genomic (G+C)% values (low or high) sustain those values largely by maintaining constant, extreme, third codon position (G+C)% values, which may be dubbed “gene-independent” in that they are only slightly influenced by the gene that contains them. Values for first and second codon positions, that specify amino acids, are more gene-dependent, and do not so obviously contribute to the extreme (G+C)% values of the species. The situation changes dramatically in species with intermediate genomic (G+C)% values. All three codon positions assume gene-dependent (G+C)% values, and this is most evident in the case of third codon positions (Figs. 12.5, 12.7b).

So, if first and second codon positions within a gene have low (G+C)% values, then the corresponding third codon positions will have even lower (G+C)% values (i.e. they are richer in A and T). If first and second codon positions within a gene have high (G+C)% values, then the corresponding third codon positions will have even higher (G+C)% values (i.e. they are poorer in A and T). Accordingly, a low (G+C)% gene has many AT-rich codons such as TTY (generic codon for phenylalanine). Here TTT is preferred over TTC. On the other hand, a high (G+C)% gene has many GC-rich codons such as GGN (generic codon for glycine). Here GGC and GGG are preferred over GGT and GGA.

Thus, third codon positions support and extend the $(G+C)\%$ values of first and second codon positions. As stated by Wada and Suyama in 1985 [17]:

The average $(G+C)$ -content of each gene is altered mainly by base substitution at the third site of the codon and ... from the positive correlation between ... [the three] sites, which is revealed when they are averaged over the entire length of a gene [see Fig. 11.3], every base in a codon seems to work cooperatively toward realizing the gene's characteristic value of $(G+C)$ -content.

In this way, each gene in a species of intermediate genomic $(G+C)\%$ has come to occupy a more distinctive **GC**-niche (microisochore) amongst its fellow genes. In contrast to genes in species of extreme genomic $(G+C)\%$, genes in species of intermediate genomic $(G+C)\%$ collectively span a wide $(G+C)\%$ range [18].

Intermediate $(G+C)\%$ Allows Compromise

Why are species with extreme genomic $(G+C)\%$ values so different? It is intuitively apparent that a sequence with only one base would have less ability to transmit information than a sequence with four bases. In fact, elementary information theory shows that information transmitting potential is maximized when the four bases are present in equal proportions (i.e. when PR2 is obeyed; see Chapter 14). Thus, departures from 50% (downwards or upwards) are likely to progressively compromise the ability of genomes to transmit *further* information. The trade-off seems to be that, at extreme genomic $(G+C)\%$ values (below 37% and above 68%), third codon positions serve the information needs of the species, rather than of individual genes within that species. Why *further* information? It is likely that departures from 50% *themselves* have information content, which, at the extremes, happens to be of more relevance to the species than to individual genes (Chapter 7).

In species with intermediate genomic $(G+C)\%$ values, the contribution to genic **G+C** is shared by all codon positions, with third positions making the most distinctive contributions (i.e. slope values are greatest for third positions; Fig. 12.6c). In species with extreme genomic $(G+C)\%$ values, third positions contribute little to distinctions in genic **G+C**; but relative to species with intermediate genomic $(G+C)\%$ values, first and second codon positions *increase* their contributions (Figs. 12.6a, b). Thus, the slope values for the first and second codon positions in species with extreme genomic $(G+C)\%$ values are *greater* than in species with intermediate genomic $(G+C)\%$ values. This provides for some genic differentiation into microisochores in extreme $(G+C)\%$ genomes. However, the range of this genic differentiation (see X-axis values) is less than the range of genic differentiation in intermediate $(G+C)\%$ genomes. The position of a gene in that restricted range in extreme **GC**-species is partly determined by protein-pressure (the pressure to encode a particular protein) and partly determined by local *genic* pressure for $(G+C)\%$ differentiation as a microisochore distinct from other microisochores (genes).

While the nature and location of certain amino acids in a protein may be critical for protein function, the nature and location of certain amino acids encoded by genes in extreme **GC**-species is also strongly influenced by the general *genomic* pressure for that **(G+C)%**. Along similar lines, we shall see here and in Chapter 14 that sometimes base composition pressures cannot be accommodated without changing proteins, sometimes even to the extent of *increasing* protein lengths by inserting ‘placeholder’ amino acids. The increases are sometimes evident as low complexity segments that primarily serve the needs of the encoding nucleic acid, not of the encoded protein.

Levels of Selection

The plots shown above illustrate the separate operations of *gene-level selection* and *genome-level selection* – subjects of much controversy and semantic confusion (Chapter 9). The conflict between ‘selfish genes’ and ‘selfish genomes’ is reflected in the differential survival (reproductive success) of individuals (*individual level selection*) that, when the crunch comes (i.e. extreme genomic **(G+C)%** values), can serve the species (i.e. ‘selfish genomes’). The quotation marks indicate that genes and genomes are not consciously selfish, but the dynamics of their evolution easily lends itself to this colorful implication (see Chapter 15).

Does natural selection act both on individuals and on species? There is sometimes talk of “species selection, in which whole species are the target of natural selection” [19]. Clearly, if the last few members of a waning species are collectively struck by a meteor, then the species is negatively selected by a ‘natural’ agency. The agency might equally well have been a predator that quickly eliminated them one at a time, and the distinction is rather pedantic. It is even predicted that: “Species that split into new species faster than, or become extinct slower than, other species will become more common” [19]. But, what is usually the most important *biologically* is not the number of species, but the number of individuals within each of those species. A species can diverge into several new species, and the members of each can come to occupy distinct ecological niches that each supports, say, a million individuals. It is a moot point whether a group of species collectively comprising several million individuals should be considered more successful *biologically* than an individual species of, say, one hundred million individuals.

A species that is better able to diverge into new species than another species, may end up containing more individuals in its species collective (genus) than if it had not diverged. But here natural selection still has acted on individual members of each species, not on the species themselves. It is individual *members* of a species, not individual species, that usually become “more common” by virtue of natural selection. Species themselves can become “more common” (i.e. give rise to more *child species*) when there is more effective *reproductive isolation*. This agency corresponds either to physiological selection (intrinsic cause) or natural selection (extrinsic cause; see Part III).

Dog Wags Tail

In the context of the ‘holy grail’ of Romanes and Bateson (Chapter 10), variations (mutations) of importance for achieving differentiation of species (reproductive isolation) are often independent of variations that affect genes. But **GC**-pressure acts both on intergenic DNA and genic DNA. To the extent that a gene does not ‘want’ to change the nature of its products (RNA and/or protein) it must accommodate to the pressure for their continued faithful elaboration – for example, by accepting only a synonymous mutation that does not change the nature of an encoded amino acid. To this extent the synonymous mutation, although localized in a segment we decide to call a gene (i.e. it is *genically* located), can be viewed as *functionally* “non-genic.”

On the other hand, a mutation primarily affecting a gene could locally distort, thus marginally affect, the **(G+C)%** of a genome. If natural selection leads to an accepted genic change, meaning a changed gene product, then there may be pressure for some compensatory local increase or decrease in **(G+C)%**. Thus, while both functionally and locationally “genic,” the mutation, albeit only marginally, would also be functionally “non-genic” [20].

As part of a speciation event, two divergent lines may progressively differ in **(G+C)%** and the parts of genes under functional constraint (e.g. first and second codon positions) will tend to oppose this. Accordingly, the rates of mutation acceptance at first and second codon positions will usually be less than the rates of mutation acceptance at third codon positions, introns, and intergenic DNA (Fig. 2.5). Exceptions may be genes (i.e. nucleic acid) that encode proteins whose substrate (i.e. target) is *itself* nucleic acid. While proteins can recognize nucleic acids generically by virtue of their common phosphate-ribose structures, many proteins are not indifferent to the composition or order of bases. So, as nucleic acids change in **(G+C)%**, the proteins that recognize nucleic acids may need to adapt functionally to ensure optimal binding to their DNA substrate. They are on an *adaptive treadmill* (Chapter 5). This implies that, when lines are diverging, the mutation rate in nucleic acid recognition proteins should be greater than in most other proteins, as is indeed found [21].

Thus, **GC**-pressure (‘dog’), while causing mutation of both non-genic and genic DNA, particularly impresses the DNA of DNA-recognition enzymes (‘tail’). Conversely, it might seem possible that mutations in genes encoding certain general DNA-recognition enzymes (e.g. DNA repair enzymes) might lead to directional changes in genomic **(G+C)%**. That the tail can wag the dog is indeed true in the case of certain ‘mutator’ genes [8]. Since these genes have a genome-wide effect on **(G+C)%**, it is tempting to consider them as candidate “speciation genes.” However, such genes are likely to be exceptions rather than the rule. The small regions of genomic DNA encoding DNA-recognition proteins (‘tail’) must, like other regions, adapt to pressure for overall changes in the large lumbering genomic **(G+C)%** (‘dog’), but are unlikely themselves to drive general **(G+C)%** changes.

AG-Pressure Versus Protein-Pressure

The responsiveness of different codon positions in bacterial genomes to **AG**-pressure (purine-loading pressure) is shown in Figure 12.8. Following Figure 12.1, average **(A+G)%** values of individual codon positions of a species are plotted against the corresponding genomic **(A+G)%** of the species. The range of genomic values is much less than in the case of **GC**-pressure (compare X-axis values with those of Fig. 12.1).

In keeping with the **RNY** rule (Chapter 7), for all species the first codon position makes the greatest contribution to genomic purine-loading, and no species has a genomic **(A+G)%** value less than 50 for this codon position (i.e. there is no pyrimidine-loading). Few bacterial genomes achieve these high values in the second and third positions.

Also in keeping with the **RNY** rule, third codon positions are predominantly pyrimidine-loaded (values less than 50%). This purine-poor position would seem

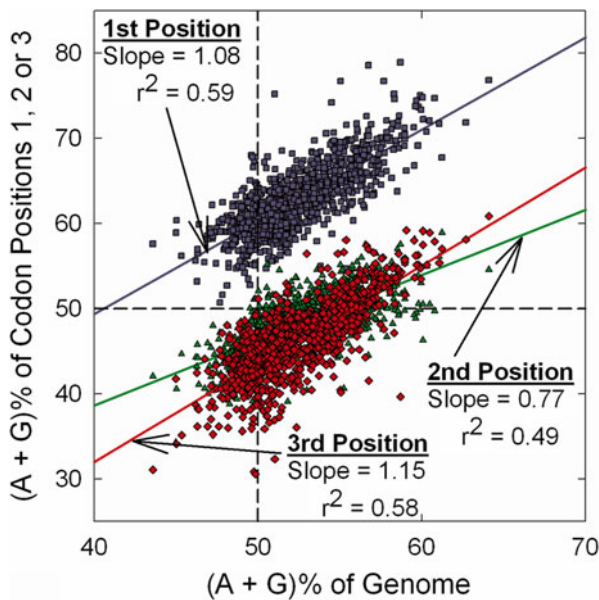


Fig. 12.8 Genomic codon position plots showing relative responses of different codon positions to **AG**-pressure in 1046 species of bacteria as in Figure 12.1. For each species there are 3 values corresponding to the average genomic **(A+G)%** of first codon positions (blue squares), the average genomic **(A+G)%** of second codon positions (green triangles), and the average genomic **(A+G)%** of third codon positions (red diamonds). These three values are plotted against an estimate of the average **(A+G)%** of the genome derived from the coding regions of all the sequenced genes. Note that slope values increase from 0.77 (second codon positions) to 1.15 (third codon positions) and, as usual, sum to 3.00

to be best poised to respond to **AG**-pressure. Indeed, from the slopes of the plots, third positions appear most responsive to **AG**-pressure so that, in species with high genomic (**A+G**)% values, third codon position values can exceed 50%. Thus, there is purine-loading, so that **RNY** becomes **RNR**, and the **RNY** rule no longer holds. This may be construed as a conflict between purine-loading pressure (“**R**-pressure”) and “**RNY**-pressure,” which is settled in these species in favor of purine-loading pressure.

As in the case of the response to **GC**-pressure (Fig. 12.1), the response to **AG**-pressure is least in the second codon position (slope value of only 0.77). However, species with high genomic (**A+G**)% values have proteins enriched in amino acids encoded by purine-rich codons [22]. This may be construed as a conflict between purine-loading pressure and protein-encoding pressure that is settled in these species in favor of purine-loading pressure, a predominant pressure in thermophiles (see later).

GC-Pressure Versus AG-Pressure

The content of (**G+C**) in a sequence segment is not independent of its content of (**A+G**) since both compositions include the base **G**. Indeed, there is a potential conflict between **AG**-pressure and **GC**-pressure (Fig. 12.9).

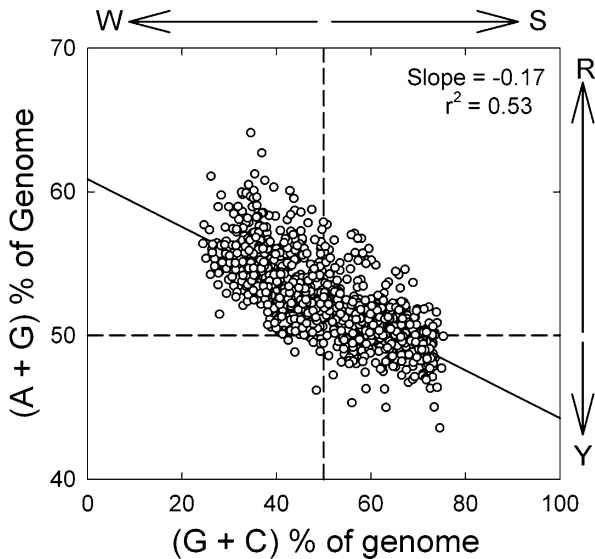


Fig. 12.9 Inverse relationship between average (**A+G**)% and average (**G+C**)% in the codons of 1046 species of bacteria. Each point represents an individual species. Horizontal arrows (top) indicate values departing from 50%: ‘upward pressure’ towards high (**G+C**)% values (i.e. high **S** values) and ‘downward pressure’ towards low (**G+C**)% values (i.e. high **W** values). Likewise, vertical arrows (right) indicate values depart from 50%: pressures towards purine-loading (**R**), and pyrimidine-loading (**Y**)

There are two ways to modulate genomic (G+C)% when the total number of bases is constant – either by changing the number of G's, or by changing the number of C's. If an increase in the number of G's is matched by a decrease in the number of C's, then there will be no overall change in (G+C)%. If one of these bases increases more than the other decreases, then there will be an overall increase. For such an increase, there must be a corresponding decrease of A's and/or T's, and other pressures may be affected. As genomic (G+C)% increases, replacing A with G, or T with C, would not affect the genomic AG-pressure (i.e. there would be no change in total purines). But if T were replaced with G, AG-pressure would increase as (G+C)% increases. Conversely, if A were replaced with C, AG-pressure would decrease as (G+C)% increases.

The latter inverse relationship is observed in bacterial genomes (Fig. 12.9), and in their viruses (bacteriophage), and in the several eukaryotes that have been examined. This indicates *a preference for trading A for C as genomic GC content increases* [23]. Why should this be? Perhaps there is a selection pressure to avoid clusters (runs) of consecutive Gs or Ts. The former might generate an excess of “G-quadruplex” structures in DNA (see Chapter 14). The latter might predispose to cross-linking of neighboring T bases (thymine-dimers), which damages DNA.

In the survey of 1046 bacterial species, 53% of the variation ($r^2 = 0.53$) could be explained in terms of the inverse relationship between (A+G)% and (G+C)%. To what extent do different codon positions contribute to this trading relationship? This can be examined by plotting the result, shown in Figure 12.9, for each codon position individually (Fig. 12.10).

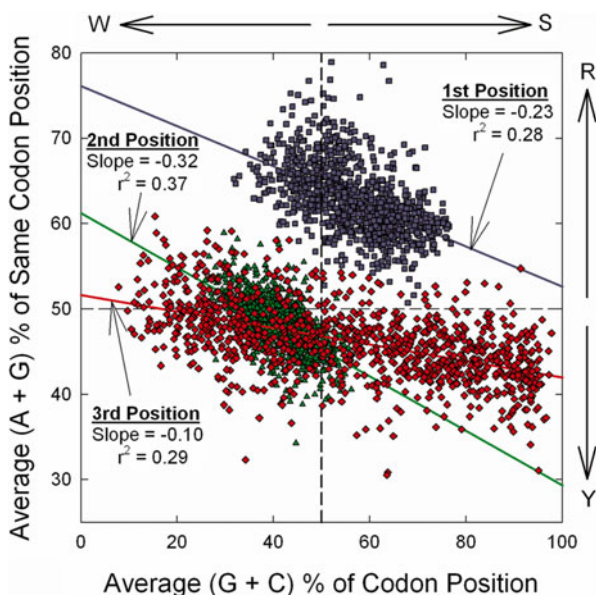


Fig. 12.10 Influence of GC-pressure on purine-loading at different codon positions. For each bacterial species, average (A+G)% values for each codon position are plotted against average (G+C)% values of the *same* codon position. Symbols are as in Fig. 12.8

Peering beneath the fog of multiple data points it can be seen that, being already highly **Y**-loaded in accord with the **RNY**-rule, genomic third codon positions are less in a position to give up **A** and acquire **C** as **(G+C)%** increases. So third codon positions appear less responsive to this aspect of **GC**-pressure (slope of only -0.10) than the first and second codon positions.

By the same token, being already highly **R**-loaded in accord with the **RNY**-rule, genomic first codon positions should best be able to trade **A** for **C**. But genomic *second* codon positions, the positions of *most* importance in determining the amino acid and *least* responsive to **GC**- and **AG**-pressures individually (Figs. 12.1 and 12.8), are *more* responsive than first positions in the trading relationship (slopes of -0.32 versus -0.23).

In other words, at extreme **(G+C)%** values (low and high), **AG**-pressure is modulated by the introduction (when **(G+C)%** is low), or the removal (when **(G+C)%** is high), of amino acids corresponding to purine-rich codons. Do these amino acids necessarily serve protein function, or are they sometimes mere ‘placeholders’?

AG-Pressure and ‘Placeholder’ Amino Acids

The reciprocal relationship between **GC**-pressure and **AG**-pressure suggests caution when trying to attribute a given effect to just one of these pressures. For example, in animals **GC**-richness and protein-shortness go together. There is a weak *negative* relationship between the **(G+C)%** of protein-encoding regions (exons) and the lengths of the proteins in those regions [24]. This is seen clearly in the case of proteins encoded by the malarial parasite *P. falciparum*. In Fig. 12.11a **(G+C)%** values of exons are plotted against the corresponding lengths in kilobases. Note that 5% of the variation in **(G+C)%** can be explained on this basis ($r^2 = 0.05$; $P = 0.0009$). When the same lengths are plotted against **(A+G)%** the relationship is *positive* and 7% of the variation can be explained ($r^2 = 0.07$; $P < 0.0001$; Fig. 12.11b). Thus, if you had to suggest which pressure might be *primarily* related to protein length you would prudently pick **AG**-pressure.

The relationship with **AG**-pressure becomes more apparent when the base compositions of different codon positions are plotted against the lengths of the corresponding protein-encoding segments. In the case of **GC**-pressure (Fig. 12.11c), first and third codon position values change little as protein lengths increase. Second codon position values show the best correlation with protein lengths (slope = -1.0%/kilobase), and 7% of the variation can be explained on this basis ($r^2 = 0.07$).

In the case of **AG**-pressure (Fig. 12.11d) the **RNY** rule prevails. In nucleic acid segments encoding both short and long proteins, first codon positions tends to be **R**-rich (i.e. high **(A+G)%**), and third codon positions tend to be **Y**-rich (i.e. low **(A+G)%**). The slopes, while modest, both support these trends. First codon position slopes are positive, indicating increasing **R**-loading. Third codon position slopes are negative, indicating increasing **Y**-loading. But it is second codon positions that are most responsive.

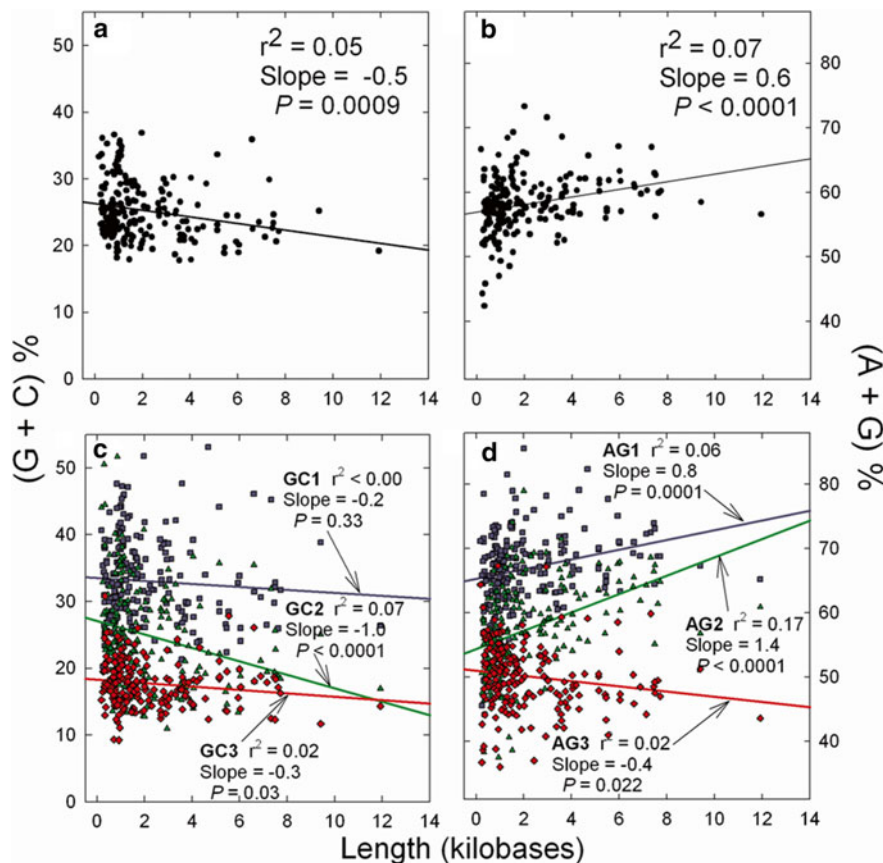


Fig. 12.11 Protein length is inversely proportional to (G+C)% and directly proportional to (A+G)% in 205 proteins encoded by chromosome 2 of *P. falciparum*. Each point refers to the protein-encoding region of a particular gene. (a) (G+C)% for all bases encoding each protein. (b) (A+G)% for all bases encoding each protein. Corresponding values for the three individual codon positions in each protein are shown for (G+C)% in (c), and for (A+G)% in (d). First codon positions (GC1, AG1; blue squares), second codon positions (GC2, AG2; green triangles), and third codon positions (GC3, AG3; red diamonds). Slopes are expressed in percentage units/kilobase of protein-encoding sequence

Second codon positions appear much freer to vary their (A+G) % values as the lengths of protein-encoding regions increase (slope = 1.4%/kilobase), and 17% of the variation can be explained ($r^2 = 0.17$). Since second codon positions are most affected, it can again be inferred that *base compositional pressures have caused amino acid-changing (non-synonymous) mutations to be accepted. Increased protein lengths appear to reflect the insertion of amino acids with R-rich codons, at the expense of S-rich (i.e. GC-rich) codons.*

In Figure 12.11 slopes differ from zero because, on average, when their base compositions are presented as percentages, nucleic acid segments encoding large proteins

somehow differ from nucleic acid segments encoding small proteins. So why cannot small protein-encoding segments contain proportionately as many **R**-rich codons as long protein-encoding segments? Then the slopes would be close to zero.

As considered later (Chapters 13 and 14), the probable explanation is that proteins tend to fold into a number of independent, highly structured, functional *domains*. Each domain may involve from 60 to 300 contiguous amino acids. Small proteins tend to have just one domain. Large proteins tend to have many domains, and it is the relatively disordered *space between domains* that most readily accepts insertions of ‘placeholder’ amino acids. These may not necessary play a role in the normal functioning of the protein. They may be there *by default* because the corresponding genes have not ‘found’ how to cater to the demands of nucleic acids in any way other than by adding interdomain codons.

By the same token, by virtue of their possession of more interdomain regions, genes encoding long proteins should more readily respond to pressures to mutate. For example, the preservation of a gene duplicate requires that it quickly mutate to avoid recombination with the source gene (see Fig. 11.1). This suggests that genes encoding long proteins might more readily form duplicates than genes encoding short proteins.

Thermophiles

We have already noted for prokaryotes that optimum growth temperature correlates both with $(\mathbf{A}+\mathbf{G})\%$ and $(\mathbf{G}+\mathbf{C})\%$, but better with the former (Chapter 11). Figure 12.12 shows this at the level of individual bases. Although an increase in $(\mathbf{G}+\mathbf{C})\%$ can stabilize RNAs that must function by virtue of their structure, there is no consistent trend towards a high genomic $(\mathbf{G}+\mathbf{C})\%$ in thermophiles. This may partly reflect a conflict with genomic **AG**-pressure, which increases with optimum growth temperature, and is accompanied by a small, but significant, decrease in genomic $(\mathbf{G}+\mathbf{C})\%$ (Fig. 11.4). Hence, a causal chain of evolutionary events might be: (i) need to adapt to high environmental temperature, (ii) purine-loading (perhaps to diminish RNA-RNA interactions; see Chapter 15), and (iii) decrease in $(\mathbf{G}+\mathbf{C})\%$. And, as might be expected from the decrease in genomic $(\mathbf{G}+\mathbf{C})\%$ as optimum growth temperature increases, the increase in genomic $(\mathbf{A}+\mathbf{G})\%$ mainly reflects an *increase in base A at the expense of base C*. Thus, the levels of the bases **T** and **G** remain relatively constant as the optimum growth temperature increases.

Since thermophiles strongly purine-load their RNAs (Fig. 6.2), it might be suspected that the interdomain regions of their proteins would have many placeholder amino acids encoded by purine-rich codons. But here another intra-genome conflict arises. The demand of stability at high temperatures tends to exclude large interdomain regions from the proteins of thermophiles. Thus, the extra purines must be accommodated within intradomain regions [24–26].

When purified, the proteins of thermophiles can be shown to be more thermally-stable than the corresponding proteins from non-thermophiles. This could be due to the *nature* of the encoded amino-acids (i.e. a bias in amino acid composition that

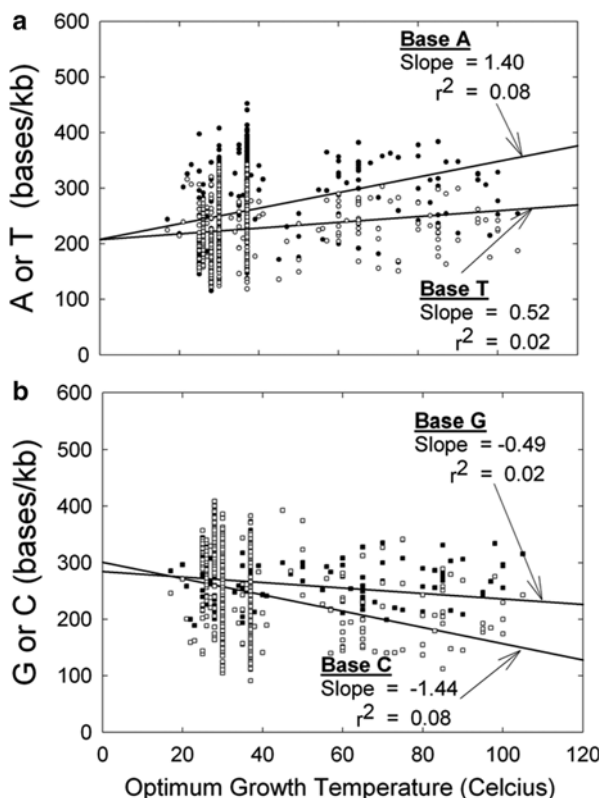


Fig. 12.12 A-for-C trading as optimum growth temperature increases. Variation of the frequencies of bases in the DNA of the coding regions of genes of each of 550 prokaryotic species (eubacteria and archaeobacteria) with the temperature require for optimal growth of that species. Each point represents a species. (a) The **W** bases: **A**, filled circle; **T**, open circle. (b) The **S** bases: **G**, filled square; **C**, open square. All slope values are significantly different from zero ($P < 0.001$). The two slope values in each graph are significantly different from each other; in (a) $P = 0.0005$, and in (b) $P = 0.0001$

favors certain amino acids), or to *the way they are arranged* (i.e. sequence). But the codons required for the amino acids might conflict with the more diffuse purine-needs of mRNAs. Thus, it is tempting to suppose that protein stability is achieved mainly by the way amino acids are arranged. Whether the needs of nucleic acid or of protein have been most influential in driving base composition in thermophiles is a subject of debate [27]. The studies of Supek and his team (Chapter 10) imply that nucleic acid compositions are often primary to amino acid compositions. Indeed, the correlation between the thermostability of an entire organism and that of its proteins, when purified, is not always high [28]. Furthermore, from a study of the purine-rich sequences encoding an enzyme (xylanase), it was concluded that “the codons relating to enzyme thermal property are selected by thermophilic force at [the] nucleotide level,” not at the protein level [29].

When a similar question arose on base composition in a malaria parasite, the trade-off was again in favor of nucleic acids (see Chapter 14). Finally, we should note that thermophilia, while not so extreme, is not confined to prokaryotes. Supporting preliminary studies [30], there are some thermophilic eukaryotes where, again, extreme **R**-loading is found [31].

We shall see that this theme of conflict *within* genomes sheds fresh light on the discoveries that protein-encoding sequences include non-coding gaps (introns) that must be removed prior to translation (see Chapter 13), and that other protein-encoding sequences include strange low complexity elements that are not removed prior to translation, and so become part of the final protein product (see Chapter 14).

Summary

The composition and order of bases in the nucleic acids of members of a particular biological species reflect the demands both of the genome phenotype (e.g. **GC**-pressure) and of the classical phenotype (e.g. protein-pressure – the pressure for a particular composition and ordering of amino acids in a protein). Base compositions of third codon positions, being less constrained by protein-pressure, can provide indices of non-classical pressures, so facilitating both the identification and the calibration of conflicts. Since changes in genetic fitness (number and reproductive health of descendents) can result from changes in the classical phenotype and/or the genome phenotype, a mutation appearing neutral with respect to the classical phenotype can nevertheless affect fitness by changing the genome phenotype – and vice-versa. Within a group of species (taxonomic group), *genomic* (**G+C**)% values can cover a wide range that is most evident at third codon positions. However, within a genome, *genic* (**G+C**)% values can also cover a wide range that is, again, most evident at third codon positions. The dispersion of (**G+C**)% values among *genes within a genome*, matches the dispersion of (**G+C**)% values among *genomes within a taxonomic group*. This suggests that both dispersions are driven by a common factor – the need to *avoid* recombination. This would facilitate duplication, either of genes (gene duplication), or of genomes (speciation or polyploidy). Third codon positions usually *relate more to the genes that contain them than to the species*. However, extreme (**G+C**)% values (low or high) constrain the potential of a genome to transmit information that is strictly pertinent to individual genes. Third codon positions then tend to maintain a constant (**G+C**)%, *thus relating more to the species than to the genes that contain them*. Each gene in an intermediate (**G+C**)% genome has come to occupy a discrete (**G+C**)% niche (microisochore) amongst its fellow genes, which collectively span a wide (**G+C**)% range. Genes in extreme (**G+C**)% genomes rely mainly on first and second codon positions for differentiation as microisochores, which collectively span a narrow (**G+C**)% range. These observations reveal a bioinformatic basis for both group selection (speciation) and the utility of a short mitochondrial DNA segment as a species ‘barcode.’

References

1. Holliday R (1968) Genetic recombination in fungi. In: Peacock WJ, Brock RD (eds) *Replication and Recombination of Genetic Material*. Australian Academy of Science, Canberra, pp 157–174 [The junction between recombining DNAs is named after this author.]
2. Galton F (1876) A theory of heredity. *Journal of the Anthropological Institute* 5:329–348 [He had trouble thinking that “germs” (genes) could be interrelated linearly: “It is difficult to suppose the directions of the mutual influences of the germs to be limited to lines, like those that cause the blood corpuscles to become attached face to face, in long rouleaux, when coagulation begins.”]
3. Bateson G (1979) *Mind and Nature. A Necessary Unity*. E. P. Dutton, New York, p. 21
4. Schaap T (1971) Dual information in DNA and the evolution of the genetic code. *Journal of Theoretical Biology* 32:293–298
5. Grantham R (1972) Codon base randomness and composition drift in coliphage. *Nature New Biology* 237:265 [Grantham’s bold generalizations back the point (see Prologue) that one does not need the full text of a dictionary to work out how a dictionary works.]
6. Grantham R, Perrin P, Mouchiroud D (1986) Patterns in codon usage of different kinds of species. *Oxford Surveys in Evolutionary Biology* 3:48–81
7. Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: mutational bias, translation selection, or both? *Biochemical Society Transactions* 21:835–841
8. Cox EC, Yanofsky C (1967) Altered base ratios in the DNA of an *Echerichia coli* mutator strain. *Proceedings of the National Academy of Sciences USA* 58:1895–1902
9. Muto A, Osawa S (1987) The guanine and cytosine content of genomic DNA and bacterial evolution. *Proceedings of the National Academy of Sciences USA* 84:166–169
10. Forsdyke DR (2004) Regions of relative GC% uniformity are recombinational isolators. *Journal of Biological Systems* 12:261–271
11. Lee J-C, Mortimer JR, Forsdyke DR (2004) Genomic conflict settled in favour of the species rather than of the gene at extreme GC% values. *Applied Bioinformatics* 3:219–228
12. Clare EL, Kerr KCR, Königslöw TE von, Wilson JJ, Hebert PDN (2008) Diagnosing mitochondrial DNA diversity: applications of a sentinel gene approach. *Journal of Molecular Evolution* 66:362–367 [A Consortium for the “Barcode of Life” has been established at the Smithsonian Museum of Natural History in Washington.]
13. Min XJ, Hickey DA (2007) DNA barcodes provide a quick preview of mitochondrial genome composition. *PLOS One* 3:e325
14. Stoeckle MY, Thaler DS (2014) DNA barcoding works in practice but not in (neutral) theory. *PLOS ONE* 9:e100755
15. Wada A, Suyama A, Hanai R (1991) Phenomenological theory of GC/AT pressure on DNA base composition. *Journal of Molecular Evolution* 32:374–378
16. Johnstone B (1987) Genes on the production line. *New Scientist* 113 (1550):39–40 [David Cyranoski wrote (2009 *Nature* 460:171–172): “In the 1970s Wada ran into many sceptical biologists when he was one of the first to envision large scale automated genomic sequencing. But even as these technologies were ramping up elsewhere, Japan’s bureaucrats stalled, its genomics fell behind, ... The unfolding of Wada’s failed efforts are described in a book aptly titled *A Defeat in the Genome Project*.”]
17. Wada A, Suyama A (1985) Third letters in codons counterbalance the (G+C) content of their first and second letters. *Federation of European Biochemical Societies Letters* 188:291–294
18. D’Onofrio G, Bernardi G (1992) A universal compositional correlation among codon positions. *Gene* 110:81–88
19. Orr HA (2004) A passion for evolution. *The New York Review of Books* 51(3):27–29
20. Forsdyke DR (2009) Scherrer and Josts’ symposium. The gene concept in 2008. *Theory in Bioscience* 128:157–161
21. Paz A, Kirzhner V, Nevo E, Korol A (2006) Coevolution of DNA-interacting proteins and genome ‘dialect.’ *Molecular Biology & Evolution* 23:56–64

22. Lao PJ, Forsdyke DR (2000) Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Research* 10:228–236
23. Mortimer JR, Forsdyke DR (2003) Comparison of responses by bacteriophage and bacteria to pressures on the base composition of open reading frames. *Applied Bioinformatics* 2:47–62
24. Rayment JH, Forsdyke DR (2005) Amino acids as placeholders. Base composition pressures on protein length in malaria parasites and prokaryotes. *Applied Bioinformatics* 4:117–130
25. Lin F-H, Forsdyke DR (2006) Prokaryotes that grow optimally in acid have purine-poor codons in long open reading frames. *Extremophiles* 11:9–18
26. Sabath N, Ferrada E, Barve A, Wagner A (2013) Growth temperature and genome size in bacteria are negatively correlated, suggesting genomic streamlining during thermal adaptation. *Genome Biology & Evolution* 5:966–977
27. Goncarencu A, Ma B-G, Berezovsky IN (2014) Molecular mechanisms of adaptation emerging from the physics and evolution of nucleic acids and proteins. *Nucleic Acids Research* 42:2879–2892
28. Dehouck Y, Folch B, Rooman M (2008) Revisiting the correlation between proteins' thermo-resistance and organisms' thermophilicity. *Protein Engineering, Design and Selection* 21:275–278
29. Liu L, Wang L, Zhang Z, Wang S, Chen H (2012) Effect of codon message on xylanase thermal activity. *Journal of Biological Chemistry* 287:27183–27188
30. Lambros R, Mortimer JR, Forsdyke DR (2003) Optimum growth temperature and the base composition of open reading frames in prokaryotes. *Extremophiles* 7:443–450
31. Jain K, Krause K, Grewe F, Nelson GF, Weber APM, Christensen AC, Mower JP (2014) Extreme features of the *Galdieria sulphuraria* organellar genomes: a consequence of polyextremophily? *Genome Biology & Evolution* 7:367–380

Chapter 13

Exons and Introns

Ideas in the theory of evolution can be used in situations far removed from biology. Similarly, information theory has ideas that are widely applicable to situations remote from its original inspiration.

Richard Hamming (1980) [1]

First a digression on nuclei. Organisms can be divided into those whose cells do not have a nucleus, the single celled “prokaryotes” (Greek: *pro* = before; *karyon* = nucleus), and those whose cells have a nucleus, the single- or multi-celled “eukaryotes” (Greek: *eu* = good or normal; *karyon* = nucleus). Prokaryotes include species of bacteria (eubacteria) and archaea (archaeobacteria) – a bacteria-like group sometimes found in extreme environments (e.g. hot springs). Eukaryotes (eukarya) include all species of animals and plants, both single-celled (protozoa, protophyta) and multi-celled (metazoa, metaphyta) – that’s you and me!

The spelling suggests that organisms without a nucleus, somewhat like modern bacteria or archaea, evolved *before* organisms with their DNA encapsulated within a nuclear membrane. However, as mentioned in Chapter 8 when wondering whether isochores appeared early or late, prokaryotes and eukaryotes are believed to have evolved from a common ancestor, whose properties we can now only infer. It may appear parsimonious to imagine an ancestor without a nucleus. But parsimony may be in the mind of the beholder. Unlikely, but maybe the ancestor had a nucleus [2]?

In the 1960s many eukaryotic RNAs were found to be first synthesized as ‘giant’ RNAs, a large part of which was then ‘wasted’ as the RNAs were ‘processed’ to generate mature cytoplasmic products. Considering a ‘gene’ as corresponding to a unit of transcription, it appeared that a large part of a gene might contain redundant information. First there were observations on eukaryotic ribosomal RNAs (rRNAs). These were found to be transcribed from their DNA genes as long precursor RNAs that were then processed by the removal of apparently functionless internal ‘spacer’ sequences [3]. However, prokaryotic (bacterial) rRNA genes were more compactly organized. Did rRNA genes in the first organisms to evolve have the spacer sequences, which decreased later in prokaryotes (‘spacers-early’), or were the long spacer sequences acquired later in eukaryotes (‘spacers-late’)?

A similar processing was later found in the case of eukaryotic messenger RNAs (mRNAs). Again, bacterial mRNA genes were more compactly organized. By 1970 it had been shown that many freshly minted giant nuclear RNAs contained sequences

corresponding to smaller mature cytoplasmic mRNAs. A simple model was that each giant RNA consisted of an mRNA sequence that was *flanked* by long redundant RNA segments [4]. It turned out, however, that instead of one DNA sequence segment compactly encoding one protein (as in most prokaryotic genes), eukaryotic genes often encode proteins in segments that are interrupted by segments that do not usually encode proteins. Thus, if a mature ‘sense’ mRNA be represented as:

Maryhadalittlelambitsfleecewaswhiteassnow (13.1)

Then the corresponding giant precursor RNA would contain segments of what would appear to be “nonsense:”

Maryhaxqvhxmgeqzdalittlelaqwxgtpscrbitsfleecewazgfxnyqswwhiteassnow (13.2)

Since an entire gene was transcribed, the internal RNA sequences (e.g. *xqvhxmgeqz*) derived from DNA “introns,” had to be removed from the initial transcript. What remained in the processed RNA – the mRNA – was derived from DNA “exons” (Fig. 13.1) [5]. Thus, a protein-encoding gene consists of exons and introns.

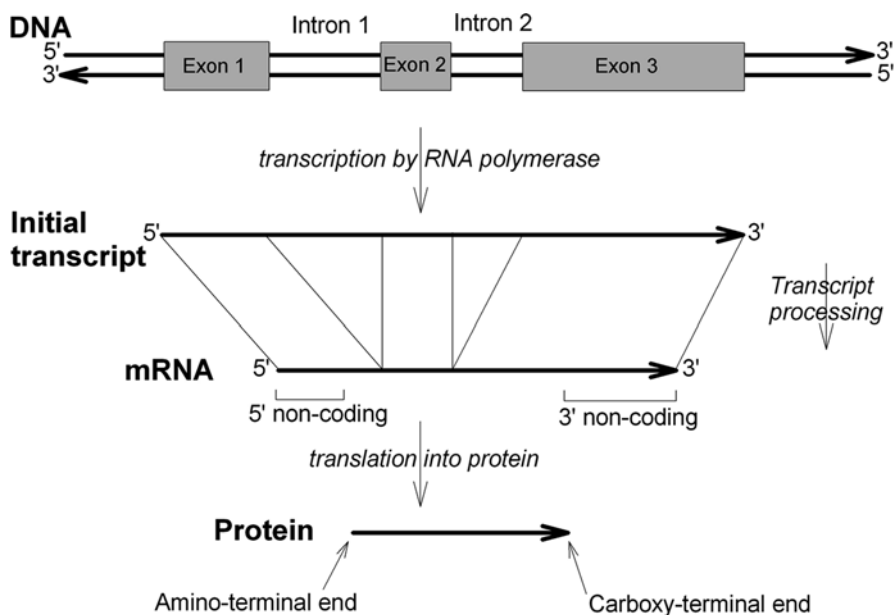


Fig. 13.1 Transcription and processing of mRNA from an eukaryotic gene. Sequences that correspond to the final mRNA are marked as boxes on the duplex DNA. The enzyme RNA polymerase initiates transcription at the left end of exon 1 using the bottom DNA strand as template, and terminates transcription at the right end of exon 3. Intron sequences are spliced out from the initial transcript, so that the mature mRNA consists of a 5' non-protein-coding region, a central protein-encoding region, and a 3' non-protein-coding region. The central region is translated into protein. A gene may be narrowly construed as corresponding to the segment of DNA that is transcribed. A broader definition would include other segments (e.g. the promoter region where RNA polymerase initially binds to DNA), which may lie outside the transcribed segment. Williams gives a different definition (Chapter 11)

Since the ‘spacer’ phenomenon had already been described for rRNA it should not have been a surprise that it also applied to mRNAs; but many, myself included, were surprised at its generality.

This chapter presents evidence that introns are a way of resolving intragenomic conflicts between different forms and levels of information. It turns out that cases of extreme intragenomic conflict, as when genes in ‘arms races’ evolve rapidly under positive selection, demonstrate quite clearly that “information theory has ideas that are widely applicable” [1].

Introns Interrupt Information

Although there were many attempts, it proved difficult consistently to associate exons with domains of protein structure or function [6, 7]. Introns interrupt genetic information *per se*, not just protein-encoding information. Thus, during transcript processing intron-encoded information is removed both from parts of mRNA precursors that encode protein, and from parts of mRNA precursors that do not encode protein (i.e. from 5′ and 3′ non-coding sequences). Certain special RNAs which, like rRNAs, are not translated to give a protein product (see Chapter 17), *also have introns*.

Exons have a narrow size range, with a peak at about 100 bases (Fig. 13.2) [8]. If coding, this would correspond to 33 amino acids. So proteins, often containing hundreds of amino acids, are usually encoded by long genes with many exons [9]. There appears to be a limit to the extent to which a genome will tolerate a region that both encodes protein, and is purine-loaded (Chapter 6). When that limit is reached, protein-encoding and purine-loading functions are arrested, and can restart only after a ‘decent’ interval – namely an intron interval. Introns are generally longer than the exons that surround them (Fig. 2.5). Each exon being separated from its neighbors by introns, the sum of intron lengths tends to increase proportionately as the sum of exon lengths increases [10].

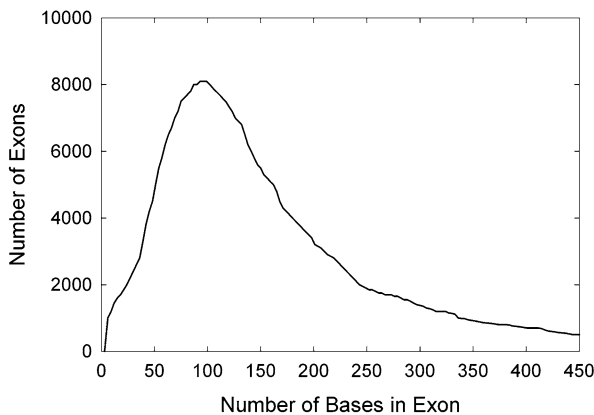


Fig. 13.2 Distribution of exon sizes in eukaryotic genes. The number of exons in different size categories increases to a maximum and then progressively declines. Most exons are about 100 bases in length. (Data are from the ExInt database.)

Introns First

What functions, if any, do introns have? Did introns appear early or late in the evolution of biological forms? If introns can be dispensed with in bacteria, then perhaps they have no function. Alternatively, whatever function introns have, either is not necessary in bacteria, or can be achieved in other ways by bacteria. Since members of many bacterial species appear to be under intense pressure to streamline their genomes to facilitate rapid replication, if it were possible they should have dispensed with any preexisting introns and/or should have been reluctant to acquire them. On the other hand, if introns play a role, and/or do not present too great a selective burden, then organisms under less pressure for genome compaction might have retained preexisting introns, or might have acquired them.

An early origin of introns is suggested by the fact that humans share the positions and sequences of many introns with a marine worm, indicating that their common ancestor had the same introns [11]. But introns did not appear so that at some remote future date organisms with exons would be advantaged. Nature is not so prescient. Although something playing a role at one point in time can later come to serve an unrelated role (see Chapter 18), in general, evolution does not work this way. Sometimes a random event (genetic drift) provides an evolutionary toe-hold, but for something as widespread and drastic as introns there would need to be some immediate selective advantage.

Knowing the function of introns seemed critical for sorting this out. There were many ingenious suggestions. Some thought introns were just another example of the apparently useless ‘junk’ DNA that appeared to clutter the DNA of many organisms (see Chapter 15). Others thought that introns might have facilitated the swapping of protein domains to generate new proteins [5, 12]; but this could not easily explain how introns *initially* arrived.

However, the notion of ‘message’ sequences interrupted by ‘non-message’ sequences is familiar to those working on noise affecting electronic signal transmission. Information scientist Richard Hamming pointed out that in such systems non-message sequences can have an error-checking function that permits the receiver to detect and correct errors in message sequences [1]. Could introns have a similar error-checking role [13]? In Australia, biochemist Darryl Reanney proposed that prior development of recombination-mediated error-correction capacities would have been critical for genomes later to evolve protein-encoding capacity. In other words, introns were neither “early” nor “late.” They were “*first*” [14, 15]!

The order of bases in nucleic acids has been under evolutionary pressure to develop the potential to form stem-loop structures, which might facilitate “in-series” or “in-parallel” error-correction by recombination (Chapter 2). This means that genomic sequences convey more than one level of information. Furthermore, as predicted in 1893 by the discoverer of DNA, Johann Miescher, a sexual process that brings molecules from separate sources together, could facilitate the mutual correction of errors (see Chapters 3 and 17). However, the need to participate in the process of error-detection and correction can result in redundancies and various constraints (Chapter 4).

The proposed error detection and correction process requires an alignment of two sequences, which itself depends on an initial ‘homology search’ (Chapter 10). For a successful homology search there must be sufficient similarity between sequences. One outcome of this is that segments of DNA link up (“recombine”) with other segments. In the course of this swapping of segments, errors can be detected and corrected (“gene conversion;” Fig. 11.2). The process is referred to as *recombination repair*, which distinguishes it from a variety of other repair processes that will not be considered here [16]. The adaptive value of recombination repair is likely to be very great. So, if it *could* have arisen early in the evolution of primeval biological forms (perhaps in an ‘RNA world’) prior to the evolution of protein-encoding capacity (i.e. if it were chemically feasible), then it *would* have arisen. Williams noted in 1966 [17]:

The existence of genetic recombination among the bacteria and viruses, and among all of the major groups of higher organisms, indicates that the molecular basis of sexuality is an ancient evolutionary development. Our understanding of the structure of the DNA molecule makes recombination at this level easy to visualize. In a sense sex is at least as ancient as DNA. ... I would agree, therefore, ... that sexual reproduction is as old as life, in that the most primitive living systems were capable of fusion and of combination and recombination of their autocatalytic particles. Modern organisms have evolved elaborate mechanisms for regulating this primitive power of recombination and for maximizing the benefits to be derived from it.

Given an early evolution of recombination, protein-encoding capacity would then have had to *intrude* into the genomes of biological forms that were *already* adapted for recombination. Would this intrusion have been readily accepted? Or would protein-encoding capacity have had to elbow its way forcibly into primeval genomes?

Protein Versus DNA

Although the degeneracy of the genetic code provides some flexibility as to which base occupies a particular position, there may still be a conflict between the needs of a sequence both to encode a protein and to respond to other pressures. In accordance with the “treasure your exceptions” principle (see Epilogue), situations where protein-encoding and/or other pressures are exceptional should be particularly informative.

Exceptional protein-encoding pressure is apparent in the case of genes under very strong positive Darwinian selection. In Chapter 7 competition among speakers for the attention of an audience provided a metaphor for positive selection. In that case, speakers were positively selected if they could communicate rapidly to a fixed audience by overcoming idiosyncrasies of accent. Under biological conditions, however, ‘the audience’ is not fixed. Positive selection often occurs under conditions where both ‘speakers’ and ‘audience’ are rapidly changing. This includes genes affected by ‘arms races’ between predators and their prey.

For example, snake venom may decrease the rodent population (prey) until a venom-resistant rodent line develops and expands (i.e. a mutant line arises with this selective advantage). Now, while the rodent population expands, the snake population

(predators) decreases because its members cannot obtain sufficient food (i.e. rodents). This decrease continues until a line of snakes with more active venom, which can overcome the resistance, develops and expands (i.e. a mutant line arises with this selective advantage). This population of snakes now expands, and the rodent population begins to fall again.

The cycle constitutes a biological arms race, and influences particular gene products. Parts of venom proteins which are important for toxicity are required to change so rapidly in response to this strong pressure from the environment (i.e. from rodents), that the corresponding genes can no longer afford the luxury of both encoding the best proteins (primary information) and attending to other pressures (secondary information; Fig. 13.3).

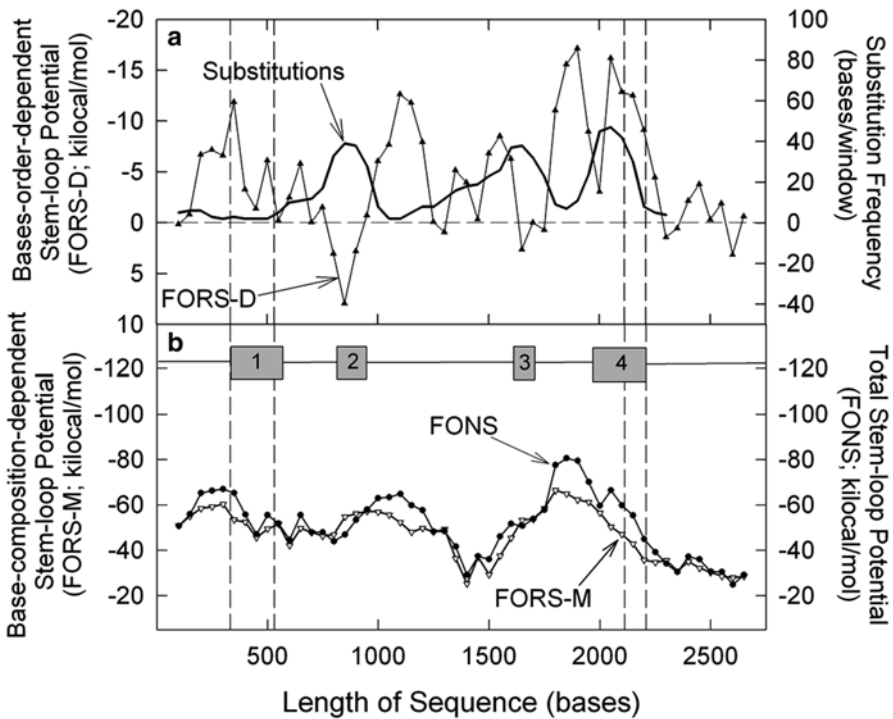


Fig. 13.3 High base substitution frequency and low base order-dependent stem-loop potential in exons of the rattlesnake gene encoding the basic subunit of venom phospholipase A_2 (PLA_2), which is under positive Darwinian natural selection. Base substitutions (continuous line in (a)) are compared with values for FORS-D (closed triangles in (a)), FORS-M (open triangles in (b)), and FONS (closed circles in (b)). Values were determined for overlapping 200 base windows moved in steps of 50 bases. Boxes in (b) indicate the location of the four exons, with dashed vertical lines showing, consecutively, the beginning of exon 1, the beginning of the protein-coding part of exon 1, the end of the protein-coding part of exon 4, and the end of exon 4. High stem-loop potential scores in negative kcals. Substitutions are base differences relative to the rattlesnake PLA_2 acidic subunit gene. The two genes are likely to have arisen by duplication from a common ancestral gene

The genes must encode better proteins even at the expense of their abilities to respond to other pressures. Accordingly, under extreme positive selection pressure the rate of sequence change in protein-encoding regions is high. When sequences of similar venom proteins (e.g. phospholipase A₂) from two snake species (or from duplicated genes within a species) are compared, *great differences* are found in the protein-encoding parts of exons (i.e. low sequence conservation). In contrast, *small differences* are found in introns, and in the 5' non-coding and 3' non-coding parts of exons (i.e. high sequence conservation). This is a dramatic reversal of the more usual situation where, in genes under classical negative ('purifying') Darwinian selection, exons are conserved much more than introns (i.e. introns display more variation; Fig. 2.5).

What is being conserved in snake venom introns? Analysis of fold potential as it affects base order (FORS-D; Chapter 5) reveals that base order-dependent stem-loop potential is *low* in exons (where sequence conservation is low) and *high* in introns (where sequence conservation is high; Fig. 13.3a). Base order-dependent stem-loop potential appears to have been conserved in introns; indeed, there is an inverse (reciprocal) relationship between base substitution frequency and base order-dependent stem-loop potential. When base order-dependent stem-loop potential is high (i.e. more negative kcal/mol), base substitution frequency is usually low (i.e. sequence conservation is high).

This suggests that the pressure to adapt the protein sequence (requiring non-synonymous codon changes) has been so powerful that base order has not been able to support base order-dependent stem-loop potential in the same exon sequence (Fig. 5.8). Instead, stem-loop potential is *diverted* to introns, which are appropriately conserved (fewer base substitutions than the surrounding exons). This is in keeping with the hypothesis that early in evolution protein-encoding potential was imposed on prototypic genomes that had *already* developed stem-loop potential. For this imposition to succeed without disturbing the general distribution of stem-loop potential, proteins had to be encoded in the fragments that we now call exons [18–21]. Thus, in the general case, introns were 'first.'

Another example of positive selection is the genome of the AIDS virus (HIV-1), which can be viewed as a predator with us as its prey (Chapter 10). Its 'teeth' are its exceptional ability to mutate itself to near oblivion, so disarming our adaptive immune system. Here, an inverse correlation between substitutions and base order-dependent stem-loop potential can be observed when substitutions and fold potential are displayed along the genome (Fig. 13.4). At first glance the data appear as a confused jumble of lines. But when the paired values from along the sequence are plotted against each other, a significant inverse correlation emerges (Fig. 13.5c) [22, 23].

Thus, sequences varying *rapidly* in response to powerful environmental selective forces ('arms races') appear unable to order their bases to favor the elaboration of higher order folded structures (of a type that, in eukaryotes, might mediate meiotic chromosomal interactions and recombination repair; Chapter 11). So, the encoding of nucleic acid stem-loop structure can be relegated to non-protein-encoding regions – introns, 5' and 3' non-coding regions, and non-genic DNA – (the favored options in less compact genomes). Alternatively the relegation can be

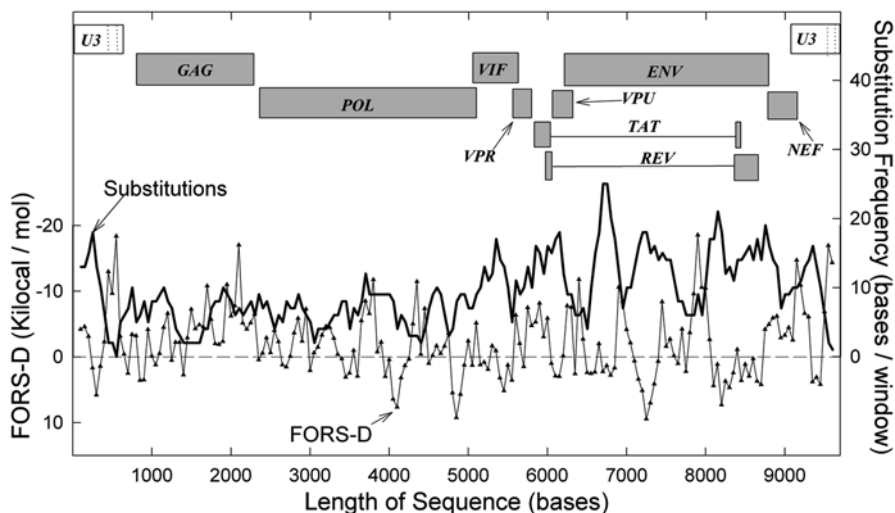
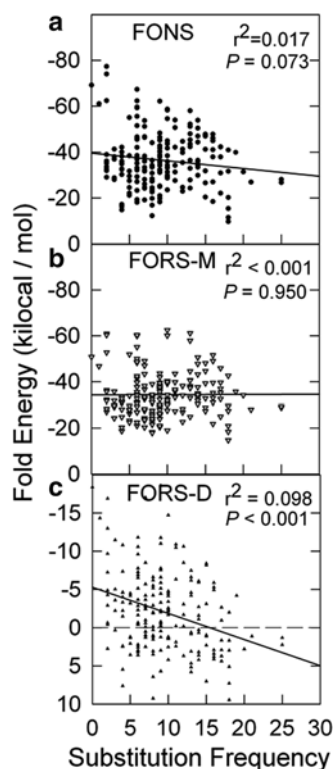


Fig. 13.4 High base substitution frequency and low base order-dependent stem-loop potential in regions of the AIDS virus genome that are under positive Darwinian natural selection. The various genes are shown as gray boxes with their abbreviated names. Two genes, *TAT* and *REV*, each have two exons that are shown linked by continuous lines that represent introns. Thus, one gene's intron can be another gene's exon. The distribution of base substitutions for virus "subtype" HIVSF2 relative to virus "subtype" HIVHXB2 is shown as a continuous line. Base order-dependent stem-loop potential (FORS-D) is shown as small filled triangles, with high potential corresponding to high negative values (kcal/mol). All values are for 200 base overlapping windows moved in steps of 50 bases

Fig. 13.5 In the AIDS virus genome there is (c) an inverse relationship between base order-dependent potential (FORS-D) and substitution frequency, but (b) no detectable relationship between base composition-dependent stem-loop potential (FORS-M) and substitution frequency. Shown in (a) are FONS values (the sum of FORS-D and FORS-M values). Although only approximately 10% of the variation in substitutions can be explained by base order-dependent stem-loop potential (shown by the r^2 value in (c)), the downward slope is significantly different from zero (shown by the low P value in (c))



to less rapidly evolving protein-encoding sequences (or parts of such sequences) where there is some flexibility in codon or amino acid choice (the main option for compact genomes).

Functionally important regions are conserved in genes evolving slowly under classical negative selection (e.g. there is a low local frequency of accepted base substitutions in protein-encoding regions that are under this selection pressure). If the rate of evolution has been slow, then there may have been more opportunities to arrive at an appropriate compromise with base order-dependent stem-loop potential. Thus, a relationship between base order-dependent stem-loop potential and sequence variability may be less evident.

On the other hand, the demands of faithful reproduction of a protein, with negative selection of individuals bearing mutations affecting its functionally most important parts, can leave the co-encoding of stem-loops not only to non-protein-encoding regions (e.g. introns), but also to regions encoding functionally less important, *and hence more variable*, parts of proteins (such as the protein surface, which cell water can readily access) [24, 25]. In this case, high base order-dependent stem-loop potential can correlate directly (not inversely) with high substitution rates (variability) when similar (homologous) sequences from different species are compared. Indeed, comparison of certain human and mouse oncogenes (*FOSB*) has revealed a positive correlation between base order-dependent fold potential and substitution frequency [26].

Achilles Heels

The AIDS virus has caused millions of deaths annually, and subtypes resistant to therapeutic agents have become more prevalent. Given the profiles shown in Figure 13.4, can you suggest potential targets for therapeutic attack – targets that are *least* likely to differ between different pathogens? Note that the region to the left of the *GAG* gene in the AIDS virus has the lowest base substitution rate (i.e. it is the most *conserved* part of the genome). This coincides with a major peak in base order-dependent stem-loop potential (indicated by high negative FORS-D values). This suggests that there is here a special need to conserve stem-loop potential [22]. Indeed, this is near the “dimer linkage” stem-loop sequence, which is necessary for the co-packaging of two AIDS virus genomes as part of a process that resembles meiotic pairing (Chapter 10). This may yet prove to be the Achilles heel of the AIDS virus, once the problem of ‘flushing out’ latent forms from host genomic reservoirs is solved [27, 28]. A highly conserved dimer linkage sequence could also prove to be the Achilles heel of the hepatitis C virus, a non-retroviral RNA virus that persistently infects 3% of the world population [29].

We humans may also have Achilles heels. Our genomes are rich in palindromes conferring stem-loop potential, which is usually to our advantage (see Chapters 10 and 17). But sometimes the palindromes involve oligonucleotide repeats, which can be to our disadvantage (see Chapter 14). This is particularly so when the repeats are **AT**-rich and can readily extrude from duplex DNA in cruciform stem-loop configurations.

At times when homologous chromosomes may not be precisely aligned, palindromic regions containing **AT**-rich repeats are prone to recombine by ‘kissing’ interactions with other, *non-allelic*, palindromic regions containing **AT**-rich repeats. Thus, there may be ‘cut-and-paste’ translocations (transpositions) between *non-homologous* chromosomes (“illegitimate recombination”). These can be detected in normal sperm samples [30, 31]. Whether they are pathological depends partly on the frequency of the translocations. This depends on the lengths of the **AT**-rich repeats, which varies (polymorphism). It is likely that, by virtue of these Achilles heels, many individuals with long palindromic regions containing **AT**-rich repeats have been eliminated by natural selection.

Mirror Repeats

In Chapter 4 we encountered “inverted repeats” with palindromic properties *at the duplex level* (see also Fig. 2.4). For example a ‘top’ single-strand in a duplex might read:

$$5' \text{ AAAAACCCGGGTTTT } 3' \quad (13.3)$$

Here **AAAAACCC** in the 5′ half of the ‘top’ strand is repeated on the *complementary* strand (not shown here), where it pairs with the **GGGTTTTT** sequence in the 3′ half of the top strand. Such sequences seem to serve a DNA level function, since they facilitate the extrusion of stem-loop secondary structures (Fig. 5.2). Single-strands also contain “direct repeats” that might, for example, read:

$$5' \text{ AAAAACCCAAAAACCC } 3' \quad (13.4)$$

Here **AAAAACCC** is repeated. Single-strands can also contain “mirror repeats” that might read:

$$5' \text{ AAAAACCCCCCAAAA } 3' \quad (13.5)$$

Here **AAAAACCC** can be considered the ‘mirror’ of **CCCCAAAA**. In these two cases, (10.4) and (10.5), the repeats occur in the *same* strand. Direct repeats and, especially, mirror repeats, have the potential to *oppose* local stem-loop formation, and so to oppose any DNA level function that stem-loops might serve.

The possibility of some consistent relationship between distinctive protein structures and the corresponding nucleic acid sequences has long been surmised [32]. Remarkably, mirror repeats are found at particular locations in exons. Their locations correlate with the boundaries of various structural elements in proteins; indeed, mirror repeats can *predict* where such structural elements will occur [33]. In this case it appears that a conflict between protein and DNA has been won by protein. By preventing stem-loop formation, mirror repeats should prevent local recombination and thus preserve the local integrity of the DNA encoding a structural element in a protein.

We know that a protein, by ‘insisting’ (through natural selection) on having a particular amino acid at a particular position in its sequence, requires that the corresponding gene have a suitable codon at a particular position in its sequence. Now we see that a protein, by ‘insisting’ (through natural selection) on having a particular structural element (e.g. alpha-helix, beta-strand), also requires that the corresponding gene have appropriately positioned (albeit often imperfect) mirror repeats. These are both examples of “protein-pressure,” but usage of the latter term generally relates to the determination that a particular amino acid must occupy a particular position.

RNA versus DNA

Usually a particular DNA sequence is transcribed into an identical RNA sequence, except that RNA molecules have **U** (uracil) rather than **T** (thymine); but these are chemically similar bases. So it is not surprising that, in broad features, computer-derived secondary structures for an RNA molecule (using dinucleotide pairing energy tables for RNA bases), are similar to the structures derived for the corresponding DNA (using dinucleotide pairing energy tables for DNA bases; Table 5.1). Yet there are genes with no protein product. The gene products are RNAs with specific functions that, like proteins, are often dependent on the secondary (and higher order) structures they adopt (usually selected for at the *cytoplasmic* level). If such structures also sufficed for structure-dependent functions at the DNA level, then there might be no need for introns in genes for non-protein-encoding RNAs.

The fact that there are spacers or introns in such genes, implies that sequences generating the stem-loop secondary structures that suffice for function at the RNA level, do *not* suffice for, and may even *conflict* with, the sequences needed for stem-loop secondary structures that function at the DNA level in the nucleus. Since patterns of RNA stem-loops are influenced by the purine-loading of loops (the selective force for which probably operates at the cytoplasmic level; Chapter 6), then purine-loading pressure (which would constrain stem-loop patterns in exons) should support stem-loop pressure in provoking the splitting of what might otherwise have been large exons.

Overlapping Genes

Strict adherence to the **RNY** rule (Chapter 7) would dictate that, of the three possible triplet reading frames in the mRNA synonymous ‘coding’ strand of DNA, the one that best fits the **RNY** pattern would be the actual reading-frame. However, sometimes it is expedient for genes to overlap, either entirely or in part, and in this case one of the genes, if transcribed in the same direction, can use another, non-**RNY**, reading-frame. This applies to some of the genes of the AIDS virus, which are co-orientated – all transcribed to the right (Fig. 13.4).

In some circumstances, genes transcribed in different directions may overlap. Thus, the ‘top’ strand may be the coding strand of one gene, and the ‘bottom’ strand may be the coding strand of another gene. Again, one reading frame is **RNY** and two are non-**RNY** (Fig. 7.2), and any of the three may be employed.

Whatever the transcription direction, in overlapping genes the region of the overlap can come under extreme protein-encoding pressure, and this might conflict with other pressures. Indeed, consistent with the case made here, base order-dependent folding potential decreases where genes overlap [34].

Simple Sequences in Malaria

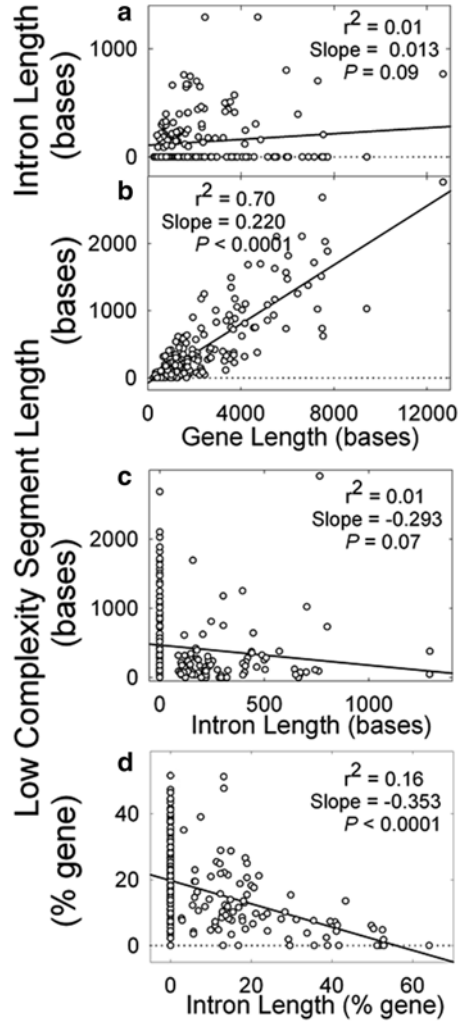
As in the above examples of genes under *exceptional* protein-encoding pressure, genomes under another exceptional pressure, **GC**-pressure, should also be highly informative. The genome of the most lethal malaria species, *Plasmodium falciparum*, satisfies this requirement, being under strong ‘downward **GC**-pressure.’ It is one of the most **AT**-rich species known (i.e. very low (**G+C**)%; Fig. 12.7a). Another unusual feature of *P. falciparum* is that many proteins are *longer* than their equivalent proteins (homologs) in species that have less extreme genomic (**G+C**)% deviations. This reflects the acquisition by the *P. falciparum* proteins of low complexity ‘simple sequence’ segments that have no known function. Simple sequence at the protein level (i.e. runs of amino acids from a limited range of the twenty possible amino acids) is encoded by simple sequence at the nucleic acid level (i.e. runs of bases from a limited range of the four possible bases; see Chapter 14).

There are other unusual features of the *P. falciparum* genome. Unlike many eukaryotic genomes [10], there is poor correlation between the length of a gene and the combined lengths of the introns of that gene (Fig. 13.6a). Yet there is a close correlation between the length of a gene and the combined lengths of low complexity segments in that gene (Fig. 13.6b). In this respect the low complexity elements appear like introns; however, unlike introns, they are not removed during processing of the RNA transcript.

Introns and low complexity segments are not interchangeable in that, as absolute intron length increases, there is little decline in length of low complexity segments in a gene (Fig. 13.6c). But when the lengths are expressed as a proportion of gene length, a reciprocal relationship emerges (Fig. 13.6d). Whereas in many eukaryotic genomes intron locations show no relationship to protein functional domains (since if splicing is accurate intron location is irrelevant to the protein), low complexity segments in *P. falciparum* do, of necessity, predominate between functional domains.

If low complexity segments have no, or minimal, function at the protein level, do they reflect a function at the nucleic acid level? Low complexity segments in *P. falciparum* are usually of high (**A+G**)%, and so they contribute to purine-loading. Introns, in contrast, tend to be of low (**A+G**)% (i.e. tending towards pyrimidine-loading).

Fig. 13.6 In the genome of the malaria parasite, *P. falciparum*, low complexity segment length, not intron length, correlates best with gene length. Shown for chromosome 2 are relationships of lengths of introns (a), and low complexity segments (b), to the lengths of the corresponding genes, and relationships of either absolute (c), or percentage (d), lengths of low complexity segments to the lengths of introns in the same genes. *P* values indicate the probability that the slope values are not significantly different from zero (i.e. a low *P* value indicates high significance)



It will be shown in Chapter 14 that, when contributing to **AG**-pressure (purine-loading pressure), low complexity segments can countermand fold potential. But when **AG**-pressure is not extreme, low complexity segments contribute to fold potential. In this respect, they do resemble introns.

Thus it would appear likely that spacers and introns arose early in evolution because they are preferential sites for the encoding of the stem-loop structures in DNA that are necessary for initiating recombination and, hence, for error-detection and correction. While, in extreme cases, by virtue of this function, introns are conserved more than exons (Fig. 13.3), in the general case, introns evolve more rapidly (are conserved less) than exons (see Fig. 2.5). By diversifying in this way,

introns change their patterns of stem-loop extrusion (Fig. 10.4) and so facilitate the anti-recombination necessary for the survival of genic or genome duplicates (i.e. speciation).

Multiple Pressures

The multiple, potentially conflicting, pressures affecting both the genome phenotype and the conventional (classical Darwinian) phenotype are summarized in Figure 13.7 and Table 13.1.

The proteins of *P. falciparum* provide an extreme example of nucleic acid level pressures that affect protein sequence. However, proteins of all organisms are, to a degree, manifestations of nucleic acid level pressures. The correlation between gene length and content of low complexity segment (Fig. 13.6b) is probably general, with low complexity segments seeming to serve pressures for purine-loading, rather than for protein function. In a species where purine-loading was a dominant nucleic acid level pressure, proteins should be longer (and hence generally larger) than their homologous proteins in other species where purine-loading was less dominant. Since there is a reciprocal relationship between GC% and purine-loading (Fig. 12.9), then proteins should tend to be smaller in GC-rich species (Fig. 12.11) unless such GC-richness itself required addition of GC-rich low complexity segments. Since GC-rich species tend to have small proteins, it is likely that purine-loading is generally the dominant nucleic acid pressure in this respect (see Chapter 14).

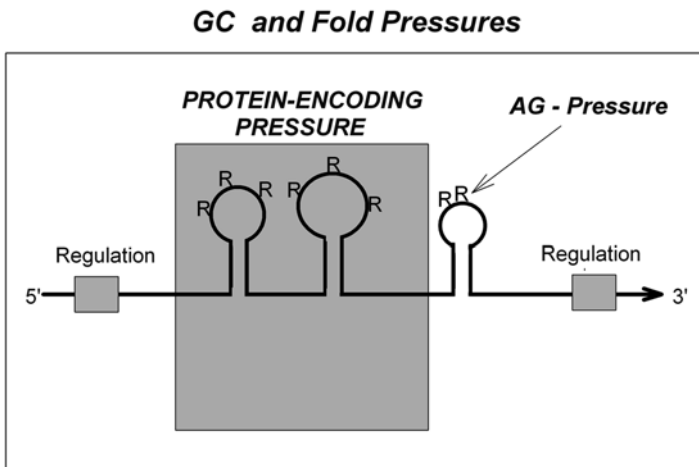


Fig. 13.7 Summary of potentially conflicting pressures operating at the mRNA level. The genome-wide pressures, GC-pressure and fold (stem-loop) pressure, influence the entire mRNA (shown as a thick horizontal arrow with loops that is enclosed in the large outer box). Purine-loading pressure (AG-pressure) is a local pressure that also influences the entire mRNA. Small gray boxes indicate potential sites for the binding of regulatory factors that sometimes preferentially locate to the 5' and 3' non-coding regions. The large gray box indicates the central, protein-encoding region (ORF), where protein-pressure is deemed to operate

Table 13.1 Summary of multiple, potentially conflicting, levels of information in genomes, and their relationships to Chargaff’s rules

Environmental selective factors	Selection for mutations which:	Primary effect on DNA function	Biological result	Observed features of modern DNA
Classical phenotypic selective factors	Change encoded proteins	Protein-encoding can constrain other functions	Change in classical phenotype	Changed base-pair in accordance with Chargaff's first parity rule
Competitors with more efficient translation, and intracellular pathogens	Purine-load RNAs	Purine-loading can constrain other functions	Efficient translation, and no "self" double strand RNA formation	Chargaff's cluster rule and Szybalski's transcription direction rule
Mutagens	Promote DNA stem-loop potential	Recombination promoted	Error detection and repair	Chargaff's second parity rule and genome-wide stem-loop potential
Classical phenotypic selective factors	Impair homology search between recently duplicated genes	Recombination between similar genes impaired	Gene duplication	Chargaff's (G+C)% rule
Differences in "reprotype" (Recombinationally "not-self" sexual partners)	Impair homology search between DNAs of species members whose sequences are diverging	Meiotic recombination impaired	Species duplication (speciation)	

Similarity Scores

Just as two nucleic acid sequences can be compared – base-by-base – by aligning them together, so two protein sequences can be compared – amino acid-by-amino acid – by aligning them together. But nucleic acid pressures can affect the success of protein alignment analyses based on the scoring schemes described in standard bioinformatics texts. For example, since the amino acid tryptophan is highly conserved in proteins but glutamic acid is not, a tryptophan match between two protein sequences scores more than a glutamic acid match between two protein sequences [35]. Thus, two protein sequences may have a low similarity score because tryptophans do not match. Where there is tryptophan in one sequence, an alignment program may score a mismatch, or place a gap, in the other sequence. However, tryptophan may be present at a certain position in a particular protein sequence because **TGG** (the codon for tryptophan) plays some role at the nucleic acid level that is not needed in the homologous gene from another species. At the protein level tryptophan is a mere ‘placeholder.’

Thus, gaps in the alignment of homologous protein sequences may occur if nucleic level pressures on the corresponding DNA sequences differ. To avoid this “problem,” protein alignment algorithms may depend on data from closely related

organisms (e.g. “PAM matrices”), or use short “blocks” of amino acids (e.g. “BLOSUM matrices”) corresponding to regions that are highly conserved at the protein level (e.g. active sites of enzymes). Trade-offs between competing pressures will be further explored in Chapter 14.

Summary

If genome space is finite with little, if any, DNA that is not functional *under some circumstance*, then potential conflicts between different forms of genomic information must be resolved by appropriate trade-offs. These trade-offs sometimes require that genes accommodate spacers, introns, and simple sequence elements. The nature and extent of the trade-offs varies with the biological species. Study of trade-offs is facilitated in genes or species where demands are exceptional (e.g. genes under positive selection pressure to adapt proteins, genes that overlap, and species under extreme downward or upward GC-pressures). Spacers and introns are likely to have existed early in evolution because they are preferential sites for the stem-loop structures that are necessary for initiating recombination and, hence, error-detection and correction. Genes, as recognized today, would have arisen in sequences *already* adapted for these purposes. Purine-loading pressure would have supported protein-pressure in provoking the splitting into introns of what might otherwise have been large exons. From this perspective we can understand why the genes of the malaria parasite are extraordinarily long, and we can identify the potential Achilles heel of the AIDS virus as the dimer-linkage sequence that is essential for the copackaging of disparate genomes, so allowing recombination repair in a future host.

References

1. Hamming RW (1980) *Coding and Information Theory*. Prentice-Hall, Englewood Cliffs
2. Booth A, Doolittle WF (2015) Eukaryogenesis, how special really? *Proceedings of the National Academy of Sciences USA* 112:10278–10285
3. Federoff NV (1979) On spacers. *Cell* 16:687–710
4. Scherrer K (2003) The discovery of ‘giant’ RNA and RNA processing. *Trends in Biochemical Sciences* 28:566–571
5. Gilbert W (1978) Why genes in pieces? *Nature* 271:501
6. Weber K, Kabsch W (1994) Intron positions in actin genes seem unrelated to the secondary structure of the protein. *EMBO Journal* 13:1280–1288
7. Stoltzfus A, Spencer DF, Zuker M, Logsdon JM, Doolittle WF (1994) Testing the exon theory of genes: evidence from protein structure. *Science* 265:202–207
8. Sakharkar M, Passetti F, Souza JE de, Long M, Souza SJ de (2002) ExInt: an Exon Intron Database. *Nucleic Acids Research* 30:191–194
9. Blake C (1983) Exons – present from the beginning. *Nature* 306:535–537
10. Naora H, Deacon NJ (1982) Relationship between the total size of exons and introns in protein-coding genes of higher eukaryotes. *Proceedings of the National Academy of Sciences USA* 79:6196–6200

11. Raible F, et al. (2005) Vertebrate-type intron-rich genes in the marine annelid *Platynereis dumerilii*. *Science* 310:1325–1326
12. Liu M, Grigoriev A (2004) Protein domains correlate strongly with exons in multiple eukaryotic genomes – evidence of exon shuffling? *Trends in Genetics* 20:399–403
13. Forsdyke DR (1981) Are introns in-series error-detecting codes? *Journal of Theoretical Biology* 93:861–866
14. Reanney DC (1987) Genetic error and genome design. *Cold Spring Harbor Symposium on Quantitative Biology* 52:751–757
15. Forsdyke DR (2013) Introns first. *Biological Theory* 7:196–203
16. Bernstein C, Bernstein H (1991) *Aging, Sex and DNA Repair*. Academic Press, San Diego
17. Williams GC (1966) *Adaptation and Natural Selection. A Critique of Some Current Evolutionary Thought*. Princeton University Press, Princeton, pp 133–138 [Because RNA can play both templating and structural roles, it is an attractive candidate for being ‘the primordial life molecule.’ Hence we speak of an early “RNA world.” Information storage and templating (its digital roles) would have later devolved to DNA (with mRNA acting as a short-lived intermediate), and many structure-dependent functions (its analog roles) would have devolved to proteins. Nevertheless, in various forms RNA retains its flexibility today.]
18. Forsdyke DR (1995). A stem-loop ‘kissing’ model for the initiation of recombination and the origin of introns. *Molecular Biology & Evolution* 12:949–958
19. Forsdyke DR (1995) Conservation of stem-loop potential in introns of snake venom phospholipase A₂ genes. An application of FORS-D analysis. *Molecular Biology & Evolution* 12:1157–1165
20. Leicht BG, Muse SV, Hanczyc M, Clark AG (1995) Constraints on intron evolution in the gene encoding the myosin alkali light chain in *Drosophila*. *Genetics* 139:299–308
21. Forsdyke DR. (1996) Stem-loop potential: a new way of evaluating positive Darwinian selection? *Immunogenetics* 43:182–189
22. Forsdyke DR (1995) Reciprocal relationship between stem-loop potential and substitution density in retroviral quasispecies under positive Darwinian selection. *Journal of Molecular Evolution* 41:1022–1037
23. Zhang C-Y, Wei J-F, He S-H (2005) The key role for local base order in the generation of multiple forms of China HIV-1 B/C intersubtype recombinants. *BMC Evolutionary Biology* 5:53
24. Alvarez-Valin F, Tort JF, Bernardi G (2000) Nonrandom spatial distribution of synonymous substitutions in the GP63 gene from *Leishmania*. *Genetics* 155:1683–1692
25. Bustamente CD, Townsend JP, Hartl DL (2000) Solvent accessibility and purifying selection within proteins of *Escherichia coli* and *Salmonella enterica*. *Molecular Biology & Evolution* 17:301–308
26. Heximer SP, Cristillo AD, Russell L, Forsdyke DR (1996) Sequence analysis and expression in cultured lymphocytes of the human *FOSB* gene (*G0S3*). *DNA Cell Biology* 12:1025–1038
27. Forsdyke DR (1991) Programmed activation of T-lymphocytes. A theoretical basis for short term treatment of AIDS with azidothymidine. *Medical Hypothesis* 34:24–27 [The HIV hypermutation strategy to ‘outwit’ our adaptive immune defenses makes a mockery of attempts to vaccinate.]
28. Williams SA, et al. (2005) Prostratin antagonizes HIV latency by activating NF-kappaB. *Journal of Biological Chemistry* 279:42008–42017 [HIV may have an Achilles heel, but first latent HIV must be ‘flushed’ from the genome using “inductive therapy.”]
29. Shetty S, Kim S, Shimakami T, Lemon SM, Mihailescu M-R (2010) Hepatitis C virus genomic RNA dimerization is mediated via a kissing complex intermediate. *RNA* 16:913–925
30. Kurahashi H, Inagaki H, Yamada K, Ohye T, Taniguchi M, Emanuel BS, Toda T (2004) Cruciform DNA structure underlies the etiology for palindrome-mediated human chromosomal translocations. *Journal of Biological Chemistry* 279:35377–35383
31. Lu S, Wang G, Bacolla A, Zhao J, Spitser S, Vasquez KM (2015) Short inverted repeats are hotspots for genetic instability: relevance to cancer genomes. *Cell Reports* 10:1674–1680
32. Thanaraj TA, Argos P (1996) Protein secondary structural types are differentially coded in messenger RNA. *Protein Science* 5:1973–1983

33. Lang DM (2005) Imperfect DNA mirror repeats in *E. coli TnsA* and other protein-coding DNA. *Biosystems* 81:183–207
34. Barrette IH, McKenna S, Taylor DR, Forsdyke DR (2001) Introns resolve the conflict between base order-dependent stem-loop potential and the encoding of RNA or protein. Further evidence from overlapping genes. *Gene* 270:181–189
35. Grantham R (1974) Amino acid difference formula to help explain protein evolution. *Science* 185:862–864 [To construct a “PAM matrix,” the observed frequency of interchanges between two amino acids is divided by the expected interchanges calculated by multiplying the respective frequencies of each amino acid in the data set. There being 20 amino acids, a 20×20 matrix is generated. Of the 400 values, 20 are on the diagonal and the remaining 380 are duplicates, so that 190 values form the final matrix. Two proteins whose amino acid differences generate a low total PAM score would be held to be closely related evolutionarily.]

Chapter 14

Complexity

All perception and all response, all behaviour and all classes of behaviour, all learning and all genetics, all neurophysiology and all endocrinology, all organization and all evolution – one entire subject matter – must be regarded as communicational in nature, and therefore subject to the great generalizations or ‘laws’ which apply to communicational phenomena. We therefore are warned to expect to find in our data those principles of order which fundamental communication theory would propose.

Gregory Bateson (1964) [1]

Some things are more complex than others. We feel intuitively that highly complex things are associated with more information than less complex things. But this will not suffice for *Homo bioinformaticus*. Information and complexity must be measured. When this is done, amazing “principles of order” emerge from large amounts of seemingly disparate data.

Scoring Information Potential

Imagine a DNA sequence that is a mix of just two bases, **A** and **T**. Given this sequence, and an appropriate code, you can ‘read’ the information it contains. However, if the sequence consists of just one base, either **A** or **T**, this is not possible. A string of **A** bases, in the sense we have been discussing in this book, conveys little information. Sprinkle in a few **T**s and the potential of the sequence to convey information increases. Sprinkle in a few more and information potential increases further, reaching a maximum when **A**s and **T**s are at equal frequencies. Since the bases are alternatives to each other, we can express their frequencies in terms of just one of the bases. So, when **A**% is zero (i.e. **T**% is 100), information potential is zero. When **A**% is 50 (i.e. **T**% is 50), information potential is maximum. When **A**% is 100 (i.e. **T**% is zero), information potential is zero again.

Biological DNA sequences tend to display intrastrand parity between complementary bases, such as **A** and **T**, so information potential is maximized (Chapter 4). Thus, violations of PR2 (i.e. violations of base equifrequency) tend to decrease the information potential of DNA sequences. “Violation” is a strong word, implying that

something unwanted has happened. But whatever affects base parity in a DNA strand, might *itself* impart information of possible adaptive value to that strand. This might include information relating to secondary structure (Chapter 5), or to purine-loading (Chapter 6), or to DNA ‘accent’ (Chapter 10). Thus, it is the *remaining* information potential (after these are accounted for) that is decreased by a lack of parity.

This lack of parity can be measured simply in percentage terms (e.g. departures from 50% A), or can be related to conventional measures of information content – the binary digits or “bits” of the information scientist (Chapter 2) [2]. Each unit scores as 1 bit in the case of a two unit sequence (e.g. a sequence containing an even mix of only two bases), 2 bits in the case of a four unit sequence (e.g. a sequence containing an even mix of four bases), and 4.322 bits in the case of a 20 unit sequence (e.g. a sequence containing an even mix of twenty amino acids; see Appendix B). Thus, in a two unit sequence, a segment containing only one of the bases would score as zero for each base, whereas a segment containing an even mix of both bases would score two for each base. In a two unit sequence, segments can be scored as to their information content on a scale from zero bits/base to two bits/base. Here we will deal with measurement of information potential in percentage terms.

In various contexts, the quality measured can be referred to either as “uncertainty,” or “entropy,” or “complexity.” This also relates to “compressibility,” since a simple sequence is easier to characterize in terms that allow a decrease in total symbol number (e.g. “AAAAAAAAAA” can be written in compressed form as “10A”). On this basis, one linear DNA molecule can be described as less complex (more compressible) than another. Furthermore, a given linear DNA molecule can be considered to consist of segments of varying complexity. Low complexity segments often have large deviations from the even parity of PR2 – the frequencies of the Watson-Crick pairing bases can approach 0% and 100%. High complexity segments often have small deviations from even parity – the frequencies of the Watson-Crick pairing bases are closer to 50%. However, a low complexity segment can have even parity. For example, **ATATATATAT** is of low complexity, but has even parity with respect to **A** and **T**; and **CGATCGATCGATCGAT** is of low complexity, but has even parity with respect to all four bases.

The New Bioinformatics

The ‘old’ bioinformatics is largely concerned with aligning different sequences to determine the degree of similarity. If a newly discovered gene in biological species *A* is found to display a high degree of sequence similarity with a well-characterized gene in species *B*, then this may provide a clue to the function of the product of the new gene in species *A*. If many genes in species *A* show similar degrees of similarity to genes in species *B*, then it is possible that the genes are homologous (orthologous) and the species have diverged relatively recently (in evolutionary time) from a common ancestral species. Thus, the ‘old’ bioinformatics is (i) a gene-product-orientated bioinformatics serving those who seek to identify and manipulate gene-products,

and (ii) a phylogenetic bioinformatics serving those who seek to classify species on the basis of evolutionary relationships.

Confusing these tasks are low complexity segments, sometimes referred to as “simple sequences.” Two genes that are not related evolutionarily may have *similar* simple sequences. These similarities are often sufficient to be picked up in alignment analyses. So simple sequences are distracters – an embarrassment that can generate spurious high-scoring alignments between genes that are not actually related. The parts of genes responsible for specific gene function usually correspond to specific sequences of high complexity. Accordingly, programs with names such as “Repeat masker” and “Seg” are widely used to identify low complexity segments and mask them prior to conducting alignment analyses (using other programs with names such as “Blast,” “Clustal,” and “Genalign”).

On the other hand there is EB which is primarily concerned with the understanding of biological evolution. This ‘new’ bioinformatics views simple sequences as having the potential to provide profound insights. The goals of the ‘old’ and the ‘new’ bioinformatics differ, but programs such as Seg can greatly aid the evolutionary bioinformaticist, as will be shown here.

Protein or Nucleic Acid Level Function?

Low complexity at the nucleic acid level correlates with low complexity at the protein level. For example, since the codon **TTT** in DNA encodes the amino acid phenylalanine (Table 7.1), then a sequence containing only **T** bases (poly(d**T**)) would encode a protein containing only the amino acid phenylalanine (i.e. polyphenylalanine). On coming across a polyphenylalanine segment within a protein, we are faced with another chicken-and-egg problem. Would the polyphenylalanine have conferred some advantage to the function of the protein, with the automatic consequence that a **T**-rich nucleic acid sequence was selected in the course of evolution? Alternatively, did poly(d**T**) confer some advantage at the nucleic acid level, with the automatic consequence that the protein sequence became phenylalanine-rich? Thus, a primary selective influence operating at one level could be the cause of low complexity at both levels.

Honghui Wan and John Wootton in the USA [3] noted for a wide range of organisms that **GC**-rich genomes and **GC**-poor (**AT**-rich) genomes have more low complexity segments in their proteins than genomes in which the quantity of **G+C** is approximately equal to the quantity of **A+T** (Fig. 14.1).

While the cause of **GC**-pressure has been controversial, there is agreement that it operates genome-wide both on genes and on intergenic DNA. On the other hand, the pressure to encode a protein operates only on exons. Again, we are faced with a chicken-and-egg problem. If protein-pressure were driving (**G+C**)%, then only exons should be affected. We might postulate that (**G+C**)% is somehow infectious and, once exons have arrived at a particular value, this spreads into introns and non-genic DNA. But it is more likely that **GC**-pressure (with certain matching k-mer frequencies) is itself primary (Chapter 10). At the extremes – low or high – (**G+C**)% biases the amino

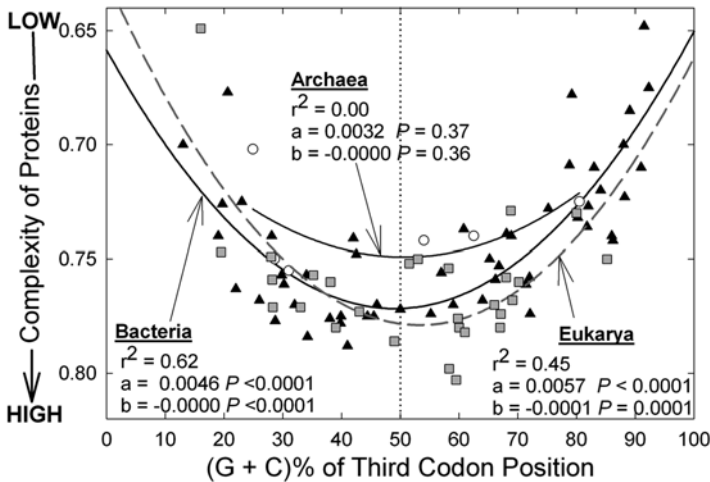


Fig. 14.1 Variation of protein complexity with species (G+C)% for three major groups of organisms: bacteria (triangles); archaea (circles); eukarya (squares). Each point corresponds to a particular species within a group. Values for protein complexity decrease from high to low on the Y-axis (note the *descending* scale). More high complexity proteins are found in species with 50% (G+C) and more low complexity proteins are found in species with extreme (G+C) percentages (low or high). This is exaggerated here by showing the base composition of the codon position that is least constrained by protein-pressure (third position). The points were fitted to curves (quadratic parabolas of general form $Y = Y_0 + aX + bX^2$), and parameters of the curves are shown in each figure with corresponding probabilities (P). The fit for archaea is not significant, but the curve seems to match those for bacteria and eukarya (eukaryotes). The data are from Wan and Wootton [3], who measured the “median global complexities” of protein sequences from a variety of extensively sequenced genomes

acid composition of proteins towards certain amino acids, so generating low complexity segments in proteins (i.e. departures from the usual twenty amino acid mix).

Nucleic acid pressures can be stronger than protein pressures, even to the extent that some amino acids may be mere ‘placeholders’ (Chapter 13). If a low complexity segment in a protein-encoding region of DNA primarily serves a nucleic acid level function, then the corresponding amino acid sequence in the encoded protein must follow this. On the other hand, if the low complexity segment primarily serves a protein-level function, then the nucleic acid sequence need only follow this with respect to the bases in codons that determine the nature of the encoded amino acid. These are the bases in first and second codon positions.

Bases in third codon positions usually do not determine the nature of the encoded amino acid and so are free to respond to pressures other than the pressure for having a particular amino acid at a particular position in a protein. For example, although phenylalanine is encoded by **TTT**, it is also encoded by **TTC**. Accordingly, if selection operated at the protein level, a **TTC**-based polymer [poly(**TTC**)], might satisfy protein-encoding demands (protein-pressure) for polyphenylalanine just as well as a polymer containing a mixture of **TTCs** and **TTTs**, or just as well as a pure **TTT**-based polymer [poly(**T**)]. However, by virtue of the contained **C**, poly(**TTC**) might satisfy other pressures, such as **GC**-pressure, better than poly(**T**).

Base Pair Pressures

GC-pressure is one of a variety of pressures appearing to operate directly on nucleic acid sequences. By “base pair pressure” is meant a pressure due to a nucleic acid segment’s content of a particular pair of bases (i.e. the segment’s base *composition*, irrespective of base order). So we are not here concerned with dinucleotide (2-mer) frequencies (Chapter 4). From four bases there are six combinations of two different bases, consisting of three mutually exclusive pairs (**GC**, **AT**; **AG**, **CT**; **AC**, **GT**). Implicit in values for (**G+C**)% in DNA segments are values for the reciprocating base pair, (**A+T**)%. Implicit in values for (**A+G**)% (purine-loading pressure) are values for the reciprocating base pair, (**C+T**)%. Likewise **AC**-pressure reciprocates with **GT**-pressure.

‘Pressure’ is a useful intuitive concept implying some force or tendency with a directional (vectorial) component. The terms “**GC**-pressure,” “**AG**-pressure,” and “**AC**-pressure” are used here to indicate evolutionary forces for *departures* from base equifrequencies. However, currently there is no strictly defined usage in the literature. Thus, in some contexts, the effect of **GC**-pressure (i.e. pressure for a change in (**G+C**)%) is considered to increase over the range 0% (**G+C**) up to 100% (**G+C**), while the effect of “**AT**-pressure” is considered to decrease correspondingly from 100% (**A+T**) down to 0% (**A+T**). Here we avoid the latter term (**AT**-pressure), which is implicit in the former (**GC**-pressure). In the present context the effect of **GC**-pressure is considered to increase both from around 50% (**G+C**) up to 100% (**G+C**) (“upward **GC**-pressure”) and from around 50% (**G+C**) down to 0% (**G+C**) (“downward **GC**-pressure”). Similar considerations apply to **AG**-pressure and **AC**-pressure.

If you were ‘the hand of Nature’ and wanted to increase two bases in the coding region of an existing mRNA from scratch, how would you go about it? Take purine-loading for example. The easiest option would be to take advantage of third codon position flexibility so that, despite the **RNY** rule, all codons that can have a purine in this position acquire one (if they do not already possess one). Thus, in the case of leucine with six alternative codons of general form **YYY** and **YYR** (see Table 7.1), you would pick **YYR**. In the case of serine (**YYY**, **YYR** or **RRY**) there is room for some modulation according to the extent of the need for purine-loading. You could pick **YYR** or **RRY**. If selective use of codons with **R** in third positions did not suffice, or **RNY**-pressure was insistent, then you could see if the protein would remain functional when you changed from amino acids with **Y**-containing codons to amino acids with similar properties that had more **R** in their codons. For example, the small hydrophobic amino acid alanine (**RYN**) might be replaced by glycine (**RRN**), which is also small and hydrophobic.

Alternatively, you could try deleting an amino acid with a **Y**-rich codon to see if the protein could manage without it. In so doing you would achieve a relative enrichment for **R**. Or you could insert an amino acid with an **R**-rich codon into some, perhaps innocuous, position in the sequence. Hopefully the protein would not mind. If the latter strategy succeeded, but the mRNA was still insufficiently purine-loaded, then you could try inserting more amino acids with **R**-rich codons next to

the first one. Because many proteins consist of an assembly of independent structural domains, it is indeed possible to insert long segments of amino acids between these domains without interfering with overall protein function. Each inserted amino acid is a placeholder.

You could arrange that the amino acid composition of these inserted segments would be biased towards amino acids with **R**-rich codons. But, some amino acids with **R**-rich codons might be better than other amino acids with **R**-rich codons (i.e. would interfere least with the normal protein function). Which amino acids would be least offensive? Whatever amino acids you choose, the selective use of particular amino acids means that the sequence would be relatively simple. Would the order of the amino acids in the inserted, low complexity, simple sequence be important, or irrelevant? Now let's see how Nature has actually gone about the task. The genomes of Epstein-Barr Virus (EBV) and of *P. falciparum* – both under exceptional **GC**-pressure (upwards and downwards, respectively) – are most informative.

Epstein-Barr Virus

EBV (human herpesvirus 4), the agent of infectious mononucleosis (“mono” or “glandular fever”), has various forms of latency in its human hosts. The initial infection may be asymptomatic. Most people are infected by the age of twenty. Almost certainly the virus is within you at this moment, but you will remain symptom-free. The latent state is associated with the transcription of only a few RNAs from the circular virus genome (“latency-associated transcripts”).

There are various types of latency, but all forms of latency have one RNA transcript in common – the transcript encoding the EBNA1 (Epstein-Barr Nuclear Antigen-1) protein. This contains a long simple-sequence element that can be removed experimentally without interfering with known protein functions. Since the EBV genome appears to be under pressure for compactness (there are few introns and little intergenic DNA), the simple-sequence element should only be present if, *under some circumstance*, it confers an adaptive advantage at the protein level and/or at the level of the encoding nucleic acid (either DNA or RNA). Indeed, an element that increases the length of a gene places the integrity of the gene in jeopardy, since it increases vulnerability to disruption by recombination. One function of the element appears to be the purine-loading of the corresponding mRNA to an extent that cannot be achieved, because of protein-encoding constraints, by the more complex (non-simple sequence) parts of the mRNA [4, 5].

The EBV genome is **GC**-rich and, consistent with the reciprocal relationship between **(G+C)%** and **(A+G)%** (Fig. 12.9), generally has a *low* **(A+G)%** in its genes. So the mRNA-synonymous “coding” DNA strands of *most* EBV genes are *pyrimidine-loaded*. The EBNA1-encoding gene is a notable exception. It is highly *purine-loaded*, mainly by virtue of a long low complexity segment (Fig. 14.2a). The reason for this special feature of the EBNA1-encoding gene will be considered in Chapter 15.

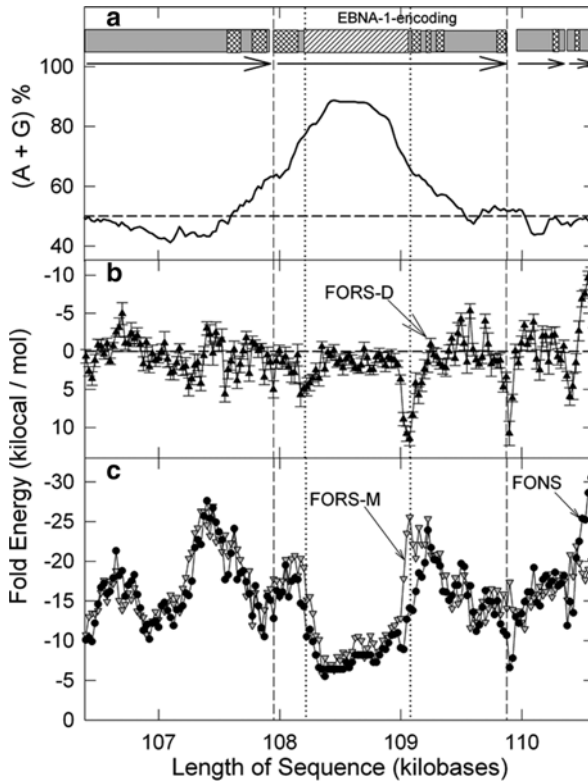
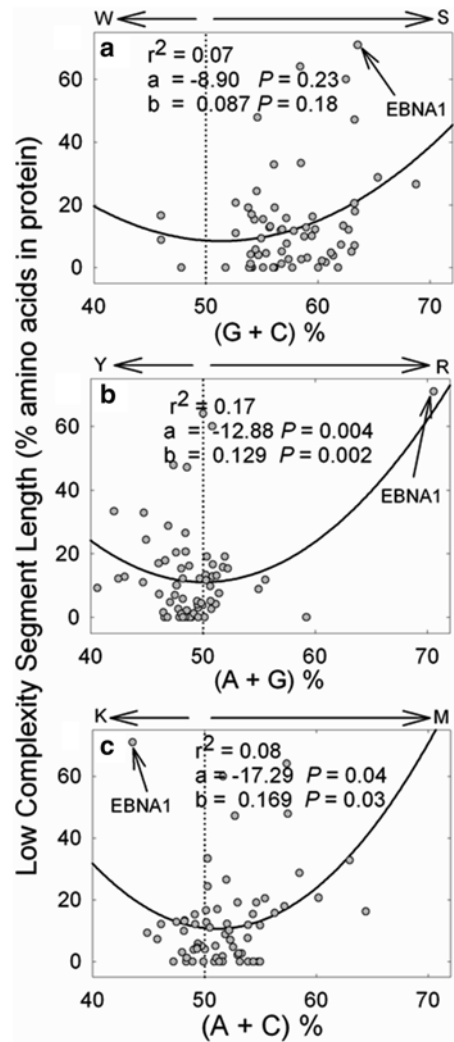


Fig. 14.2 The long low complexity element in the gene encoding the EBNA1 antigen of Epstein-Barr virus (EBV) is purine-loaded (a), and has low folding energy (b, c). In (a) four genes are shown as gray boxes with four horizontal arrows indicating transcription directions. Vertical dashed lines indicate the borders of the EBNA1-encoding gene. The region with the long low complexity, simple-sequence, repetitive element in the EBNA1 gene is shown as a white box with diagonal stripes (demarcated by vertical dotted lines). Other low complexity elements in the EBNA1 and neighboring genes are shown as crosshatched boxes. In (a) purine-loading is shown as $(A+G)\%$ (continuous line). These values are greater than 50% in the EBNA1 gene (purine-loaded), and are less than 50% in neighboring genes (pyrimidine-loaded). In (b), values for the base order-dependent component of the folding energy (FORS-D) are shown with standard errors. In (c), values are shown for the base composition-dependent component of the folding energy (FORS-M), and for the total folding energy (FONS). Low complexity segments were detected using the Seg program

Different EBV genes, and hence the corresponding proteins, vary in their content of low complexity sequence. By eye you can tell from Figure 14.2a that the EBNA1-encoding gene has about 70% low complexity sequence (see boxes with diagonal or hatched lines). When values for percentage low complexity in individual EBV proteins are plotted against values for the various base pressures, points are widely scattered. However, they can be fitted to curves (quadratic parabolas) as shown in Figure 14.3. Ultimately, it may turn out that more exotic equations apply, but the quadratic approximation leads to interpretations that seem to make sense.

Fig. 14.3 Variation of low complexity segment length in EBV proteins with content of (a) **G+C**, (b) **A+G**, and (c) **A+C**, in the corresponding genes. The angled arrows indicate the gene encoding EBNA1 protein. Points were fitted to curves (parameters shown). Horizontal arrows pointing to **W** and **S** indicate progressive enrichment for either the **W** bases (**A+T**) at low (**G+C**)%, or the **S** bases (**G+C**) at high (**G+C**)%. **Y** and **R** indicate enrichment for either the **Y** bases (**C+T**) at low (**A+G**)%, or the **R** bases (**A+G**) at high (**A+G**)%. **K** and **M** indicate enrichment for either the **K** bases (**G+T**) at low (**A+C**)%, or the **M** bases (**A+C**) at high (**A+C**)%



At base compositions around 50% most points are located near minima. Thus most genes have very little low complexity sequence (i.e. they are mainly of high complexity, with equifrequencies of bases that are irregularly ordered). The exceptional low complexity value, corresponding to the EBNA1-encoding gene, is at one extreme of the curves [6]. The scatter of points, either to the left or right of the vertical dotted line (marking 50%), shows the general tendency of EBV genes towards **S**-loading (high **G+C**; Fig. 14.3a), and **Y**-loading (low **A+G**; Fig. 14.3b). The maverick EBNA1 gene is both highly **S**-loaded and highly **R**-loaded. The plots become more informative when the base pressures exerted by different codon positions are plotted independently (Fig. 14.4).

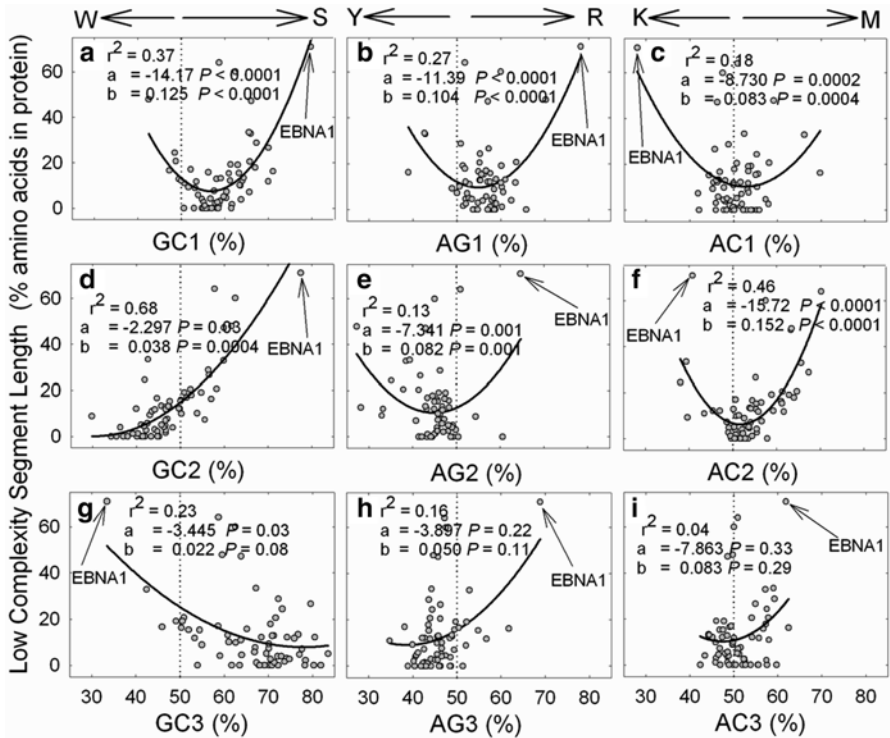


Fig. 14.4 Variation of low complexity segment length in EBV proteins with the content of either **G+C** (a, d, g), **A+G** (b, e, h), or **A+C** (c, f, i), at first (a, b, c), second (d, e, f) and third (g, h, i) codon positions (e.g. on the X-axis **GC1** refers to the (**G+C**)% of first codon positions). Details as in Figure 14.3

By comparing the different plots, sense emerges from a bewildering scatter of points. The amino acid-determining first and second codon positions tend to be the best *predictors* of the amount of low complexity segment. Points for first and second codon positions best fit the curves, so you would have more faith in predicting the degree of complexity from values for these positions than from values for third codon positions. This can be determined by examining the various values shown in each figure.

Values for the square of the correlation coefficient (r^2) can range from zero to one and indicate how much of the observed variability (i.e. 0% to 100%) can be attributed to low complexity segment content – 68% in the case of the **G+C** content of the second bases of codons as in Figure 14.4d, and 27% in the case of the **A+G** content of the first bases of codons as in Figure 14.4b.

The equations of quadratic parabolas have two key determinants (a and b); the probability values (P) associated with these (ranging from zero to one) indicate how likely it is that the match is *not* a good one (i.e. low P values indicate a *good* match of points to the curve). Third codon positions – those least likely to relate to the

encoding of amino acids – tend to have the highest P values (i.e. there is a *poor* match of points to the curve; Figs. 14.4g–i).

With the caveat that in some species third codon positions can affect the speed or accuracy of translation (Chapter 7), third positions can be regarded as controls (i.e. they provide a frame of reference). You can even think of them as mini-introns. Like non-genic DNA segments, *whether they are part of low or high complexity segments, they should be equally free to respond to pressures other than the encoding of amino acids*. Thus, the base compositions of third codon positions (X-axes) should be relatively independent of the quantity of low complexity segment in protein (Y-axes). Indeed, as expected, values for r^2 are low and values for P are high.

On the other hand, first and second codon positions, when in *high* complexity segments, primarily play the amino acid-encoding role and *are restrained* from responding to other pressures. But when in *low* complexity segments, first and second codon positions, if encoding mere placeholders, *would not be so restrained*. Consistent with this, the low complexity segment in the maverick EBNA-1-encoding gene appears free to serve other pressures by virtue of the base content not only of third codon positions, but also of the first and second codon positions.

The flexibility of third codon positions would have been insufficient to meet the demands of pressures other than the encoding of proteins. So first and second codon positions would have been *recruited* to assist in dealing with these demands (c.f. Chapter 12). In simple-sequence elements, this assistance would not be to the detriment of the amino acid-encoding role of first and second codon positions, because the nature of the encoded amino acids would be largely irrelevant to the function of the protein.

We now go back to Figure 14.2 to consider DNA secondary structure. In the DNA encoding the EBNA1 repetitive element there are up to 800 purines per kilobase so that there are relatively few pyrimidines to pair with purines for formation of the stems of stem-loop secondary structures (Chapter 5). Thus, the potential for forming secondary structure is greatly decreased as measured by the FONS value (Fig. 14.2c). This value decomposes into a base order-dependent component (FORS-D; Fig. 14.2b) and a base composition-dependent component (FORS-M; Fig. 14.2c). Compared with neighboring regions, in the region of the repetitive element base order contributes poorly to the folding potential (assessed as negative kilocalories/mol). Indeed, base order supplements base composition in *constraining* the folding potential (so most FORS-D values, being below the horizontal dashed line, are slightly positive; Fig. 14.2b).

Since secondary structure in DNA is likely to enhance recombination (Chapter 10), this predicts *decreased* recombination between regions of the gene corresponding to the amino-terminal and carboxy-terminal ends of the EBNA1 protein. The gene would otherwise be more vulnerable to fragmentation by recombination, due both to the increase in gene length resulting from the presence of the low complexity element, and to the increased probability that an EBV low complexity element will find a willing partner in the recombination minuet (i.e. another low complexity element with a similar sequence), either elsewhere in its own genome, or in the host cell's genome.

Apart from the low complexity *repetitive* element in the gene encoding the EBNA1 protein, there are other low complexity elements (*non-repetitive*) in neighboring genes (Fig. 14.2a). Thus, the immediate upstream gene encodes two small elements. The two immediate downstream genes each have only one small element. However, in many respects the EBNA1-encoding gene is *different from all other EBV genes*, most of which are not expressed during latency. Indeed, there are more similarities with genes of the parasite responsible for the most malignant form of malaria – *Plasmodium falciparum*.

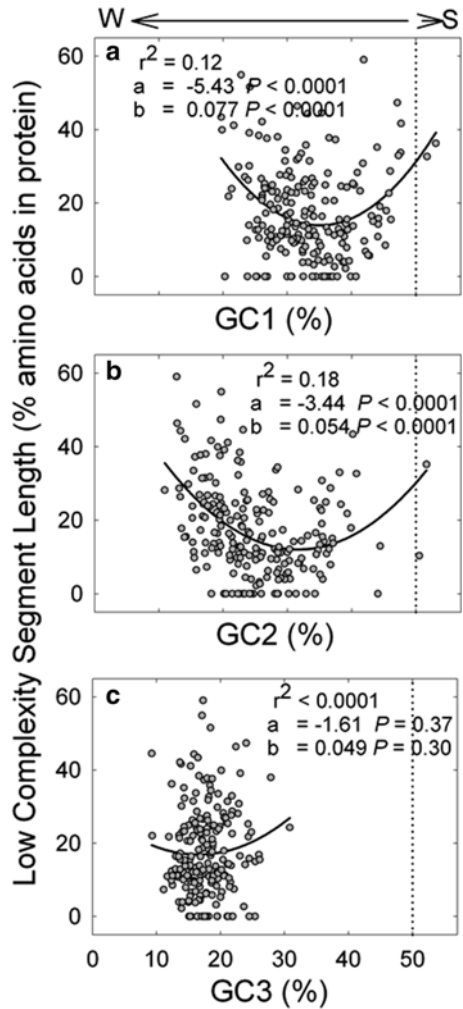
Malaria

Low complexity segments with biased amino acid composition are abundant in the proteins of *P. falciparum* (Chapter 13). In Italy, Elizabetta Pizzi and Clara Frontali wondered whether these low complexity protein segments have special properties that provide a functional explanation for their existence [7]. They compared the properties of amino acids in low complexity sequences with the properties of amino acids in high complexity sequences. The amino acid biases in low complexity segments did not correlate with properties of individual amino acids (hydrophobicity, volume, flexibility). However, a correlation was found with the **A** content of the codons. There was a correlation, not at the protein level, but *at the nucleic acid level*.

In the exceptionally **AT**-rich malaria parasite genome the third codon position is dominated by intense downward **GC**-pressure, making it one of the most **AT**-rich eukaryotic species (Fig. 12.7a). Being less influenced by the nature of the encoded amino acid, third codon positions have more flexibility to respond to **GC**- and **AG**-pressures. However, this flexibility can be insufficient to meet the demands both of downward **GC**-pressure and of upward **AG**-pressure (purine-loading pressure). So first and second codon positions are recruited. This threatens to change the nature of the encoded amino acid, which could then disturb protein function. So, again, a solution is to create extra protein segments that lie *between* functional protein domains. These extra segments should neither play a role in, nor impede, normal protein function. The amino acids in these segments are mere placeholders. *Thus, the low complexity segments represent evolutionary adaptations allowing genes, through the compositions of first and second codon positions (as well as third codon positions), to respond to extreme base compositional demands that could not be met by high complexity DNA segments alone (despite their flexibility at third codon positions).*

Consistent with this, as in the case of EBV, the quantity of low-complexity sequence correlates well with the base composition of first and second codon positions, but, as expected, less well with that of third codon positions. Figure 14.5 shows that the percentage content of low complexity segment in malaria parasite genes relates to the **(G+C)%** of each gene in a way that is dependent on codon position. By eye you can tell, from the clustering of points relative to the X-axis, what average **(G+C)%** values for different codon positions are likely to be. Average **(G+C)%** values become progressively less as codon position changes from first

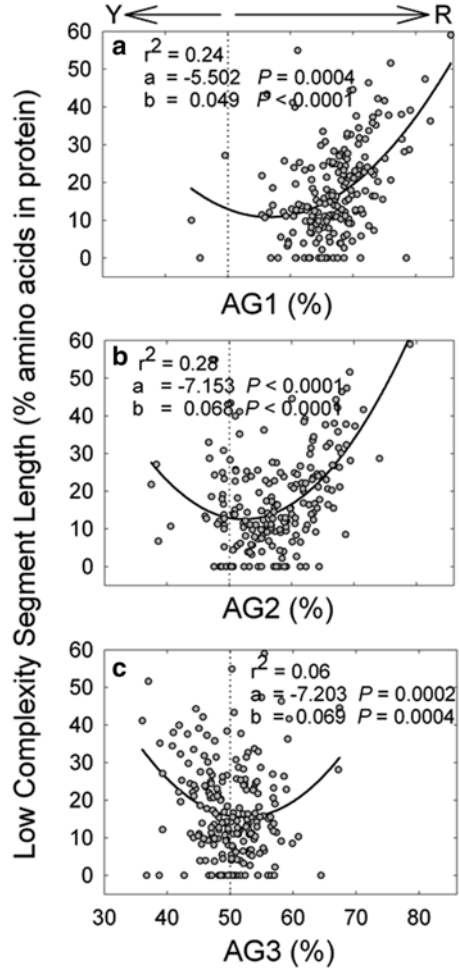
Fig. 14.5 Variation of low complexity segment length, in the proteins encoded by chromosome 2 of *P. falciparum*, with the (G+C) content of first (a), second (b) and third (c) codon positions (expressed as GC1, GC2 and GC3 on the X-axis). Parameters of second order linear regression plots are shown in each figure. Horizontal arrows pointing to W and S indicate the enrichment for either the W bases (A+T) at low values of (G+C)%, or the S bases (G+C) at high values of (G+C)%



position (Fig. 14.5a) to third position (Fig. 14.5c). The extreme value of the third position (around 18% G+C) indicates that it has the most flexibility to respond to downward GC-pressure. This is largely *independent* of the percentage of low complexity segment (indicated by the low r^2 value, and high P values). Thus, whatever their content of low complexity sequence, on the basis of their third codon position (G+C)% values genes contribute strongly to downward GC-pressure.

On the other hand, data points for first and second codon positions fit to curves with minima corresponding to genes with low percentages of low complexity segment (Figs. 14.5a, b). Most data points relate to the leftward limbs of the curves. For first and second codon positions the curves show that acquisition by certain genes of low (G+C)% values, correlates with an increased content of low complexity segment. *The more AT-rich a gene, the more low complexity segment it contains.* In the case of second codon positions 18% of the variation can be explained on this basis ($r^2 = 0.18$).

Fig. 14.6 Variation of low complexity segment length in proteins encoded by chromosome 2 of *P. falciparum*, with the (A+G) content of first (a), second (b) and third (c) codon positions. Horizontal arrows pointing to **Y** and **R** indicate the enrichment for either the **Y** bases (C+T) at low values of (A+G)%, or the **R** bases (A+G) at high values of (A+G)%



Similar results are seen with **AG**-pressure (Fig. 14.6). Departure from equipfrequency of bases (50% purine: 50% pyrimidine) is greatest in the case of first codon positions. Average (A+G)% values become progressively less as codon position changes from first position (Fig. 14.6a) to third position (Fig. 14.6c). This is consistent with the **RNY** rule (purine for first codon base, pyrimidine for third codon base; see Chapter 7). Values for first and second codon positions can best be fitted to the curves; most data points relate to rightward limbs (Figs. 14.6a, b). *The more AG-rich a gene, the more low complexity segment it contains*. There is a positive correlation between a gene's purine content (i.e. upward AG-pressure) and its content of low complexity sequence. In the case of second codon positions, 28% of the variation can be accounted for on this basis ($r^2 = 0.28$).

Values for third codon positions show poor correlations (i.e. $r^2 = 0.06$; Fig. 14.6c). So, as expected, the distinction between genes with low and high percentages of

low complexity segment length is least for this position. Note that the correlations (r^2 values) in the case of purine-loading (Figs. 14.6a, b) are better than in the case of GC-loading (Figs. 14.5a, b). This is consistent with the notion that purine-loading is usually the *dominant* pressure with respect to increasing the length of a protein by insertion of low complexity elements (Chapters 12 and 13).

Conflict with Fold Potential

The long low complexity sequence in the EBNA1 gene appears to countermand the potential of the encoding DNA duplex to extrude single-stranded DNA that adopts a stable stem-loop secondary structure (Fig. 14.2). Low complexity sequences in *P. falciparum* genes might be expected to act similarly – but not necessarily so. A prominent example is the gene encoding the circumsporozoite protein (CSP; Fig. 14.7). This is expressed in the form of the malaria parasite (sporozoite) that is initially injected by infected mosquitoes and enters liver cells. A host immunized against parts of the CSP should be able to nip the infection in the bud.

So we should not be surprised to learn that the two end segments of the protein (N-terminal and C-terminal segments) correspond to CSP antigens that vary between different individual parasites (i.e. are polymorphic) and are evolving rapidly under positive Darwinian natural selection. There is one low-complexity, non-repetitive element, and a central low-complexity simple sequence element encoding repeats of either eight amino acids (the octapeptide NVDPNANP) or four amino acids (the tetrapeptide NANP; Fig. 14.7a). Downstream of the single exon CSP-encoding gene is a three exon gene encoding an unidentified (therefore referred to as a “hypothetical”) protein. Here the upstream half of the small middle exon has a low complexity, non-repetitive, element.

The codons of the low complexity segments in the CSP-encoding gene tend to be A-loaded, and contribute to the generally high level of purine-loading (i.e. the curve in Fig. 14.7a is mainly above 50% (A+G) in the region of the gene). In regions corresponding to the N- and C-terminal segments this is achieved mainly by A bases, with a small contribution from G bases (the two bases are not shown separately in the figure). For the middle segment this is achieved *entirely* by A bases, and there is an *excess* of C compared with G (not shown in the figure). The inclusion of the low-complexity regions can be considered as a device to enhance purine-loading, as in the case of the EBNA1 antigen (Fig. 14.2). However, the fact that the low value for G seems to work *against* this suggests other factors are in play.

Figures 14.7b and 14.7c show the potential of windows in the segment to form secondary structure. The base order-dependent component (FORS-D) is impaired (i.e. tends to be less negative) in regions of the CSP gene corresponding to the N- and C-terminal segments of the protein; this is characteristic of sequences under positive selection [8]. Consistent with the high purine-loading of the repetitive middle segment, the base *composition*-dependent component of the folding potential

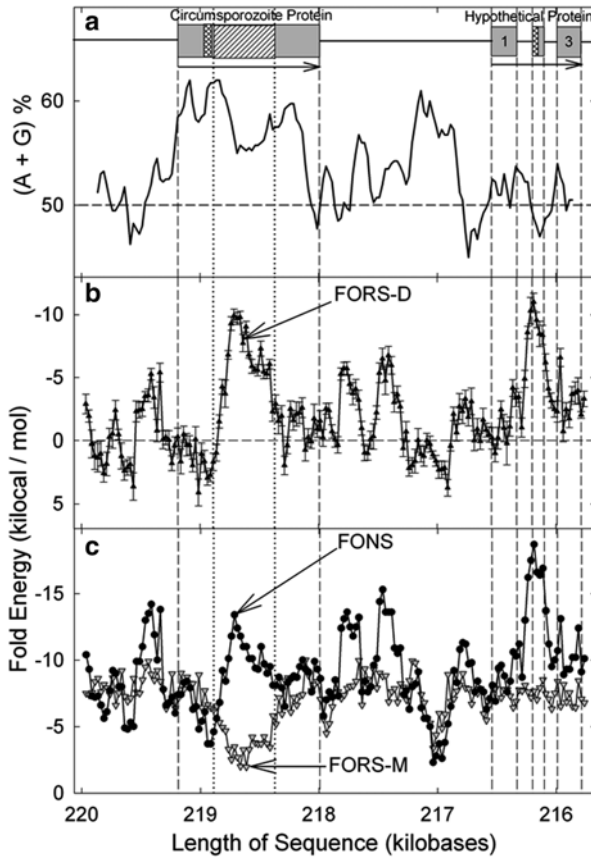


Fig. 14.7 The long low complexity element of the *P. falciparum* circumsporozoite protein-encoding gene supports two nucleic acid level functions – purine-loading, and base order-dependent stem-loop potential. Details are as in Figure 14.2. Vertical dotted lines indicate the boundaries of the long low complexity, simple-sequence, repetitive element. Vertical dashed lines indicate the exon borders of the single exon CSP gene and of a three exon gene corresponding to an unidentified “hypothetical” protein. Note, from the genome sequence numbering on the X-axis, that the sequence was reversed so that the ‘bottom’ strand is the analyzed strand (mRNA-synonymous coding strand) and genes formally designated in GenBank as transcribing to the left are shown as transcribing to the right

(FORS-M) is greatly *decreased* in the middle segment (Fig. 14.7c); but this is more than compensated for by an *increase* in the base *order*-dependent component (FORS-D; Fig. 14.7b).

Hence, the overall stem-loop potential (FONS) in this region is high (Fig. 11.7c). This suggests that recombination between the two ends of the gene might be *enhanced* (see later). In general, stem-loop potential in genes tends to be relegated to introns, especially in the case of genes under positive selection where introns may be more

conserved than exons (Chapter 13). In the case of the CSP gene, what might have been an intron appears to have been retained and expressed in the protein sequence, perhaps because of an adaptive value at the mRNA level (i.e. purine-loading) and/or at the protein level.

Similar considerations may apply to the second exon of the downstream gene. Whereas there is purine-loading (i.e. **AG** greater than 50%) both in the single exon CSP gene and in the first and third exons of the downstream gene (i.e. Szybalski's transcription direction rule is obeyed), the second exon is pyrimidine-loaded (i.e. **AG** is less than 50%). Much of the second exon consists of a low complexity element, which strongly supports base order-dependent stem-loop potential (high negative FORS-D value).

Roles of Low Complexity Segments

Since low complexity is often associated with departure from base-equiprequencies, it is not surprising that proteins with high percentages of low complexity segments correspond to the extremes of base pair percentages (Fig. 14.1). The issues are whether low complexity results from selective pressures, and if so, whether these pressures operate primarily at the nucleic acid level, or at the protein level. A sequence primarily generated in response to a nucleic acid level pressure might secondarily have acquired a protein level function, thus making it difficult to determine which was primary. A clear assignment of nucleic acid level roles can assist this task.

Unlike the first and second codon positions, the base composition of third codon positions (the “mini-introns”) correlates poorly with the percentage of low complexity segment in proteins, especially in the case of (**G+C**)% (Fig. 14.5). This suggests that, for this position, low and high complexity segments are *equally* free to respond to base compositional pressures. First and second positions are constrained by other pressures. In high complexity protein segments, mainly corresponding to conserved globular domains, these pressures include pressure to retain specific protein function (protein-encoding pressure). On the other hand, the encoding of amino acids by low complexity segments is largely determined by the base compositional pressures. Such segments mainly correspond to non-globular hydrophilic surface domains that do not play a critical role in a protein's structure and specific function.

Yet, some protein-level function might still be of value. In the case of the CSP gene, association with enhanced stem-loop potential (Fig. 14.7) would be expected to promote recombination and hence enhance CSP polymorphism, a feature conferring protection against host immune defenses. The long low complexity segment in the gene encoding the EBNA1 protein does not enhance stem-loop potential (Fig. 14.2), so probably functions solely as a means of purine-loading the mRNA to avoid formation of double-stranded RNA, which would alert the host's immune system (see Chapter 15).

Short Tandem Repeats

The low complexity segments contain either non-repetitive, or repetitive, amino acid sequences, the latter being prevalent in long low complexity elements. That in the EBNA1 protein has a repeat of two glycines and an alanine. If Nature's 'purpose' is merely to purine-load, then there should be no particular virtue in having placeholder amino acids *ordered* in this way:



Surely, a disordered sequence might serve just as well – for example:



The ordering (periodicity) might somehow help the insert remain innocuous with respect to the function of the protein. Hence it would be selected for (i.e. natural selection would prefer viruses with ordered amino acids in the low complexity element, over viruses with disordered amino acids in the low complexity element). However, in some cases the ordering can be shown to depend on the 'mechanical' (i.e. chemical) process by which the repeat was first generated. In fact, the 'ratchet' can get out-of-hand and the insert can grow so long that it begins to interfere with protein function, often by making the protein insoluble. Clear examples of this are found in the "trinucleotide expansion" diseases, which often affect central nervous system function, and include the human brain disorder known as Huntington's disease.

In Huntington's disease there is a long repeat of the trinucleotide **CAG** in the 'top' mRNA-synonymous strand of a protein-encoding exon. The corresponding mRNA thus contains a long **CAG** tract, which is in the codon reading frame (i.e. it is read as **CAG**, not as **AGC** or **GCA**; Fig. 14.8a). Since **CAG** is a codon for the amino acid glutamine, the protein contains a long polyglutamine tract.

The **CAG** tract in the DNA is likely to have expanded because it can form an internal stem-loop structure (Fig. 14.8b). The stem can form because of the ordering of the **C** and **G** bases. This periodicity is conducive to stem-formation. The **A** bases are displaced and do not positively support the stability of the structure. A possible mechanism for the expansion, that may operate both during mitosis and meiosis, is shown in Figures 14.8c–h [9].

In normal humans the gene that, when changed, causes Huntington's disease, has a short **CAG** tract (less than 84 bases). This can be presumed to satisfy the normal need for purine-loading. However, as shown in Figure 14.8 the tract can increase in length. In fact, the longer it gets, the more likely it is to adopt the stem-loop structure, so facilitating a further expansion. Once it reaches a critical length (greater than 84 bases) disease features begin to appear. Initially, these features appear *late* in life, so that there may have been time to produce children. Yet, within a generation, the tract can expand, so in these children the tract may be longer than in the affected parent. This causes the disease to be more severe and to begin earlier in life.

Note, however, that **CAG** tracts can adopt stem-loop secondary structures with long stems not only at the DNA level, but also at the RNA level. The lengths of double stranded RNA (dsRNA) so generated should be sufficient to trigger intracellular alarms, thus activating the immune system [11, 12]. Perhaps, dsRNA is responsible for (or is a necessary cofactor in) the development of Huntington's disease? Consistent with this, in some other trinucleotide expansion diseases that affect the nervous system, (e.g. spinocerebellar ataxia) the expansions are *not* in protein-encoding regions. The pathogenic agents appear to be RNA transcripts [13]. We will consider this in Chapter 15.

Many diseases associated with intracellular, and sometimes extracellular, protein aggregates are brain diseases that slowly progress over many years. Notable examples are the “prion” diseases (e.g. “mad cow disease”), and Alzheimer's disease (see Chapter 17). An obvious explanation would be that the proteins are brain-specific. Yet, often the proteins are *not* brain-specific. The proteins are expressed in non-neural tissues, but aggregates are not seen and tissue function is not impaired. Perhaps of relevance is the fact that the brain is “immunologically privileged,” being partially protected from the immune system by membranes (“blood-brain barrier”). This raises the possibility that the immune system can recognize cells with aggregates in non-neural tissues and destroy them. The cells can then be replaced by the division of neighboring cells or of “stem cells” (whose primary role is replacement of ‘worn out’ cells). But somehow this is not possible in neural tissues (see Chapter 16).

AC-Pressure and Asymmetric Fold Potential

The **AT**-rich DNA sequence 5'AAAAAAAAATTTTTTTT3' obeys PR2 and has the potential to fold into a stem-loop (hairpin-like) configuration. In an antiparallel duplex containing this sequence in its ‘top’ strand, the ‘bottom’ strand – 3'TTTTTTTTAAAAAAAA5' – has the same base composition and the same potential to fold into a stem-loop configuration. Thus, there is *symmetry* in the folding energies of top and bottom strands. Furthermore, the two strands would display the same buoyant density when subjected to centrifugation in alkaline density gradients (Chapter 11; the alkali keeps the strands separate).

In contrast, the **AC**-rich sequence 5'AAAAAAAAACCCCCCCC3' (top strand) does not obey PR2, has little potential to fold, and has a different base composition from the bottom strand – 3'TTTTTTTTGGGGGGGG5'. However, since **G** can form a weak Watson-Crick base-pair with **T**, the latter sequence retains some potential to fold. Thus, there is *asymmetry* in the potential folding energies of top and bottom strands. Furthermore, by virtue of their composition differences, the two strands display different densities in alkaline density gradients. When banded in such gradients one strand can be regarded as a “satellite” of the other. Indeed, a significant class of repetitive sequences was discovered by virtue of this satellite property [14].

(**A+C**)%, reciprocating with (**G+T**)%, hardly seems cast for a spectacular role (Figs. 14.3c, 14.4c, f, i). Yet, as a sequence grows richer in **A** and **C** its potential to extrude stem-loop structures *declines*. **A** forms base pairs poorly with **A**, and **C**

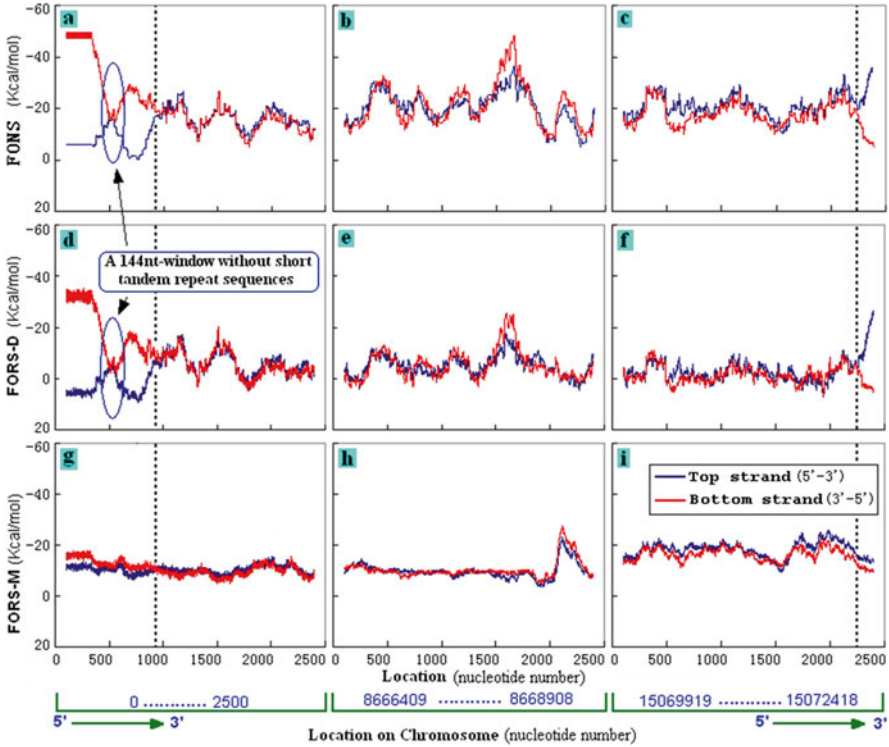


Fig. 14.9 Telomeres display strand asymmetry. Segments (2.5 kb) from the 5' end (a, d, g), middle (b, e, h) and 3' end (c, f, i) of chromosome 1 of *C. elegans* were folded to determine FONS values (a, b, c), FORS-D values (d, e, f), and FORS-M values (g, h, i). Telomeric boundaries are indicated by the vertical dashed line. For further details see Figure 5.9.

forms base pairs poorly with **C**. **A** and **C** form base pairs poorly with each other. Often they come as short tandem repeats – referred to as “microsatellite” sequences – the simplest being of the form **ACACACAC** ... , reciprocating with **GTGTGTGT** ... in the complementary strand. Since base composition and order both influence folding symmetry, the degree of symmetry/asymmetry is modified when bases are ordered in this fashion.

If an **AC**-rich region transitions from classical duplex form to extruded stem-loop form, there is asymmetry of structure [15]. Figure 14.9 contrasts the folding potentials of the **GT**- and **AC**-rich ends of chromosomes (telomeres) with that of an internal chromosome segment similar to that shown in Figure 5.9. The telomeres contain heptameric tandem repeats, with **TTAGGC** in the bottom strand at the 5' telomere (Fig. 14.9a) retaining more fold potential (FONS) than its complementary repeat – **GCCTAA**. The converse applies at the 3' telomere (Fig. 14.9c). In both cases, the difference in folding energies is due more to *base order* (Figs. 14.9d, f) than to *base composition* (Figs. 14.9g, i).

Repetitive Sequences and Speciation

Apart from telomeres, microsatellite repeats are also found in centromeres and are dispersed at various other chromosomal locations. In most cases strand asymmetry is evident [15]. Inspired by the concept of ‘selfish DNA,’ in 1980 Orgel and Crick held that repetitive sequences were autonomous entities that made “no specific contribution to the phenotype” [16]. But according to the model advanced in Chapter 10, this asymmetry should impede meiotic pairing and recombination. Hence divergence into species would be favored. Indeed, it was argued in 1981 that, since repetitive sequences tend to vary (a) more than unique sequences and (b) more between species than within a species, they “could act as a blunt instrument driving speciation by reproductive isolation alone, without reference to adaptation” [17]. Thus, repetitive sequences could be more important at the species level, than at the level of the individual phenotype [18].

However, a species population usually displays *much* microsatellite variability (polymorphism). The individual organisms that comprise a species differ greatly *from each other* in their microsatellites. This can be sufficient to affect chromosome lengths, and can even be used to ‘finger-print’ individuals for forensic purposes [19]. *Yet those organisms remain within species limits* [20]. While, paradoxically, **GT**-richness *favors* recombination [21, 22], it is a specialized form of *somatic* recombination. In prokaryotes recombination may be associated with ‘**GT**-rich islands’ (approx. 1 kb) that contain strand-specific “crossover hotspot instigator” sequences – CHI sequences (**GCTGGTGG**) [23]. For their function CHI sequences require a **GT**-rich base composition, but the order of bases is also important. In higher eukaryotes specific **GT**-rich sequences (**GGGGCTGGG**) embedded in **GT**-rich regions are important in the strand-biased recombination events that allow immunoglobulin class switching for antibody formation (see Part V). The folding asymmetry can also facilitate access of a mutagenic enzyme (deoxycytosine deaminase) to the ‘bottom’ **AC**-rich strand, so facilitating antibody diversification [24].

Thus it seems probably that regions adapted for special forms of somatic recombination would not influence germ-line meiotic recombination. Indeed, short tandem repeats are found in recombination ‘cold-spots’ (regions of suppressed recombination [25], which include centromeres [26]. High variability (associated with microsatellite content) would fail to register in terms of the pairing incompatibility between homologous chromosomes that is expected to precede speciation (Chapter 10). Consistent with this, there is suggestive evidence that PR2 ‘accounting’ between chromosomes is *uninfluenced* by their content of short tandem repeats [27].

On the other hand, noting an association of some microsatellite DNAs with meiotic recombination ‘hotspots,’ Wayne Wahls has suggested a role in the termination, rather than initiation, of meiotic recombination [28]; some characteristic of satellites is held to arrest the branch migration that may follow formation of Holliday junctions (Figs. 11.2, 11.4). Thus, recombination might launch in a flanking non-satellite region and terminate within the satellite due to some inhibitory influence.

Meiotic Hotspots

Meiotic recombination in eukaryotes also has its hotspots – regions (approx. 2kb) where recombination is greatly enhanced. Again, at hotspot centers there are specific base sequences (13 bases) with surrounding “islands” showing regularities in base composition. Sequences in mammals, which are recognized by specific proteins (e.g. PRDM9), can show upstream **R**-loading and downstream **Y**-loading – a skew transition similar to that seen at sites of initiation of DNA replication (e.g. Fig. 6.4) [29].

A major difference is that replication initiation sites remain relatively constant, so the enzymes of DNA replication (e.g. DNA polymerases) are not in an adaptive treadmill (Chapters 5, 12), their genes having to mutate to keep up. However, while recombination initiation sites may endure for many generations, eventually they disappear presumably due to the *directional* conversion (Fig. 11.2) that accompanies recombination (see Chapters 17, 18). Thus, “hotspots seed their own destruction” [30] and fresh sites with different sequence characteristics emerge. Accordingly, proteins such as PRDM9 are themselves transient, displaying extensive polymorphism (allelic variation) over time. Their genes do have to mutate to keep up.

Long-Range Periodicities

In Chapter 7 we encountered short-range **RNY** periodicities at the codon level. Indeed, in protein-encoding regions the level of every third base tends to remain constant (Fig. 11.3). Here we have encountered the local periodicities displayed by low complexity elements, whether they are in coding or non-coding regions. Are there long-range periodicities?

By virtue of their structure we sometimes encounter long-range periodicities in written texts. In this book every 10 to 20 pages the word “Chapter” appears in enlarged type with a number following it. This also applies to DNA texts. For example, the DNA double-helix is normally slightly unwound (Chapter 6). Where other pressures do not countermand, every 10-11 Watson-Crick base pairs there are dinucleotide sequences (‘bendability signals’) that support this. The signals are weak but the periodicity becomes quite clear when long sequences are examined [31]. The number of base pairs corresponds to about one turn of the double-helix. Periodicities occurring at intervals of 200 bases, 400 bases, and even as long as 500 kilobases, have been related to aspects of chromosome structure [32, 33].

Summary

Protein segments that contain few of the possible twenty amino acids, sometimes in tandem repeat arrays, are referred to as containing “simple” or “low complexity” sequence. Many proteins of the malaria parasite, *P. falciparum*, are longer than their

homologs in other species by virtue of their content of such low complexity segments that have no known function; these are interspersed among segments of higher complexity to which function can often be ascribed. If there is low complexity at the protein level then there is low complexity at the corresponding nucleic acid level (often seen as a departure from equipfrequency of the four bases). Thus, low complexity may have been selected *primarily* at the nucleic acid level and low complexity at the protein level may be secondary. The amino acids in low complexity segments may be mere placeholders. The amino acid composition of low complexity segments should then be more reflective than that of high complexity segments on forces operating at the nucleic acid level – such as **GC**-pressure, **AG**-pressure, **AC**-pressure, and fold pressure. Consistent with this, for amino acid-determining first and second codon positions, low complexity segments show significant contributions to downward **GC**-pressure (decreased percentage of **G+C**) and to upward **AG**-pressure (increased percentage of **A+G**). When not countermanded by high contributions to **AG**-pressure, which locally decrease fold potential, low complexity segments can also contribute to fold potential. Thus they can influence recombination within a gene. Short tandem repeat sequences under **AC**-pressure violate PR2 and are extruded *asymmetrically* as stem-loops from DNA duplexes. This may favor specialized forms of somatic recombination, but probably does not affect meiotic pairing of chromosomes. These observations have implications for our understanding of malaria, infectious mononucleosis, and brain diseases in which protein aggregates accumulate.

References

1. Bateson G (1964) The logical categories of learning and communication. In: *Steps to an Ecology of Mind*. Paladin, St. Albans (1973) pp 250–279
2. Sibbald PR (1989) Calculating higher order DNA sequence information measures. *Journal of Theoretical Biology* 136:475–483
3. Wan H, Wootton JC (2000) A global complexity measure for biological sequences. AT-rich and GC-rich genomes encode less complex proteins. *Computers & Chemistry* 24:71–94
4. Cristillo AD, Mortimer JR, Barrette IH, Lillicrap TP, Forsdyke DR (2001) Double-stranded RNA as a not-self alarm signal: to evade, most viruses purine-load their RNAs, but some (HTLV-1, EBV) pyrimidine-load. *Journal of Theoretical Biology* 208:475–491
5. Forsdyke DR (2002) Selective pressures that decrease synonymous mutations in *Plasmodium falciparum*. *Trends in Parasitology* 18:411–418
6. Xue HY, Forsdyke DR (2003) Low complexity segments in *Plasmodium falciparum* proteins are primarily nucleic acid level adaptations. *Molecular & Biochemical Parasitology* 128:21–32
7. Pizzi E, Frontali C (2001) Low-complexity regions in *Plasmodium falciparum* proteins. *Genome Research* 11:218–229
8. Forsdyke DR (1996) Stem-loop potential: a new way of evaluating positive Darwinian selection? *Immunogenetics* 43:182–189
9. Figueroa AA, Delaney S (2010) Mechanistic studies of hairpin to duplex conversion for trinucleotide repeat sequences. *Journal of Biological Chemistry* 285:14648–14657
10. Suhr ST, Senut M-C, Whitelegge JP, Faull KF, Cuizon DB, Gage FH. (2001) Identities of sequestered proteins in aggregates from cells with induced polyglutamine expression. *Journal of Cell Biology* 153:283–294

11. Tian B, et al. (2000) Expanded CUG repeat RNAs form hairpins that activate the double-stranded RNA-dependent protein kinase PKR. *RNA* 6:79–87
12. Peel AL, Rao RV, Cottrell BA, Hayden MR, Ellerby LM, Bredesen DE (2001) Double-stranded RNA-dependent protein kinase, PKR, binds preferentially to Huntington's disease (HD) transcripts and is activated in HD tissue. *Human Molecular Genetics* 10:1531–1538
13. O'Rourke JR, Swanson MS (2009) Mechanisms of RNA-mediated disease. *Journal of Biological Chemistry* 284:7419–7423
14. Flamm WG, Walker PM, McCallum M (1969) Some properties of the single strands isolate from the DNA of the nuclear satellite of the mouse (*Mus musculus*). *Journal of Molecular Biology* 40:423–443
15. Zhang C, Xu S, Wei J-F, Forsdyke DR (2008) Microsatellites that violate Chargaff's second parity rule have base order-dependent asymmetries in the folding energies of complementary DNA strands and may not drive speciation. *Journal of Theoretical Biology* 254:168–177
16. Orgel LE, Crick FH (1980) Selfish DNA: the ultimate parasite. *Nature* 284:604–607
17. Robertson M (1981) Gene families, hopeful monsters and the selfish genetics of DNA. *Nature* 293:333–334
18. Flavell RB (1982) Sequence amplification, deletion and rearrangement: major sources of variation during species divergence. In: Dover GA, Flavell RB (eds) *Genome Evolution*. Academic Press, San Diego, pp 301–323
19. Jeffreys AJ (1985) Individual-specific 'fingerprints' of DNA. *Nature* 316:76–79
20. Biémont C (2008) Within species variation in genome size. *Heredity* 101:297–298
21. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. *Nature Reviews Genetics* 5:435–445
22. Majewski J, Ott J (2000) GT repeats are associated with recombination on human chromosome 22. *Genome Research* 10:1108–1114
23. Lao PJ, Forsdyke DR (2000) Crossover hot-spot instigator (CHI) sequences in *Escherichia coli* occupy distinct recombination/transcription islands. *Gene* 243:47–57
24. Huang F-T, Yu K, Balter BB, Selsing E, Oruc Z, Khamlichi AA, Hsieh C-L, Lieber MR (2007) Sequence dependence of chromosomal R-loops at the immunoglobulin heavy-chain Smu class switch region. *Molecular & Cellular Biology* 27:5921–5932
25. Gvozdev VA, Kogan GL, Usakin KA (2005) The Y chromosome as a target of acquired and amplified genetic material in evolution. *BioEssays* 27:1256–1262
26. Talbert PB, Henikoff S (2010) Centromeres convert but don't cross. *PLOS Biology* 8:e1000326
27. Forsdyke DR, Zhang C, Wei J-F (2010) Chromosomes as interdependent accounting units. The assigned orientation of *C. elegans* chromosomes minimizes the total W-base Chargaff difference. *Journal of Biological Systems* 18:1–16
28. Wahls WP (1998) Meiotic recombination hotspots: shaping the genome and insights into hypervariable minisatellite DNA change. *Current Topics in Developmental Biology* 37:37–75
29. Pratto F, Brick K, Khil P, Smagulova F, Petukhova GV, Camerini-Otero RD (2014) Recombination initiation maps of individual human genomes. *Science* 346:826
30. Wahls WP, Davidson MK (2011) DNA sequence-mediated, evolutionarily rapid redistribution of meiotic recombination hotspots. *Genetics* 189:685–694
31. Trifonov EN, Sussman JL (1980) The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proceedings of the National Academy of Sciences USA* 77:3816–3820
32. Trifonov EN (1998) 3-, 10.5-, 200-, and 400-base periodicities in genome sequences. *Physica A* 249:511–516
33. Schieg P, Herzel H (2004) Periodicities of 10-11 bp as indicators of the supercoiled state of genomic DNA. *Journal of Molecular Biology* 343:891–901

Part V
Conflict Between Genomes

Chapter 15

Self/Not-Self?

*Thrice is he armed that hath his quarrel just,
But four times he who got his blow in fust.*

Josh Billings, His Sayings (1866) [1]

Given the “struggle for existence” an organism might hope to outshine its competitors by virtue of the excellence of its positive characters (e.g. greater speed, longer neck). But an organism that could also (i) identify, defend against, and attack its foes, and (ii) identify and support its friends, would seem to have an even greater advantage. To identify is to allow discrimination – friend or foe? – the former classification being non-dangerous (potentially unharmed), the latter classification being dangerous (potentially harmful). For this, friend and foe must have distinguishable characters and the organism must both detect these characters, and know which characters are likely to associate with non-danger (friend) and which characters are likely to associate with danger (foe). After this discrimination, there must then be a response (either non-alarm, or alarm) that, in turn, would lead to a variety of other responses of an adaptive nature (e.g. escape from, or destroy, the foe).

All this is a rather long-winded way of saying that successful organisms are street-smart and look out for number one! However, there are semantic pitfalls that can easily confuse. For example, it is easy to believe that the initial act of character recognition is the same as danger recognition. But an organism does not initially recognize danger; it recognizes and classifies characteristics (attributes, signals) in a source organism, and on the basis of this discrimination draws conclusions regarding the potential harmfulness of that source.

When grazing sheep hear a sound in the wood they know it is not from one of themselves. So the sound has arisen from not-self. The first discrimination is a self/not-self decision. In one step the sheep detect the sound and register it as not-self. The not-self signal may alone suffice to make them move away from the wood as a precautionary measure. If the source of the sound is subsequently seen with the attribute “shepherd,” the sheep relax. If the attribute is “wolf,” then alarms sound (i.e. bleating and running). The initial *recognition* (detect and register as not-self or self) and *classification* (friend or foe) events are usually quiet. The subsequent alarm is noisy – an amplification of the initial quiet signal. The alarm itself will be independently recognized and responded to; but this is secondary, and does not require the specificity of the quiet primary recognition event. It is true that an individual deaf

sheep may respond only when it sees its companions running. So it is responding primarily to a general alarm signal. But the primary event in the sheep collective is the recognition of the not-self signal. The intuition to move away on this basis alone makes sense. Even if the source turns out to be the shepherd, one day they are going to be mutton, and the shepherd is going to have a hand in this!

An added complication, with which we are here most concerned, is that the foe may be within an organism's own body. Here, the characters recognized, and the sensors that do the recognizing, are likely to be chemical – receptors that should bind “not-self” molecular groupings, but should not bind “self” molecular groupings. Irrespective of its location, be it external or internal, the foe is part of the environment to which the organism has had to adapt through the generations. Likewise, the foe has had to counter-adapt.

Thus, the full understanding of the genome of any biological species requires an understanding of the genomes of the species with which it has coevolved. A virus, for example, is not going to just sit around and let itself be tagged as “foe.” It is going to exploit its power of rapid mutation to fool its host into believing it is “self.” Step by step, the virus will mutate to look more and more like its host, and if each step confers an advantage it will continue on this path. However, if as it continues it encounters progressively stiffer host defenses, then it will discontinue. Thus, hosts whose immunological forces are poised to attack “near-self” versions of not-self (a subset of not-self), rather than not-self *per se* (the entire set of not-self – formidable in range), should be at a selective advantage (see Chapter 16).

Homology Search

Human sexual reproduction is an example of a multistep discriminatory process, the sensitivity of which increases step by step, culminating in a homology search that is the final arbiter of friend or foe. First your mother, possibly discriminating among a number of suitors, chose your father. Then, one of his spermatozoa was recognized as “self” (a ‘friend’ to the extent that it was from a member of the same species) by receptors on the surface of your mother's egg (ovum). Accordingly, your father's spermatozoon was able to enter your mother's ovum to form a zygote. Next, through their gene products, your parental genomes attempted to work with each other to enable the single zygotic cell to multiply and develop into you, an adult. For this there had to be an absence of not-self discrimination (i.e. any one maternally-encoded product might have failed to cooperate with one or more paternally-encoded products, leading to developmental failure (hybrid inviability).

In the words of Samuel Butler [2]:

This power of cutting logical Gordian knots and fusing two contradictory or conflicting statements into a harmonious whole – is one without which no living being whether animal or plant could continue to live for a single day, or indeed could ever have come into existence. This holds good concerning all things that have a reproductive system, but it is most easily seen in the case of those forms that are reproduced by parents of different sexes.

The first thing which the male and female elements must do when they unite to form offspring is to fuse the conflicting memories and inconsistent renderings of their two parents into a single consistent story.

All this must have proceeded satisfactorily or you would not now be reading this. But you should not be complacent. As noted in Part III, depending on your sex and stage of life, the chances are that at this very moment in your gonad your parental genomes are repeatedly completing the initial act of conjugation that the parents themselves initiated decades earlier [3, 4]. The genomes appear to conjugate during meiosis, which results in the production of new gametes (Fig. 9.6). In the meiotic minuet your 22 maternally-derived autosomal chromosomes pair with your corresponding 22 homologous paternally-derived autosomal chromosomes, and 1 maternally-derived sex chromosome (X-chromosome) pairs with 1 paternally-derived sex chromosome (X if you are female, Y if you are male). Between each pair of homologs there is a search for similarity (“homology search”). A high or moderate degree of similarity of base sequence registers as “self” (i.e. friend) and the dance continues (the usual situation). However, below a certain threshold value for similarity, the base sequence registers as “not-self” (i.e. foe). The music stops and meiosis, and hence gamete production, fails. So, although unlikely, you could be sterile (a ‘mule’).

The latter sterility (hybrid sterility) should not be regarded as unnatural since, as related in Chapters 9 and 10, through variations in DNA sequences Nature confers sterility barriers on some species members; this allows branching into new species whose members, though natural selection, *can explore phenotypes that were not possible within the confines of the parent species*. We refer to species that have recently diverged as “allied,” implying a limited compatibility. But we are here concerned with conflict between species that cannot be so construed. Potential pathogens and their hosts are long diverged from their common ancestor, and hence are “foreign” to each other. However, so sensitive are immune systems that, even *within a species*, there is foreignness. Genetic identity is usually required for successful grafting between members of a species (allografting).

Antibody Response

The human embryo gains its nourishment from its mother by way of the placenta. But the embryo is, in some respects, a graft. It has a paternal component that is foreign to the mother. For this reason the evolution of placentation required that there be barriers preventing an immune response by the mother against the embryo. In theory, this would require that paternally-encoded immunogenic macromolecules not access the maternal immune system, and/or that maternal antibodies, which can bind those macromolecules, not be able to penetrate the placental barrier.

Initial formation of the zygote requires that the integrity of one of the mother’s cells, the ovum, be breached by a paternal cell (spermatozoon). The ovum is permissive for this invasion by a specialized foreign (not-self) cell. As long as it is derived

from a member of the same species it is seen as self. Should there subsequently be an intracellular self/not-self recognition event, the usual outcome would be that the paternal genome would be classified as friend rather than as foe. At this level of discrimination the paternal genome would be registered as likely to be unharmed.

However, the invading of female ova by foreign male spermatozoa of the *same* species is a very special case. Organisms and their cells become hosts to foreign organisms of *different* species (viruses and bacteria). Molecules, or parts of molecules, which can mark an organism as self or not-self are “antigens,” or antigenic “determinants,” or “epitopes.” The immune recognition events elicited by these can occur both inside and outside cells.

An organism is said to have undergone a specific antibody response when, within its extracellular body fluids, there is an increase in the concentration of protein antibodies able to combine specifically with eliciting antigens. Antibodies normally function outside cells and have an intracellular existence only within the cells of the lymphoid system (“B cell” lineage) where they are produced; they are rapidly secreted from these source cells and encounter antigens while remaining within the confines of the body of the source organism. This means that usually antibody molecules from one organism do not escape the body perimeter to locally assist its fellow organisms; an exception is milk (colostrum), which may transfer maternal antibodies to the newborn.

The antibody response to foreign (not-self) antigens has six consecutive steps. These six steps would appear fundamental to an internal biological defense system where there must be self/not-self discrimination and an adaptive response:

1. There is a randomization process whereby an organism acquires the potential to generate a wide repertoire of antibodies of varying specificities (i.e. with abilities to bind a range of antigens, *be they self or not-self*).
2. The potential to bind self-antigens is eliminated by the negative selection of self-reacting elements (i.e. the antibody repertoire is purged of self-reactivity, thus creating ‘holes’ in the repertoire).
3. The repertoire is molded to ‘dissuade’ pathogens from mutating to take advantage of these ‘holes’ (positive selection for “near-self;” see Chapter 16).
4. When the presence of a not-self antigen is detected by virtue of its reaction with specific antibody, then the concentration of that antibody increases.
5. The interaction between not-self antigen and specific antibody triggers alarms so that the host organism responds adaptively to eliminate the source of the not-self antigen.
6. The host organism learns from the encounter (i.e. acquires memory) so that, if challenged again by the same antigen, there is a better response (e.g. faster and quantitatively greater).

These are six steps fundamental to processes of adaptive self/not-self discrimination, which include processes for *intracellular* self/not-self discrimination. We should note that, just as the above-mentioned sheep in the meadow responded immediately to the self/not-self discrimination signal without the potential harmfulness of the source being determined, so a not-self antigen at the surface of a bacterium that

has been killed by some physical or chemical treatment, usually suffices to trigger an adaptive immune response when dead bacteria are injected into a host organism (steps 4-6). Detection and registration as not-self occur in one step because the repertoire has been pre-screened to avoid detection of self (steps 1-3). The response is not conditional on attribute recognition (i.e. classification as potentially harmful or non-harmful).

This trigger-happy tendency is fortunate, since it allows prophylactic immunization with attenuated or dead organisms, so conferring advanced protection against subsequent challenge with the live organism. Thus, our bodies do not allow bacteria to wander around like tourists. Even if likely to be non-harmful, bacteria, dead or alive, are recognized as not-self and responded to. Touring bacteria are not required to 'break windows' before the host organism will respond by making specific antibodies. On the other hand, trigger-happiness may result in self being mistaken for not-self (autoimmunity).

Antibody Variable Genes

In principle, the ability to generate, purge, and mold, elements of an immunological repertoire (steps 1-3) could occur, together or independently, either during the life of an organism (i.e. during somatic time), or over multiple generations (i.e. during evolutionary time). In the latter case the organism would be born with an appropriately purged and molded repertoire (i.e. the repertoire would be *innate*, having been inherited through the germ line). In the case of the antibody response there is an innate repertoire of variable region genes, but the generation of diverse antibody molecules from individual variable region genes involves extensive further variations, so that essentially steps 1-3 all occur somatically.

The necessity for this is apparent from the fact that two diploid parents are likely to have independent germ-line histories, so that within their diploid child a paternally-donated repertoire would not have been purged of antibodies with the potential to react with maternal antigens, and a maternally-donated repertoire would not have been purged of antibodies with the potential to react with paternal antigens. This problem is eliminated if the purging occurs during somatic time. "*Self*" is *defined afresh* when a new organism is produced from two parents of different genetic constitutions. An additional advantage is that each individual is likely to have a *unique repertoire*, so that a pathogen that somehow evaded the repertoire of one host could not count on an equally easy ride in its next host.

Generation of an antibody repertoire during somatic time provides an opportunity to develop antibodies of high specificity for provoking antigens. The chemical bonding between antigen and specific antibody is strong. Thus the antibody repertoire can encompass both range (variety of different antibodies) and high specificity (strong bonding). On the other hand, an innate repertoire would be limited by genome size. To detect not-self there might be range, but the trade-off might be less specificity. Bonding between antigen and antibody might be weak.

Perhaps this is one reason why twentieth century immunologists did not readily transferred their thinking about antibodies, essentially an extracellular phenomenon of multicellular organisms, to the idea of some intracellular antibody equivalent (see below).

A parallel can be drawn in terms of friends and acquaintances. Most of us have a few ‘friends’ with whom we are strongly bonded, and many ‘acquaintances’ with whom we weakly bond. Remarkably, quite often in the grand scheme of things it is the acquaintances that make the difference. Mark Granovetter’s study of social networks – *The Strength of Weak Ties* – argues this well [5]. We shall consider here the possibility that, in an intracellular environment, weak ties can provide a sufficient degree of specificity for self/not-self discrimination. The ‘crowded cytosol’ (see Chapter 16) greatly supports this process.

Prototypic Immune Systems

If some evolutionary adaptation is highly advantageous, then the ‘hand of Nature’ is likely to have explored possible ways to achieve it. For this, organisms must be in possession of any necessary *facilitatory* adaptations, and there must be no *countermanding* adaptations. Thus, if not countermanded by weight, wings can be highly advantageous. Necessary facilitatory adaptations are appendages. To evolve wings an appendage must ‘find’ how to trap air effectively when moving in one direction, but not when returning to its original position by moving in the opposite direction. Birds, bats and flying insects have solved this problem in different ways. The underlying adaptive principle is the same.

In this light, when celebrating the advantages of adaptive immune systems in the higher vertebrates, we should be aware that, no less than the membranous wings of insects differ from feathers, organisms perceived as low on the evolutionary scale might have immune adaptations that differ from the adaptive immune systems we are familiar with. For example, jawless vertebrates such as the lamprey have no immunoglobulin antibodies, but they have developed an analogous immune system, as have snails [6, 7]. How far down the evolutionary scale can we go before the presence of a countermanding adaptation, or absence of some facilitatory adaptation, becomes prohibitory?

Unicellular organisms are likely to have evolved some 800 million years before multicellular organisms. In 1861 the German physiologist Ernst Brücke dubbed single cells “elementary organisms” implying that many multicellular level functions might have prototypic equivalents at the unicellular level [8]. For example, the functional separation of germ cells and somatic cells, resulting in what Weismann called “the continuity of the germ-plasm” (Chapter 1), was originally construed as a phenomenon of multicellular organisms. Certainly, there is no sign of the separation in the unicellular amoeba. To reproduce itself, the amoeba enters mitosis and evenly divides the contents of its solitary nucleus between the two child cells. However, in another group of protozoa, the ciliates, there may be two nuclei one of

which can be construed as a ‘working nucleus’ and the other as a non-working ‘archival nucleus.’ When so inclined, ciliates conjugate sexually and in this process working nuclei (“soma”) are destroyed while the archival nuclei reciprocally exchange haploid products (“germ line”), and the regenerated archival nuclei then form new working nuclei.

Give such sophistication, it can hardly be doubted that, over time, some nucleic acid sequences would have ‘learned’ how to break away from one cell and enter another (“horizontal transmission”). So unicellular organisms are likely to have been challenged by many such ‘viruses’ and to have had plenty of time to adapt to this challenge. Indeed, it is likely that immune systems of multicellular organisms arose as *extensions* of systems pre-existing at the unicellular level. In broad outline, the probable characteristics of such prototypic systems can be deduced, and this might help us understand modern systems.

In a clonal unicellular population where asexual reproduction predominated, early self-destruction (i.e. apoptosis; see Chapter 16) would be a simple mechanism to prevent spread of a virus from host to host. This would diminish the opportunity for virus multiplication within a host cell, so that fewer copies of the virus would be available to infect other cells. Since members of the host population would be genetically identical, this altruistic act by one individual might allow more copies of its genes to survive than if it were not altruistic (i.e. if it struggled to cope with the virus itself, and did not self-destruct). Thus natural selection would favor organisms with genes for altruism (‘selfish genes’).

However, even such a primitive defense would require specific antibody-like sensors and discrimination between self and not-self. If released from the cell these sensors would probably be lost by diffusion and hence would not be available to help either the cell of their origin or others in the vicinity. Thus, it is unlikely that a system for the secretion of antibody-like molecules would have evolved prior to the appearance of multicellular organisms with an extracellular space within a perimeter (skin) preventing loss by diffusion. It is more likely that, if it were at all feasible for it to evolve, a system for intracellular self/not-self discrimination would have evolved *first*. Such a sensory system could consist of the multiplicity of structurally distinct macromolecules that are *normally* present within cells, of which we emphasize here two populations – ‘immune receptor’ proteins and ‘immune receptor’ RNAs (Fig. 15.1).

Many of these macromolecules would have distinct properties necessary for cell function (catalytic, structural, transporting, templating, etc.). Because of these properties both populations would be variable in type and number. For example, multiple heterogeneous mRNA species each encode a different protein product. On observing such variability we are not surprised, since there is an obvious explanation – small mRNAs encode small proteins, and large large. This tends to preempt the search for alternative explanations. However, if we consider the possibility that molecular variability might serve more than one purpose, then we can regard the variable *intracellular* mRNA and protein populations in the same way as we regard the variable *extracellular* antibody (immunoglobulin) populations. The latter act as an internal barrier against foreign pathogens (“not-self”), which ‘select’ from among the antibodies (i.e. are recognized by) those with specificity for their coat antigens.

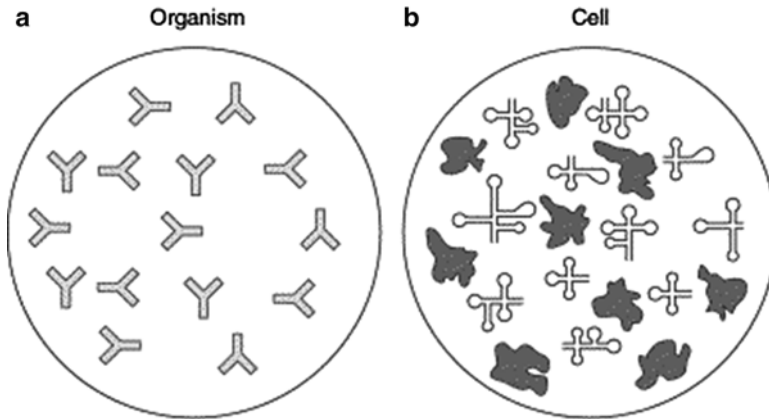


Fig. 15.1 The cell as an elementary immune organism. At left (a) represents a multicellular organism with Y-shaped antibodies of various specificities. Each antibody has a variable region at the tips of the Y-arms (not shown), and a constant region (the rest of the Y-arms and the Y-stem). At right (b) represents a unicellular organism with a repertoire of ‘antibody-like’ protein molecules (gray) and ‘antibody-like’ RNA molecules (stem-loop structures). These could be referred to as “immune receptors” implying that parts of the molecules have the potential to interact with intracellular ‘antigens’

In the case of a virus, coat proteins would be relinquished at the time of entry into a host cell, and the cell’s first line of intracellular defense – its first intracellular barrier – would be to recognize the foreign nucleic acid as “not-self,” either in its genomic form, or in the form of an early viral RNA transcript. This might implicate double stranded RNA (dsRNA). Perhaps a segment of a particular host RNA species, because it happened to have sufficient sequence complementarity, would form a duplex with viral RNA of sufficient length (at least two helical turns) to trigger an alarm response (Fig. 6.8). Thus, the variable host RNA population can be viewed as consisting of an innate repertoire of ‘RNA antibodies,’ each preselected over evolutionary time not to react with “self” RNAs, while retaining, and if possible developing, the potential to react with “not-self” RNAs [9].

There is, then, the possibility that one or more *resident* intracellular molecules would be able to bind molecules from an invading virus with sufficient affinity to tag them as “not-self,” thus initiating an adaptive intracellular ‘immune response.’ In theory, this diverse ‘immunological repertoire’ could develop either over evolutionary time, or over somatic time. In a unicellular organism, somatic time can be considered as being the period between successive cell divisions, which makes development over evolutionary time far more likely. Whatever the mechanism and timing of the diversification process, there would be a need to eliminate receptors with high affinity for self-antigens. In the case of a clonal population, these intracellular self-antigens should be common to all members of the population.

Unfortunately, given the high replication and mutation rates of viruses relative to those of their hosts, it would seem highly probable that viruses would preadapt to avoid interaction with hostile host macromolecules. What a virus had ‘learned’

(by mutation and selective proliferation) in one host, it would exploit on the next host. However, a high degree of host genome variability (polymorphism) would thwart the virus (see below). Thus, host clonality would be a hazard. Unicellular species that could diversify their genomes (i.e. differentiate and decrease clonality), while retaining their basic phenotypic identities, would be at a selective advantage.

In an elementary unicellular immune system, viruses that, through mutation, acquired the ability to inactivate host apoptotic mechanisms, would preferentially survive (by preventing the ‘altruistic’ self-destruction of their host cell). In the course of the ensuing arms race (Chapter 13), an intracellular ‘inflammatory’ host response would have evolved to limit viral activities. But in multicellular organisms, death (apoptosis) of the primarily infected cell might limit the opportunity to alert other target cells and cells of the immune system (the lymphoid system). This would prevent the initiation of the extracellular host inflammatory response. Thus, further sophistications, including the delaying or prevention of apoptosis, would be expected to have evolved in multicellular types.

Polymorphism Creates Host Unpredictability

On average, the haploid maternal and paternal contributions to your diploid genome are likely to differ from each other at least once every 0.5-2.0 kilobases, and general intraspecies differences are even more frequent. Such variations, referred to as “polymorphism,” should decrease the extent to which a pathogen from one host could anticipate the genomic characteristics of its next host. “Polymorphism” means literally “many forms” (Greek: *poly* = many; *morph* = form or shape), and the word was used initially to describe anatomical differences among members of a species – e.g. shades of skin color. However, polymorphism may be present at the genomic level without any correlated anatomical changes (i.e. no change in the conventional phenotype). When polymorphism affects proteins, it is likely to involve relative low complexity sequences corresponding to surface protein domains (usually hydrophilic and non-globular; Fig. 15.2).

Thus, these surface domains, often not critical for the specialized function of a protein, would be available for interaction with complementary molecular patterns presented by intracellular pathogens (“not-self”). But these same domains should also have the potential to react with other “self” proteins, sometimes to an extent sufficient to trigger adverse responses in the host (intracellular ‘autoimmune’ pathology). Accordingly, over evolutionary time natural selection would have favored organisms with mutations avoiding this. The intracellular ‘antibody’ repertoire would have been generated, purged and molded (i.e. fine-tuned) over evolutionary time. But would a single cell be capable of generating an adequate repertoire? In seeking an answer we should consider not just genic information, but the total information in genomes, both genic and non-genic.

So many genomic phenomena can be explained in terms of conflicts between different forms of information, that severe doubt is cast on the notion that there is

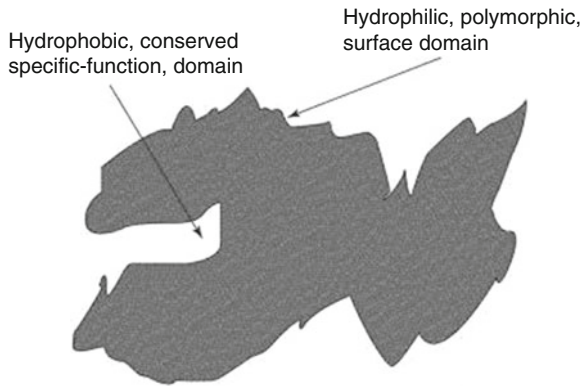


Fig. 15.2 Specific and general functions of a protein as reflected in its structure. Dedicated functions ('primary') are often associated with conserved, internal, hydrophobic, globular domains. Potential immune receptor functions ('secondary') would be associated with variable (polymorphic), external, hydrophilic, non-globular domains

'room for all' in the journey through the generations. DNA sequences, be they primarily genic or non-genic, contain many, often overlapping, informational messages (Chapter 12). Genomic channel capacity appears finite [10, 11]. We must pause when tempted to dismiss any part of it as "junk," even when confronted with organisms that appear normal after megabase deletions [12], or that seem to have far, far, more DNA than appears necessary in those that appear no less complex [13].

Junk DNA

As mentioned in Chapter 1, human beings have a special way of dealing with things that they do not understand. They invent a word for it. They then sometimes confuse word with explanation. For a long time physicians used "psychosomatic" when there was no available explanation for sets of symptoms that did not match a particular disease. Sometimes it was, indeed, 'all in the mind,' but over the years more and more so-called psychosomatic diseases have been found to have an organic basis. Likewise, for a long time changes in DNA sequences that did not appear of adaptive value were dismissed as "neutral" and ascribed to "mutational biases" (Chapter 7).

The uncomfortable fact that around 98% of our DNA appeared non-genic was blunted by a little four letter word – "junk" [14]. We appear to be the products of our genes alone, so surely the rest *must* be non-functional junk? Some organisms, such as the puffer fish, seem to manage with very little non-genic DNA, which therefore cannot be important. On the other hand, the puffer fish has retained *some* non-genic DNA, so perhaps there is an aspect of its biology that allows it to make do with less? Birds seem under a pressure to rid their genomes of superfluous DNA as shown by the reduction in the size of large introns (Fig. 2.5). So perhaps non-genic DNA and

introns are both examples of junk DNA? Perhaps some junk DNA plays a regulatory role? But would 98% of our DNA be needed for this purpose?

What other functions could junk DNA have? After all, junk (stuff in your attic that might be useful some day) is not the same as garbage (stuff you want to get rid of). Are there circumstances under which a piece of junk could become useful? When considering usefulness it is very easy, and often correct, to assume that that which is useful has been preserved by natural selection. This can lead to the argument, again often correct, that sequence features that are *conserved* between two species must have been useful. Are there circumstances where the converse might apply? Under what circumstances could sequence features that are *not conserved* between two species (i.e. features that vary) have been useful?

To be functional it is likely that non-genic DNA would have to be transcribed into RNA. Indeed it is. Investigations of the transcriptional activities of entire human chromosomes revealed a “hidden transcriptome,” corresponding to a large number of low copy number cytoplasmic RNAs. It was estimated that there is “an order of magnitude” more transcriptionally active DNA than can be accounted for by conventional genes [15–17]. Independent identification of a large number of “CpG islands” supports this (see Chapter 18).

Could these low copy number RNAs be dismissed as mere cytoplasmic ‘junk’ – an unavoidable consequence of the existence of genomic ‘junk’? To understand their role in the economy of the organism, we need to know, by analogy with known transcriptional processes, whether there are specific promoters, whether there are dedicated RNA polymerases, whether transcription occurs randomly or under specific conditions (e.g. tissue specificity), whether transcripts are diverse and include appreciable non-repetitive DNA, and whether they are translated into proteins, or function at the RNA level.

Repetitive Elements and “Selfish Genes”

Much non-genic DNA consists of repetitive elements (Chapters 2, 14). The most prominent of these in humans are the 1,090,000 *Alu* elements. Three of these, each about 300 nucleotides in length, are shown in Figure 15.3 to which we will return in Chapter 18. It so happened that the discovery of *Alu* elements came at about the same time as the re-emergence of the selfish gene concept. The temptation to explain away junk DNA as due to the ability of these parasitic ‘selfish genes’ to colonize our genomes then became irresistible.

The power of the selfish gene concept was its ability to provide an explanation for certain high social and moral qualities (altruism). Darwin had pondered [18]:

How within the limits of the same tribe did a large number of members first become endowed with these social and moral qualities, and how was the standard of excellence raised? ... He who was ready to sacrifice his life, as many a savage has been, rather than betray his comrades, would often leave no offspring to inherit his noble nature ... Therefore it seems scarcely possible ... that the number of men gifted with such virtues, or that the standard of their excellence, would be increased through natural selection, that is, by the survival of the fittest.

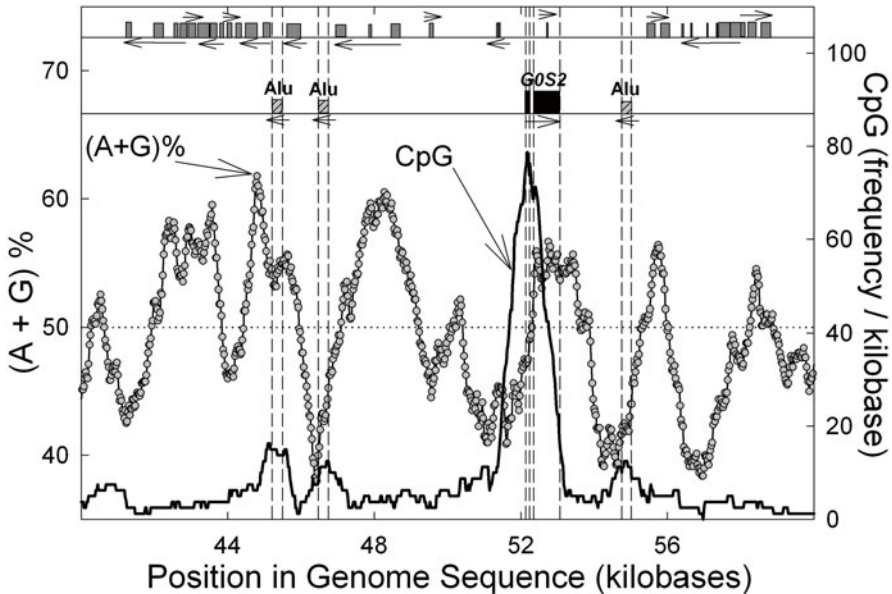


Fig. 15.3 *Alu* and other repetitive elements in a segment of human chromosome 1 containing the two exon gene, *GOS2*. Horizontal arrows indicate transcription directions. *GOS2* and three *Alu* elements are represented on the lower horizontal line by black and gray (striped) boxes, respectively. Their boundaries are demarcated by vertical dashed lines. Other repetitive elements (some of retroviral origin) are represented by gray boxes on the upper horizontal line. (A+G)% and CpG frequency (dinucleotides/kb) were evaluated for 400 base windows moving in steps of 25 bases. CpG peaks (macro or mini) are associated with *GOS2* and the *Alu* elements (see Chapter 18). Note that Szybalski's transcription direction rule is obeyed by the *rightward*-transcribing *GOS2* gene ($R > 50\%$ in top strand) and two of the three *leftward*-transcribing *Alu* elements ($R < 50\%$ in top strand)

A tribe consists of groups of people that are likely to be related, and so share common genes. A gene, or collection of genes, that, by some chain of cerebral events, encouraged altruistic behavior, would increase within the tribe if, by sacrificing himself the “savage” had caused many of his relatives (and hence their genes for altruism) to survive. Members of that group within the tribe could increase in numbers, relative to groups whose members did not exhibit altruistic behavior. Thus, selfish individuals would not survive, but selfish genes (genes governing altruistic behavior) would. This brought to light the possibility that, in some circumstances, the focus of natural selection might be at the levels of the *gene* and the *group*, rather than at the level of the *individual* (Chapter 12). However, in the present case there is perhaps a better argument than the selfish-gene argument.

Both conventional genes and repetitive elements provide promoters for the transcription of non-genic DNA. Some gene transcripts have been found longer than expected due to a failure of transcriptional termination (‘run-on’ transcription). Some classes of repetitive element contain promoters from which transcription can initiate and extend beyond the bounds of the element into neighboring genomic regions (Fig. 15.4).

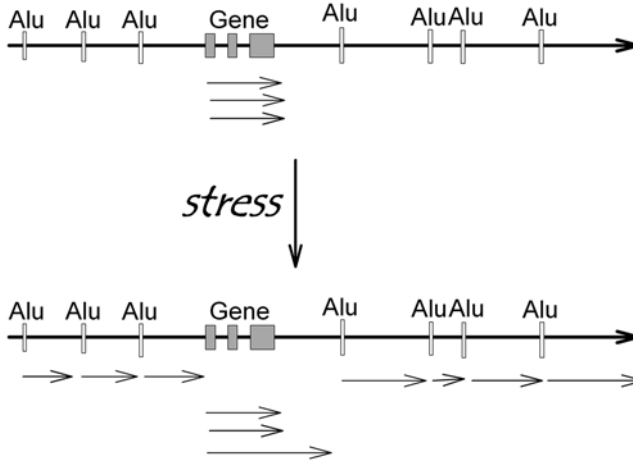


Fig. 15.4 Run on transcription from genes and repetitive elements in cells subjected to the ‘stress’ of virus infection. The direction and extent of transcription is indicated by the thin horizontal arrows. There is evidence that stress may indeed increase run-on transcription of *Alu* elements. Thus, careful controls must eliminate inadvertent stress during RNA preparation before concluding that pervasive transcription from the “hidden transcriptome” is not a general stress-induced phenomenon

Are such extended transcripts generated randomly in time? In the case of *Alu* elements, transcription (by RNA polymerase III) increases at times of cell stress (e.g. viral infection, heat-shock). Indeed, viral infection can simulate the “heat shock response” where there is induction of a set of proteins known as the “heat shock proteins” (see Chapter 16). Thus, *Alu* transcription might reflect an adaptive response to virus infection, and natural selection could have favored the spread of *Alu* and its like, for this purpose [19].

All Hands on Deck!

While some crew members are on watch, others rest. But in an emergency it is “all hands on deck!” Rest must be deferred. The gunnery officer and galley cook must hasten to shorten sail rather than attend to their ‘primary’ scheduled tasks. Likewise, a surge of ‘promiscuous’ transcription due to the stress of infection could reflect the ‘awakening’ of many genes that would not normally be expressed *at that time and place*.

For example, rubella virus infection (German measles) during early pregnancy could launch unscheduled transcription by the embryo of ‘RNA antibodies,’ some of which, unlike the gunnery officer and cook, might try to attend to their usual (primary) tasks (e.g. be translated into functional proteins). The inadvertent expression in an embryo at a critical developmental stage of *minute* quantities of the product of a key regulatory gene could suffice to generate a deformed (mutant) child

(a complication of rubella in pregnancy). But it would be a “phenocopy” mutation – the child’s germ cells would not have been affected so, if able to become an adult and reproduce, its children would be normal (see Chapter 16).

A candidate gene for the mediation of such ‘promiscuous’ transcription was found among patients suffering from a rare form of autoimmune disease [20]. In this case the disease was due to the lack of expression of the “autoimmune regulator” (AIRE) protein in the thymus where cells of the lymphoid system (“T cell” lineage) ‘audition’ (steps 1-3 above). When the AIRE protein is functional there is expression of a subset of mRNAs, which are translated into a corresponding set of self-proteins that are normally synthesized *in tissues other than the thymus*. Maturing T cells that have acquired some specificity in responding to antigen (step 1) can then be checked against self (steps 2 and 3). Only the successful auditioners are released from the thymus. If this check fails to occur, autoimmune disease may result.

The specialized thymus cells in which AIRE causes ‘promiscuous’ expression of potential self antigens are short-lived. That which is promiscuously expressed has *little opportunity* to manifest its normal tissue-specific role. But, although initially found in the thymus, the AIRE protein is also present in other tissues, so unscheduled expression of some self RNA and self proteins during rubella infection in pregnancy appears feasible [21].

The Double-Stranded RNA Alarm

Although protein molecules can recognize specific nucleic acids (and the converse), we consider here host proteins as recognizing foreign proteins and host RNAs as recognizing foreign RNAs. In the cytosol RNA molecules adopt characteristic stem-loop configurations (Figs. 5.1, 15.1b), and RNA-RNA interactions must initiate by way of a ‘kissing’ search between bases at the tips of loops. If sequence complementary is found, then two RNA species can pair, partially or completely, to generate a dsRNA (Fig. 6.8).

If a virus introduced its own RNA into a cell, would the variability among host RNA species suffice for the reliable existence of a specific host ‘immune receptor’ RNA capable of forming a segment of dsRNA with the “not-self” RNA of the virus? Calculations show this to be feasible, especially if the *entire* host genome were available for transcription [22]. Would the dsRNA be able to initiate an adaptive intracellular ‘inflammatory’ response? How would the cell prevent two “self” RNAs interacting to form dsRNA?

Formation of dsRNA has long been recognized as providing an early cellular alarm response to viral entry. Protein synthesis can be inhibited by very low concentrations of dsRNA. This involves activation of a kinase (an enzyme that adds a phosphate group to proteins) known as “dsRNA-dependent protein kinase” (PKR), and results in inhibition of the initiation of protein synthesis. This is bad news for an invading virus because it wants to synthesize viral proteins, but it is also bad news for the host cell, since the inhibition is not virus-specific, and its ability to synthesize

its own proteins is impaired. Evasive viral strategies would include the acceptance of mutations to avoid formation of dsRNA, and direct inhibition of cell components required for the formation of, or the response to, dsRNA.

Virus-infected cells produce “interferons” – proteins that can be considered part of the inflammatory response. These induce a general anti-viral state spreading together with various other protein molecules (e.g. “chemokines”) from the cell of origin to other cells. Their production was long known to be powerfully stimulated by dsRNA, for which no clear explanation was forthcoming. There is now much evidence that, both in animals and plants, another more sequence-specific ‘inflammatory’ response to dsRNA arises as part of an intracellular mechanism for self/not-self discrimination (Fig. 15.5).

Just as in the adaptive antibody protein response there is increased production of specific antibody, so, courtesy of enzymes with names such as “RNA-dependent-RNA polymerase,” “dicer” and “slicer,” there is increased production of specific ‘immune receptor’ RNAs (“short interfering RNAs”; siRNAs) as part of the specific adaptive response to a virus. In association with an “RNA-induced silencing complex” (RISC), they mediate sequence-specific strand cleavage of viral nucleic acid. Organisms with mutations in components of this system show increased vulnerability to pathogens [23].

Purine-Loading to Self-Discriminate

Given the acknowledged role of dsRNA as an intracellular alarm signal, it seemed that viruses, needing to multiply within a host cell, were in some way *obliged* to form dsRNA, possibly as a replicative intermediate. Perhaps viruses, for some fundamental reason, had not been able to avoid formation of dsRNA replicative intermediates, and their hosts took advantage of this. However, viruses with dsRNA genomes have adaptations that appear to conceal their genomes from host cell surveillance mechanisms [24]. So, even if there were an obligatory need for a double-stranded replicative intermediate, it could be concealed.

Thus, the hypothesis that dsRNA would be formed by an interaction between host RNA and viral RNA is quite plausible. This requires that the host itself not form dsRNA by interactions within or between its *own* RNA molecules (see also Chapter 18). The phenomenon of purine-loading provides a possible mechanism for avoiding the formation of ‘self-self’ RNA duplexes.

More than twenty contiguous base pairs in a duplex are needed to initiate dsRNA-dependent events (e.g. activation of PKR). But after their transcription and processing, among the RNA species of a cell (“self” RNAs) there might be two whose members, by chance, happened to have enough base complementarity *with each other* for formation of a mutual duplex of a length sufficient to trigger alarms. So there would need to have been an evolutionary selection pressure favoring mutations in RNAs that decreased the possibility of their interaction with other self-RNAs *in the same cell*, a process that might otherwise lead to the negative selection of the host organism.

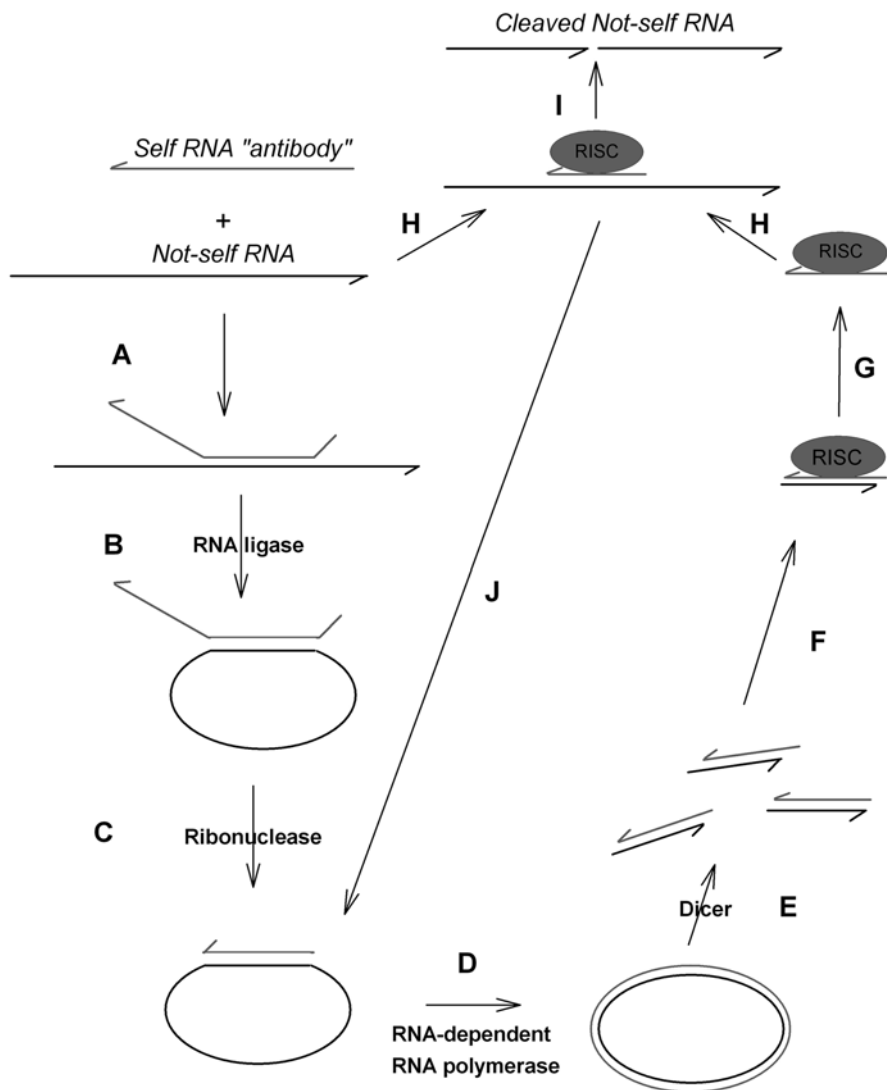


Fig. 15.5 Model for an adaptive intracellular self/not-self discriminatory immune system operating by "RNA interference" (RNAi). Foreign "not-self" RNA (black line) interacts (A) with a member of the diverse 'RNA-antibody' (immune receptor) repertoire (gray line) to produce a small, but sufficient (i.e. two helical turns) length of dsRNA. The foreign not-self RNA is circularised by an RNA ligase (B), but a single strand-specific ribonuclease (C), associated with immune receptor RNA, begins to degrade the latter RNA so preventing its circularisation. RNA-dependent RNA polymerase (D) uses the paired immune receptor RNA fragment to prime the synthesis of a longer dsRNA that, when complete, corresponds to the entire not-self RNA. Dicer (E), an enzyme (endonuclease) specific for double-stranded RNA, cuts the dsRNA into multiple small (two helical turns) dsRNA fragments. The small dsRNAs are individually bound (F) to a RISC protein complex (RNA-induced silencing complex), which (G) retains only the strand derived from the original immune receptor RNA (gray line). By virtue of the complementarity between the immune receptor derived RNA (gray line) and the not-self RNA species (black line), the RISC recognition complex interacts with *another* member of the not-self RNA species (H). The process may then recycle (J) through steps D-H to amplify the number of RISC recognition complexes. Alternatively, an endonuclease (I) in the RISC recognition complex ("slicer") cleaves and inactivates the not-self RNA, which will be further degraded by single strand-specific exonucleases (ribonucleases). This area of research is moving very rapidly and the figure is merely illustrative of the principles involved

In many cases mutations to a purine would assist this, since purines do not pair with purines. Indeed, interaction with self-RNAs seems to have been avoided by purine-loading the loop regions of these RNAs, thus avoiding the initial loop-loop ‘kissing’ reactions that precede more complete formation of dsRNA. An excess of purines, observed both at RNA and at DNA levels (in RNA-synonymous DNA strands), is found in a wide variety of organisms and their viruses (Chapters 6, 12).

Viruses that purine-loaded their genomes to match their hosts (e.g. the AIDS virus, HIV1) would seem to have adopted a sound strategy (from a viral perspective). Viruses that foolishly pyrimidine-loaded would generate pyrimidine-rich mRNAs that should readily base-pair with purine-rich host RNAs to generate corresponding dsRNAs, hence activating intracellular alarms. Apparently foolish viruses are HTLV1 (Chapter 10) and EBV (Chapter 14). However, it should be noted that these two viruses are *profoundly latent* (quiescent) in their hosts. Major transcription occurs only under provocation. HIV1 and HTLV1 (retroviruses) both have the latency option, but HTLV1 is more deeply latent, most of its human hosts living normal lives.

HTLV1 is localized to certain populations, and is particularly prevalent in Japan. Like HIV1, it inserts ‘seamlessly’ (collinearly) into the genome of its hosts. In contrast, EBV exists as an independent genomic element (“episome”). To regulate its relationship with the host cell, it transcribes an mRNA encoding the EBNA1 protein. In contrast to most other EBV genes, *this gene is purine-loaded*. Indeed, it appears to have acquired a long, purine-rich, simple sequence repetitive element solely to facilitate purine-loading (Fig. 14.2). Here there is an intriguing similarity with HTLV1 (see later).

Entropy

Purine-loading is especially high in organisms that live at high temperatures (thermophiles; Chapter 11). For this there is an “entropic” explanation. Unlike the word “junk,” an intellectual crutch that long assured the credulous that 98% of our genomes could safely be ignored, the word “entropy” brings discomfort. Physicists and chemists include entropy among the forces that can drive chemical reactions, but biologists tend not to find the concept so intuitive. Max Lauffer’s seminal *Entropy-Driven Processes in Biology* has largely been ignored [25]. All that is asked of the present reader is that he/she accept the fact that entropy can drive molecular interactions, irrespective of any understanding of it as reflecting a universal tendency for ordered systems to break down and become disordered (chaotic).

For example, when you leave a recalcitrant burned saucepan ‘in soak’ overnight before cleaning, you are letting entropy work for you. The burned food is ‘ordered’ to the extent that it is so firmly attached to the bottom of the saucepan that it cannot easily be removed manually. A night’s bombardment with water molecules serves to loosen the attachment releasing the burned food, which thus becomes ‘disordered.’ The work you have to do the following morning is less to the extent that

entropy has been at work. If you had left the saucepan overnight in hot water you would have been spared even more work. The entropy-driven component of a reaction *increases* as temperature increases. Given that increase in body temperature in infections (pyrexia) is of high adaptive value [26], entropy-driven reactions may provide a basis for this adaptation (see Chapter 16).

Exploratory ‘kissing’ interactions between hybridizing single-stranded nucleic acids involve transient base stacking interactions with the exclusion of water molecules that have become ordered around the free bases exposed when nucleic acids are in open single-stranded form. Such reactions have a strong entropy-driven energy component. Although two separate nucleic acid strands become united as a duplex and so are collectively *more* ordered, this is accompanied by the release of *many* water molecules that become collectively *less* ordered. Overall there is an increase in disorder, and this drives duplex formation.

It seems intuitive that an increase in temperature would increase the internal vibrations and velocities of chemical molecules, so that two interacting molecular species would need to be held together tenaciously at high temperatures by classical bonding forces (ionic bonds, hydrogen bonds and Van der Waal bonds). Thus, DNA duplexes can be ‘melted’ by heating to produce single strands. So it would seem that if you wished molecules to aggregate and stay together you should lower the temperature. However, the release of water molecules that are ordered around macromolecules in solution becomes greater as temperature increases. Accordingly, the entropy-driven component of chemical reactions *increases* as temperature increases. This can be greater than the pressures to break classical bonds.

So, instead of being dispersed by increasing temperature, the aggregation of macromolecules, such as single-stranded nucleic acids forming double-strand duplexes, is *favoured*. The propensity to aggregate increases until the melting temperature is reached, when the strands suddenly separate. The released open single strands immediately adopt stem-loop conformations that are strongly favored entropically at such high temperatures. It follows that the need to avoid ‘kissing’ interactions between single-stranded “self” RNA molecules progressively *increases* as temperature increases.

The physico-chemical state of the ‘crowded’ cytosol (see Chapter 16) is likely to be highly supportive of reactions with a high entropy-driven component, which include those between the codons of mRNAs and anticodons in tRNA loops. This particular interaction results in transient formation of dsRNA segments that normally do not exceed five base pairs (Chapter 7). So, protein synthesis, a process of *fundamental* importance for cell function, is optimized. But a cytosolic environment that favors mRNA-tRNA ‘kissing’ interactions should *also* favor mRNA-mRNA ‘kissing’ interactions. This could act as a ‘distraction’ impeding mRNA translation, and hence impeding protein synthesis. By generally militating against RNA-RNA interactions, purine-loading should prevent mRNA-mRNA interactions and so free mRNA molecules for interaction with the cytoplasmic components necessary for their translation (e.g. ribosomes and tRNAs).

Thus, purine-loading should facilitate protein synthesis. Because of this, organisms that did not purine-load their RNAs would be negatively selected. The early

evolution of such mechanisms to *avoid* inadvertent formation of segments of dsRNAs as a result of interactions between self-RNAs, driven by the need for efficient protein synthesis, would have created an opportunity for the subsequent evolution of mechanisms *utilizing* dsRNA as an intracellular alarm against not-self-RNAs (“RNA interference”; RNAi).

In some cases, so important has been the need to accept purine mutations, that even the protein-encoding role of genes appears to have been compromised (Chapter 12). When this compromise is not possible, then the length of the protein can be increased by inserting simple sequence repeats of amino acids with purine-rich codons between functional domains, as in the case of EBV and malaria parasites (Chapter 14). Since the extra purines often occupy third codon positions, it is possible experimentally to *decrease* purine loading without affecting the nature of the encoded protein. For EBV this *enhances* the expression of fragments of the EBNA1 protein by the host cell and triggers a T-cell response (Chapter 16). Thus, purine-loading the EBNA1 gene serves to protect EBV against immune attack [27].

In Australia, Murat and Tellem [28] proposed that this protection is due to an excess of only one of the purines in the EBNA1 gene. Sequential Gs would form a higher ordered structure in mRNA (G-quadruplex) that would slow protein synthesis. However, this does not explain why purine-loading also involves As. Intriguingly, it has emerged that another virus, HTLV1, which like EBV is both profoundly latent and GC-rich, synthesizes a protein (HBZ) that plays a role analogous to that of EBNA1 in EBV. HTLV1 predominantly transcribes RNA from one (‘top’) strand (that happens, unlike HIV1, to be *pyrimidine*-rich; Chapter 10). However, the mRNA for the HBZ protein is quite distinctive, being the *sole* HTLV1 transcript corresponding to the “*bottom*,” purine-rich, strand [29, 30]. It is envisaged that both this mRNA and EBNA1 mRNA keep their viral genomes below the radar screen of host immune defences by adopting the R-loading strategy [31].

The Hidden Transcriptome

The six fundamental steps in the extracellular antibody response have at least four analogs in the proposed RNA immunoreceptor system (Fig. 15.5). Host repertoire generation (step 1) occurs over evolutionary time. Purine-loading militates against reaction with self (step 2). Double-stranded RNA is formed when viral RNA becomes base-paired with a sufficiently long segment of host ‘immunoreceptor’ RNA, and there are enzymes that can amplify the concentrations of the dsRNA complexes (step 4). Double stranded RNA acts as an alarm signal triggering adaptive responses (step 5). A further similarity is that RNAi is not confined to its cell of origin. Although the target is intracellular, at least in plants and insects RNAi is known to transfer extracellularly to other cells [32], probably in double-stranded form [33].

Thus, many disparate observations appear comprehensible in the context of an intracellular RNA immunoreceptor system for self/not-self recognition. Although generation of variability by post-transcriptional RNA editing is feasible (Chapter 5)

[34], in essence the maximum potential repertoire of ‘RNA antibodies’ is limited by genome size. This would explain the existence of a “hidden transcriptome” [15–17]. Repetitive elements (e.g. *Alu* sequences) which change rapidly in evolutionary time, and mobile genetic elements (transposons), can be viewed as germ-line variation-generating devices that create polymorphisms so making it difficult for viruses to anticipate the genomic characteristics of future hosts. This suggests that regions transcribed as the “hidden transcriptome” should *not* be evolutionarily conserved. A comparison of libraries of DNA copies of cellular RNAs (“cDNAs”) made from humans and mice supports this [35]:

It is interesting to note this [human] type of ‘non-coding’ transcript was also found in mouse cDNA collections. ... What was significant was that [the] majority of the examined cDNAs were not evolutionally conserved. In this dataset of mouse genes, identification of 11665 similar transcripts (which would be categorized as ‘unclassified’ according to our scheme) has also been reported. This suggests that there is little conservation for these ‘unclassified’ transcripts and/or that there are huge numbers of such transcripts (at least in the order of 100000). ... We have recently examined the promoter activities of randomly isolated genomic DNA fragments on a large scale and observed that there are cryptic promoter activities throughout the genomic DNA (unpublished data). It may be possible that those cryptic promoters may act at low frequency to produce aberrant (or sporadic) transcripts.

To what extent are there analogs of the third (positive selection) and sixth (memory) fundamental steps? One way of creating memory in the extracellular antibody response would be to *retain* the original eliciting antigen in some cryptic form so that it could ‘prompt’ the immune system from time-to-time, so reminding it of the possibility that the source of the eliciting antigen might one day reappear. While some studies have suggested that trace amounts might persist [36, 37], the main memory strategy is to have a steric complement of an antigen (i.e. a specific antibody protein) and to retain an enhanced capacity to make more of that antibody.

However, in the case of not-self ‘RNA-antigens,’ persistence might be a more effective strategy. Thus, at steps D or E in Figure 15.5, an enzyme (a polymerase known as reverse transcriptase) might convert the dsRNA into dsDNA, which might then be inserted into the genome and selectively mutated (or methylated; see Chapter 18) to make it both inactive regarding any primary function, and ‘anti-near-self’ rather than ‘anti-self’ regarding its potential immune receptor function (see Chapter 16). One strand of the sequence would retain sufficient complementarity to the eliciting not-self RNA to render it of value, following transcription, in generating ‘RNA antibodies’ at times of future ‘stress’ (Fig. 15.4). If this could occur in germ-line tissue then many sequences with similarities to ancient virus sequences should be present in intergenic DNA, and traces of Lamarckian phenomena (inheritance of acquired characters) might emerge [22, 38]. For example, the HERV elements shown in Figure 15.3 are likely to be remnants of ancient retroviruses. A similar arrangement is present in prokaryotes (see below).

It is difficult to find an analog of the third step, positive selection. There should be selection of individuals in which favorable mutations preventing self dsRNA formation had been collected together by recombination. But this is not the same as the molding of lymphocyte repertoires that occurs during an individual lifetime. Nevertheless, we see in genomes that some genes have duplicated (Fig. 11.1),

sometimes due to reverse transcription of their mRNAs into the corresponding DNAs (see above). Often the duplicate is not functional in the usual phenotypic sense and is so mutated that it is dismissed as a “pseudogene.” Thus we can distinguish (i) a perfect matching self duplicate, (ii) a *near-self* duplicate, and (iii) a duplicate that had mutated so greatly that even weak hybridisation with the originating gene was eliminated. Near-self duplicates could be of great adaptive value in that, while not retaining the original function, they could still foil the attempt of a virus to mutate, step-by-step, from not-self (unambiguously foreign) towards self (not foreign; see Chapter 16).

Another example of intracellular positive repertoire selection might be a tendency over evolutionary time to accept mutations in some proteins because the mutations confer an ability to aggregate with other self proteins when they mutate away from self to ‘near-self.’ Whether RNAs generated from the hidden transcriptome would operate solely at the RNA level is not known. Their translation products might contribute to the intracellular population of protein ‘immune receptors’ (Chapter 16).

CRISPR

The case outlined above, derived from fundamental principles of immunology and bioinformatic analyses of sequences, was supported by the discovery of an analogous defense system in bacteria [39]. Many of their genomes have “clustered regularly interspaced short palindromic repeats” (CRISPR), between which are variable “spacer” sequences, each about 30 bases long. Two decades after their discovery by Atsuo Nakata and colleagues [40], there was much surprise in 2005 when the spacers were found to be segments from viruses that infect bacteria. In the course of a ‘primary’ infection a CRISPR region can acquire a “protospacer” sequence from a virus. Under optimal conditions bacteria divide in less than an hour and, if a bacterium survives the infection, the CRISPR region, with acquired spacer, is transmitted to its descendants. If there is a ‘secondary’ infection by a member of the same species of virus, this ‘memory’ can be called upon by bacterial descendants, which transcribe it to generate “prokaryotic short interfering” (psi), otherwise known as “CRISPR” (cr), RNAs. These target the complementary protospacer sequence of the infecting virus, be it in DNA or RNA. Of course, viruses mutate their protospacers to evade bacterial defenses, but the bacteria, in turn, are able to discard old spacers, and acquire new ones [41]. In a continuing arms race, some viruses have even hijacked an entire CRISPR system to retaliate in kind [42].

Such “immunospecific RNAs” (imRNAs) have also been observed in certain animals and plants [43–47]. But, bacterial systems being experimentally more tractable than those of multicellular organisms, we look here for early answers to basic questions: How does a bacterial host avoid its transcribed spacer psiRNA looping back and attacking its own source spacer, rather than attacking the protospacer in the genome of an intrusive virus? Even more basic, how is a virus sequence initially

distinguished from its host sequence? Numerous CRISPR associated (“Cas”) proteins are involved, and some of them, somehow, have to reliably detect some *difference*, perhaps related to the differential oligonucleotide compositions that distinguish species (Chapter 10), or perhaps related to some aspect of pathogen biology. For example, there is one host but, following their replication, many intracellular pathogens, only *one* of which has to err in some way to permit host capture of a spacer. Viruses are small and replicate rapidly, and so are likely to present *more* replication forks with associated base asymmetries (Fig. 6.4) than their hosts (Fig. 6.3).

In the case of *E. coli*, strand-orientation-specific CHI sequences (8-mers) are involved. These are found in **GT**-rich regions that display duplex extrusion asymmetry (Chapter 14). They moderate the fragmentation of a region where breaks in DNA are being repaired. Replication forks are particularly prone to such breaks (“hotspots”). Viruses with both *many* replication forks and *few* CHI sequences can be discriminated from their host on this basis [48].

E. coli are enriched in CHI sequences (Fig. 15.6). Some microorganisms manage well without a CRISPR system, but those that do (40% of bacteria and 90% of archaea) may use CHI-like sequences [49]. Some Cas proteins have found an important place in the genetic engineer’s toolbox [50]. Here there is the possibility of changing a mammalian embryo, so that the majority of the mature organism’s cells are ‘corrected,’ a process raising ethical concerns (germ line editing). To correct a late-emerging genetic disease the CRISP machinery has to be virally targeted to specific somatic cells (see Prologue).

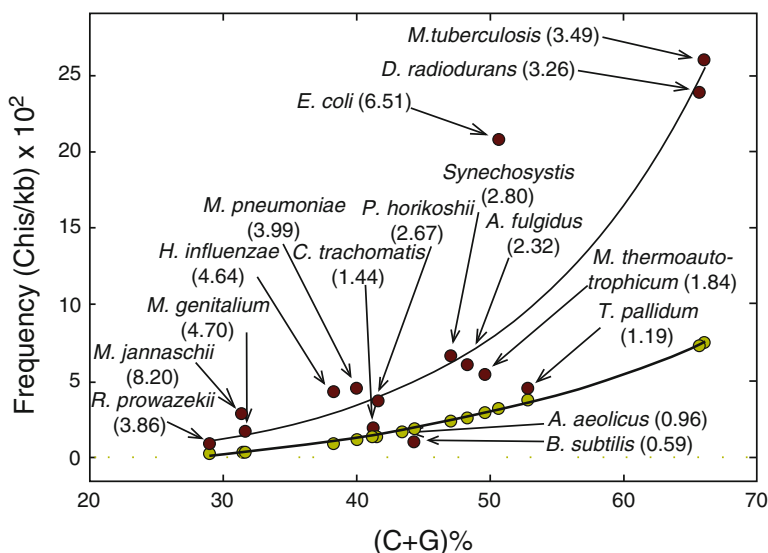


Fig. 15.6 Higher than expected frequency of CHI sequences in many bacterial genomes. Points correspond to observed frequencies (red circles) and the frequencies expected if base order were random (green circles). Organisms corresponding to each point are shown with ratios of observed to expected frequencies in parentheses. For details see [49].

Summary

The full understanding of the genome of a biological species requires an understanding of the genomes of the species with which it has coevolved, which include pathogenic species. Members of a pathogenic species that enter the bodies of members of a host species must be recognized as “not-self” so activating host immunological defenses. However, the capacity of its genomic information channel being limited, a host whose defenses are poised to attack “near-self” versions of not-self (a narrow repertoire), rather than not-self *per se* (a repertoire formidable in range), should be at a selective advantage. It is likely that immune systems of multicellular organisms are adaptations of those of unicellular organisms, which had *already* developed the capacity for self/not-self discrimination. From this perspective we can comprehend phenomena such as “junk” DNA, genetic polymorphism and the ubiquity of repetitive elements. That which is evolutionarily conserved is often functional, but that which is functional is not necessarily conserved. Variation itself may be functional. The “hidden transcriptome,” revealed by run-on transcription of genes or repetitive elements constitutes a diverse repertoire of *inherited* RNA ‘immune receptors’ with the potential to form double-stranded RNA with viral RNA ‘antigens,’ so triggering intracellular alarms. Both genic and non-genic DNA would have been screened over evolutionary time (by selection of individuals in which favorable mutations had been collected together by recombination) to decrease the probability of two complementary “self” transcripts interacting to form dsRNA segments above a critical length (about two helical turns). Thus, many RNAs are purine-loaded. The CRISPR system provides an analogous RNA immune receptor defence in microorganisms. Here numerical differences in CHI sequences and replication forks may facilitate self/not-self discrimination.

References

1. Shaw HW (1866) *Josh Billings, His Sayings*. Carleton, New York
2. Butler S (1965) *Earnest Pontifex, or The Way of All Flesh*. Howard DF (ed) Methuan, London, p. 159 [The quoted paragraph was removed from the original 1903 version.]
3. Mira A (1998) Why is meiosis arrested? *Journal of Theoretical Biology* 194:275–287
4. Johnson J, Canning J, Kaneko T, Pru JK, Tilly JL (2004) Germ line stem cells and follicular renewal in the post-natal mammalian ovary. *Nature* 428:145–150
5. Granovetter M (1983) The strength of weak ties. A network theory revisited. *Sociological Theory* 1:201–233
6. Pancer Z, Amemiya CT, Ehrhardt GRA, Ceitlin J, Gartland GL, Cooper MD (2004) Somatic diversification of variable lymphocyte receptors in the agnathan sea lamprey. *Nature* 430:174–180
7. Zhang S-M, Adema CM, Kepler TB, Loker ES (2004) Diversification of Ig superfamily genes in an invertebrate. *Science* 305:251–254
8. Brücke E (1861) Die Elementarorganismen. *Sitzungsberichte der Akademie der Wissenschaften Wien, Mathematische-wissenschaftliche Classe* 44:381–406
9. Forsdyke DR, Madill CA, Smith SD (2002) Immunity as a function of the unicellular state: implications of emerging genomic data. *Trends in Immunology* 23:575–579

10. Shadrin AA and Parkhomchuk DV (2014) Drake's rule as a consequence of approaching channel capacity. *Naturwissenschaften* 101:939–954
11. Battail G (2014) Channel capacity and channel coding. In: *Information and Life*. Springer, Dordrecht, pp. 93–131 [Essentially the same message as in the present book, but from the perspective of an information scientist.]
12. Nóbrega MA, Zhu Y, Plajzer-Frick I, Afzal V, Rubin EM (2004) Megabase deletions of gene deserts result in viable mice. *Nature* 431:988–993
13. Palazzo AF, Gregory TR (2014) The case for junk DNA. *PLOS Genetics* 10:e1004351
14. Ohno S (1972) So much 'junk' DNA in our genome. *Brookhaven Symposium on Biology* 23:366–370
15. Plant KE, Routledge SJ, Proudfoot NJ (2001) Intergenic transcription in the human β -globin gene cluster. *Molecular & Cellular Biology* 21:6507–6514
16. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, Fodor SP, Gingeras TR (2002) Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296:916–919
17. Johnson JM, Edwards S, Shoemaker D, Schadt EE (2005) Dark matter in the genome: evidence of widespread transcription detected by microarray tiling experiments. *Trends in Genetics* 21:93–102
18. Darwin C (1871) *Descent of Man, and Selection in Relation to Sex*. Appleton, New York, pp 156–157
19. Faulkner GJ, et al. (2009) The regulated retrotransposon transcriptome of mammalian cells. *Nature Genetics* 41:563–571
20. Ferguson BJ, Cooke A, Peterson P, Rich T (2008) Death in the AIRE. *Trends in Immunology* 29:306–312
21. Nishikawa Y, Hirota F, Yano M, Kitajima H, Miyazaki J, Kawamoto H, Mouri Y, Matsumoto M (2010) Biphasic Aire expression in early embryos and in thymic medullary epithelial cells before end-stage terminal differentiation. *Journal of Experimental Medicine* 207:963–971
22. Cristillo AD, Mortimer JR, Barrette IH, Lillicrap TP, Forsdyke DR (2001) Double-stranded RNA as a not-self alarm signal: to evade, most viruses purine-load their RNAs, but some (HTLV-1, EBV) pyrimidine-load. *Journal of Theoretical Biology* 208:475–491
23. Wilkins C, Dishongh R, Moore SC, Whitt MA, Chow M, Machaca K (2005) RNA interference is an antiviral defence mechanism in *Caenorhabditis elegans*. *Nature* 436:1044–1047
24. Guglielmi KM, McDonald SM, Patton JT (2010) Mechanism of intraparticle synthesis of the rotavirus double-stranded RNA genome. *Journal of Biological Chemistry* 285:18123–18128
25. Lauffer MA (1975) *Entropy-driven Processes in Biology*. Springer-Verlag, New York
26. Evans SS, Repasky EA, Fisher DT (2014) Fever and the thermal regulation of immunity: the immune system feels the heat. *Nature Reviews Immunology* 15:335–349
27. Tellam JT, Lekieffre L, Zhong J, Lynn DJ, Khanna R (2012) Messenger RNA sequence rather than protein sequence determines the level of self-synthesis and antigen presentation of the EBV-encoded antigen, EBNA1. *PLOS Pathogens* 8:e1003112 [The reciprocal experiment is to drastically change the protein while hardly changing the mRNA. This can be achieved by adding or subtracting one base (so changing the reading frame; RF), an experiment performed both by 'Nature,' and by human experimenters. If no stop codons are generated by the frame shift, then it is an "open reading frame" (ORF). See: (Kwun HJ, Silva SR da, Shah IM, Blake N, Moore PS & Chang Y (2007) Kaposi sarcoma-associated herpes virus latency-associated nuclear antigen 1 mimics Epstein-Barr virus EBNA1 immune evasion through central repeat domain effects on protein processing. *Journal of Virology* 81:8225–8235; Zaldumbide A, Ossevoort M, Wiertz EJHJ, Hoeben RC (2007) *In cis* inhibition of antigen processing by the latency-associated nuclear antigen 1 of Kaposi sarcoma Herpes virus. *Molecular Immunology* 44:1352–1360).]
28. Murat P, Zhong J, Lekieffre L, Cowieson NP, Clancy JL, Preiss T et al. (2014) G-quadruplexes regulate EBV-encoded nuclear antigen 1 mRNA translation. *Nature Chemical Biology* 10:358–364
29. Larocca D, Chao LA, Seto MH, Brunck TK (1989) Human T-cell leukaemia virus minus strand transcription in infected T-cells. *Biochemical and Biophysical Research Communications* 163:1006–1013

30. Cook LB, Elemans M, Rowan AG, Asquith B. (2013) HTLV-1: Persistence and pathogenesis. *Virology* 435:131–140
31. Forsdyke DR (2014) Implications of HIV RNA structure for recombination, speciation, and the neutralism-selectionism controversy. *Microbes & Infection* 16: 96–103
32. Saleh MC, Tassetto M, van Rij RP, Goic B, Gausson V, Berry B, Jacquier C, Antoniewski C, Andino R (2009) Antiviral immunity in *Drosophila* requires systemic RNA interference spread. *Nature* 458:346–350
33. Dunoyer P, Schott G, Himber C, Meyer D, Takeda A, Carrington JC, Voinnet O (2010) Small RNA duplexes function as mobile silencing signals between plant cells. *Science* 328:912–916
34. Barak M, Levanon EY, Eisenberg E, Paz N, Rechavi G, Church GM, Mehr R. (2009) Evidence for large diversity in the human transcriptome created by *Alu* RNA editing. *Nucleic Acids Research* 37:6905–6915
35. Ota T, et al. (2004) Complete sequencing and characterization of 21243 full-length human cDNAs. *Nature Genetics* 36:40–45
36. Garvey JS, Cambell DH (1966) Autoradiographic investigation of tissues after primary and multiple antigenic stimulation of rabbits. *Nature* 209:1201–1202
37. Li C, Buckwalter MR, Basu S, Garg M, Chang J, Srivastava PK (2012) Dendritic cells sequester antigenic epitopes for prolonged periods in the absence of antigen-encoding genetic information. *Proceedings of the National Academy of Sciences, USA* 109:17543–17548
38. Zabolotneva A, Tkachev V, Filatov F, Buzdin A (2010) How many small interfering RNAs may be encoded by the mammalian genomes? *Biology Direct* 5:62
39. Barrangou R, Marraffini LA (2014) CRISPR-Cas systems: prokaryotes upgrade to adaptive immunity. *Molecular Cell* 54:234–244
40. Nakata A, Amemura M, Makino K (1989) An unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K12 chromosome. *Journal of Bacteriology* 171:3553–3556
41. Fineran PC, Gerritzen MJH, Suárez-Diez M, Künne T, Boekhorst J, Hijum SAFT van, Staals RHJ, Brouns SJJ (2014) Degenerate target sites mediate rapid primed CRISPR adaptation. *Proceedings of the National Academy of Sciences, USA* 111:E1629–E1638
42. Seed DS, Lazinski DW, Calderwood SB, Camilli A (2014) A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* 494:489–491
43. Flegel TW (2009) Hypothesis for heritable, antiviral immunity in crustaceans and insects. *Biology Direct* 4:32
44. Bertsch C, Beuve M, Dolja VV, Wirth M, Pelsy F, Herrbach E, Lemaire O (2009) Retention of virus-derived sequences in the nuclear genome of grapevine as a potential pathway to virus resistance. *Biology Direct* 4:21
45. Tanguy M, Miska EA (2013) Antiviral RNA interference in animals: piecing together the evidence. *Nature Structural & Molecular Biology* 20:1239–1241
46. Rechavi O (2014) Guest list or black list: heritable small RNAs as immunogenic memories. *Trends in Cell Biology* 24:212–220
47. Parrish NF, et al. (2015) piRNAs derived from ancient viral processed pseudogenes as trans-generational sequence-specific immune memory in mammals. *RNA* 21:1691–1703
48. Levy A, Goren MG, Yosef I, Auster O, Manor M, Amitai G, Edgar R, Qimron U, Sorek R (2015) CRISPR adaptation biases explain preference for acquisition of foreign DNA. *Nature* 520:505–510
49. Lao PJ, Forsdyke DR (2000) Crossover hot-spot instigator (CHI) sequences in *Escherichia coli* occupy distinct recombination/transcription islands. *Gene* 243:47–57
50. Charpentier E, Marraffini LA (2014) Harnessing CRISPR-Cas9 immunity for genetic engineering. *Current Opinion in Microbiology* 19:114–119

Chapter 16

The Crowded Cytosol

*If you can talk with crowds and keep your virtue,
Or walk with Kings – nor lose the common touch,
Yours is the Earth and everything that's in it."*

Rudyard Kipling [1]

While, given the expansion of the world population, it cannot be guaranteed that it will always be so, currently one accepts as a temporary discomfort short periods of crowding, as in a crowded elevator or subway train. However, if, like most intracellular macromolecules, one were both blind and deaf, the need to communicate by touch might make crowding an option of choice. It seems likely that the first cells to evolve soon discovered the advantage of intracellular crowding, which persists to this day [2]. Indeed, the physiological environment of enzymes is very different from the environment we can normally create in the test-tube.

The French microbiologist Antoine Bechamp reported in 1864 that enzymes ("ferments") in yeast would work outside of the cell of origin [3]. More than anything else, this demystified the cellular "protoplasm" and suggested that, in principle, it should be possible to take a cell apart and then reassemble it from its individual components – like a clock. So enzymes were purified and characterized. It even became possible to purchase enzymes 'off the shelf' as protein powders. The study of their abilities to convert target molecules (substrates) to other molecules (products) kept biochemists busy for much of the twentieth century, and the fundamental chemistry of life emerged.

For this, the enzymes had to be dissolved in suitable salt solutions. But it is difficult to dissolve proteins at concentrations higher than 10 mg/ml. In contrast, cytosolic proteins are collectively at concentrations around 300 mg/ml! Within cells many water molecules are likely to be ordered (made relatively immobile) by crowded macromolecules and so are less able to diffuse freely. This means that many chemical changes within intracellular fluids that involve macromolecules are likely to be entropy-driven (Chapter 15).

Piles of Coins

The concentration of proteins within extracellular fluids is also very high (about 80 mg/ml). Cells suspended in these fluids can ‘lose solubility’ and aggregate when either the total protein concentration, or the concentrations of certain macromolecules, exceeds certain limits. Normally mammalian red blood cells (erythrocytes) appear as flat red disks suspended in blood plasma. When the concentration of proteins in plasma is increased only slightly, the red cells move, as if under the direction of unseen hands, and queue up to form ‘piles of coins’ (“rouleaux”). It is possible to watch this and demonstrate the specificity of the aggregation by light microscopy. In mixtures of red cells from different species, cells of the same species preferentially aggregate with each other (Fig. 16.1).

Thus, the entropy-driven aggregation shows specificity – ‘like’ aggregating with ‘like.’ In mixtures, red cells form homoaggregates (like with like), not heteroaggregates (like with unlike). This reflects a general tendency for shared regularities in structure, and shared molecular vibrations (resonances), to promote self-self interactions between particulate entities, be they whole cells, or discrete macromolecules (Chapter 2).

It is important to note that the aggregation is a response to an increase in total proteins in the surrounding medium and there is no direct cross-linking of cells by the proteins. The proteins are not acting as cross-linking agents (ligands; Fig. 16.2).

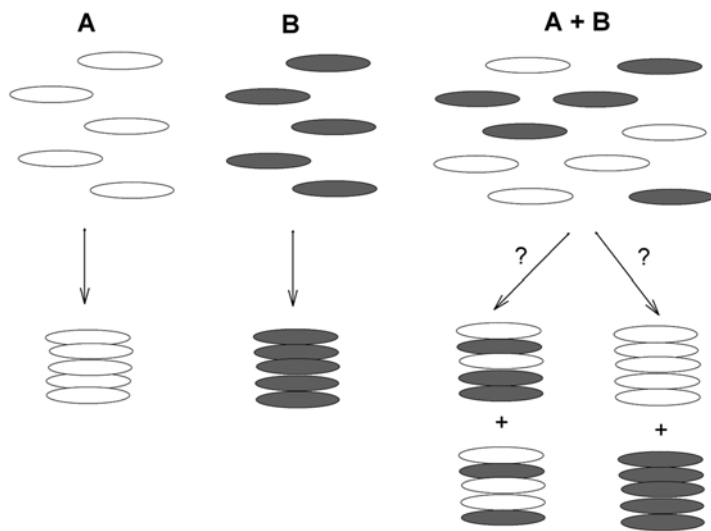


Fig. 16.1 Specificity of rouleaux formation. A rouleaugenic agent (e.g. polymerized blood albumin) is equally active in aggregating red blood cells from animal species *A* or from animal species *B*. The aggregates appear as rouleaux (‘piles of coins’). When the two cell populations are mixed and then treated with the rouleaugenic agent, each of the resulting rouleaux should either contain both *A* and *B* red cells (indicating non-specific aggregation), or all *A* and all *B* red cells (indicating specific aggregation). The latter alternative is found experimentally

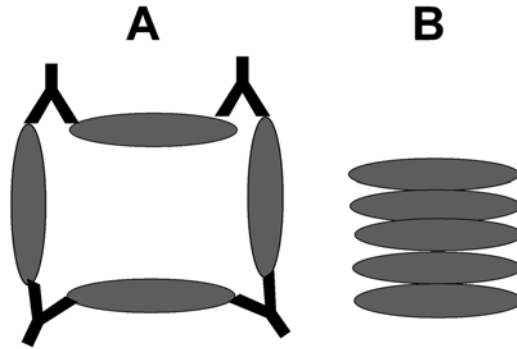


Fig. 16.2 Distinction between (A) aggregation of red blood cells by an extracellular cross-linking agent (antibody) and (B) aggregation that does not involve a cross-linking agent. In (A) the disk-shaped red cells are cross-linked by Y-shaped bivalent antibodies (not drawn to scale). The bonding here is strong and the aggregates are not readily disrupted. In (B) the red cells adopt the energetically most favorable (entropy-driven) pile-of-coins conformation (rouleaux). The bonding here is weak and the aggregates are readily disrupted. Within an organism the high protein concentration of the extracellular environment should promote the aggregation of similar cells into a common tissue – a process that, initially, might not need cross-linking agents

The increase in protein concentration can be produced non-specifically by addition of an excess of any one of a variety of proteins. Thus, rouleaux-generation is a *collective* function of proteins. The appearance of rouleaux in a blood sample provides an index of the underlying state of the plasma proteins, not of the red blood cells. However, if red cells themselves are thought of as merely large proteins, then the phenomenon of rouleaux formation can assist our understanding of the phenomenon of protein aggregation as it occurs in concentrated solutions. This can be of help when considering possible mechanisms for intracellular self/not-self discrimination [4].

Homoaggregates

A specific protein can be induced to self-aggregate so forming homoaggregates (Greek: *homos* = same). This is achieved either by increasing the concentration of surrounding proteins or by increasing the concentration of the protein itself. In some cases homoaggregate formation reflects an obvious physiological function. Thus, coat protein molecules of tobacco mosaic virus (TMV) naturally aggregate to form the outer coat that protects virus nucleic acid during passage from cell to cell. But if an enzyme aggregates, its activity often diminishes and it may become insoluble (i.e. it precipitates). To this extent, aggregation can be non-physiological and harmful to the cell.

On the other hand, proteins are unstable, with life-spans extending from minutes to days, and homoaggregation can assist protein breakdown. In this stepwise process a protein first becomes tagged (marked) as ready for degradation, and then is cleaved

into fragments (peptides). Finally the peptides are degraded to the relatively stable amino acid ‘building blocks’ from which new proteins can be assembled. This process of synthesis, breakdown and reassembly is referred to as “protein turnover” (Chapter 2).

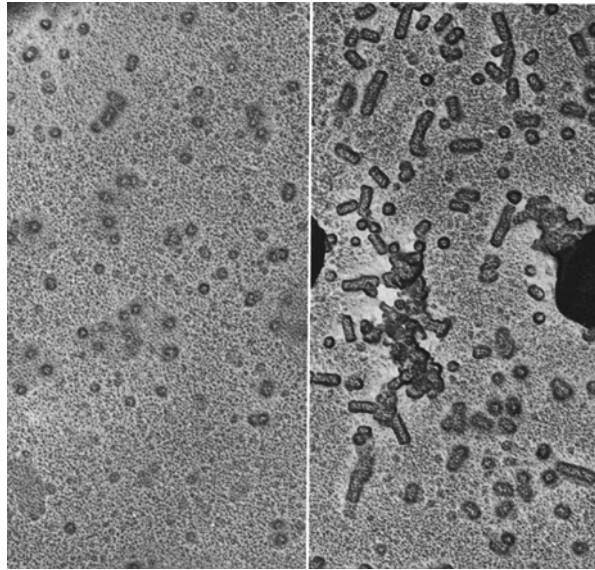
Sometimes, however, under special circumstances (see later), certain peptide fragments are not further degraded. Instead, they are united with peptide-display proteins (major histocompatibility complex proteins; MHC proteins), and taken to the surface of the cell. Here, as peptide-MHC complexes (pMHCs), they are recognized by cytotoxic lymphocytes, which destroy the cell. The underlying principle that emerged in the 1980s was amazingly simple, and I (and probably others) kicked myself for not realizing it earlier. You do not need a whole elephant to diagnose an elephant. You do not need a whole protein to diagnose a protein. Immunologists already knew this! They had coined the expressions “antigenic determinant” and “epitope” for a part of a protein (antigen) that would suffice for an extracellular immunological recognition event. Yet, no one seemed to have considered the possibility that cells might *detach* the equivalent of an antigenic determinant from a protein *intracellularly* prior to engaging in a diagnostic recognition event. This raised the question of specificity [4]. How do cells distinguish such peptides from those produced as part of regular protein turnover?

In studies with TMV, Lauffer showed that aggregation involves the liberation of water molecules bound to the macromolecules [5]. Thus, while it might appear, from the observed aggregation, that entropy was decreasing, the increase in disorder of the liberated water molecules more than compensated for the increase in order of the macromolecules. System entropy increased. If the aggregation were entropy-driven (endothermic), then it should be promoted by a small increase in temperature. Indeed, aggregation can be induced by increasing the temperature over a narrow range, much lower than would be needed to disrupt the structure of (denature) the protein (Fig. 16.3).

Collective Pressure

Homoaggregation at high protein concentrations generates specific, but relatively unstable, aggregates (i.e. only weak chemical bonding is involved). If you were to progressively concentrate a mixture of proteins (A, B, C, D) in aqueous solution then, at a certain critical concentration, one of the proteins, say A, would self-aggregate. In this process each molecule of A would loose some of its bound water. In the absence of the surrounding proteins (B, C, D) much *higher* concentrations of A would be needed for aggregation to occur. Thus, a group of proteins *collectively* exerts a ‘pressure’ (due to their binding of water) tending to force individual protein species to self-aggregate and give up their water. A protein species with the greatest tendency to self-aggregate (a function of factors such as shape, size, and initial concentration) aggregates first. As total protein concentration increases, other protein species aggregate in turn.

Fig. 16.3 Aggregation of TMV coat protein by increasing the temperature. This is an electron micrograph of a solution of coat protein (0.1 mg/ml) held at either 4°C (left) or 23°C (right). Reprinted from [5]



The concept of the ‘crowded cytosol’ implies that much intracellular water is bound to proteins, so that there is always a strong *standing* pressure to drive into homoaggregates any macromolecular species that exceeds the solubility limits imposed by the macromolecules surrounding it. Each individual macromolecular species can both contribute to, and be acted upon by, the pressure. Thus, each macromolecular species can have this *collective* function as well as a *specific* function. Both functions can affect phenotype and hence influence selection by evolutionary forces.

Concentration Fine-Tuning

It follows that the concentration of a protein is an important attribute that is not necessarily related, in any simple way, to the role the protein might normally play in the life of a cell. A protein has evolved to carry out a specific primary task. On grounds of economy, it might be supposed that evolutionary forces would have pressed for a maximization of specific activity (i.e. enzyme activity/protein molecule). This would minimize the necessary concentration of the protein. But the virtue of minimizing the concentration of a protein is not immediately apparent.

Provided its concentration is not extreme, a protein should not itself burden a cell. Indeed, if a *collective* function (e.g. the ability to exert a pressure to drive other proteins from solution) were an important attribute of a protein, then the steady-state concentration of the protein might tend towards the maximum compatible with the protein remaining in solution without self-aggregation. This might tend to decrease any tendency to maximize specific activity because the number of

molecules present could suffice for the necessary level of activity and there might be no selection pressure for them to improve on a per-molecule basis (but see later).

The steady-state concentration of a protein is determined by evolutionary forces acting on parts of the corresponding gene (i.e. base mutations) to affect factors such as mRNA transcription rate, and mRNA and protein stabilities. For example, a mutation that decreases the transcription rate of a gene decreases the concentration of the corresponding mRNA, and hence the concentration of the protein that is made by translating that mRNA.

A protein within a cell ends up with a certain specific activity, which can be less than the maximum possible. Over evolutionary time, the concentration of the protein is fine-tuned to the concentrations of its fellow travelers – the other diffusible proteins with which, from generation to generation, it has shared a common cytosol. In this circumstance, whereas normally the protein would be soluble, a mutation in the corresponding gene could result in homoaggregation and insolubility. This might provide an opportunity to register the protein as “not-self.”

Heteroaggregates

Why fine-tune? In general, as hinted at in Chapter 2, fine-tuning creates a narrower frame of reference so broadening the range of events that may be discerned as falling outside that frame. A factor favoring the precise fine-tuning of cytosolic protein concentration would be the need to discriminate self-proteins both from mutated self-proteins and from foreign proteins (such as might be encoded by a virus). We need to discriminate between “self,” “near-self,” and “not-self.” Mutated self-proteins are only slightly changed and so can be considered as near-self proteins, rather than not-self proteins. But, although near-self, they need to be registered as not-self. Virus proteins would be foreign and less likely to correspond with self (i.e. they would be not-self). But viruses that could accept mutations making their proteins *more like* their host’s proteins (i.e. that could approach near-self) might be at a selective advantage. As we shall see, this advantage would fade if host defenses were *preset* for discriminating between self and near-self in such a way that the latter would register as not-self.

A system for intracellular self/not-self discrimination could have evolved when the first unicellular organisms arose and were confronted with the first prototypic viruses (Chapter 15). A mutated self-protein might have lost activity either as a direct result of the mutation (e.g. the mutation might have affected the active center of an enzyme) or because the mutation had decreased its solubility – perhaps manifest as homoaggregate formation. However, when mutated, a resulting structural (or vibrational) change might, by chance, have *created* some degree of reactivity with one or more of the many *other* diffusible protein species within the same cell. Hence, the mutated self-protein might ‘cross-seed’ the aggregation of unrelated proteins. Heteroaggregates might form (Greek: *hetero* = different). There could then be a loss of function not only of the primarily mutated protein, but *also* of the coaggregated

proteins. Thus, a mutation in one protein might affect the functions of other specific proteins in unpredictable ways, generating complex mutational phenotypes (pleiotropism). Certain clinical syndromes may be sets of such diverse, often not obviously related, altered phenotypic characters that are observed in disease states.

To the extent that a primary mutation does not result in heteroaggregate formation, then the function of a cell where the mutation has occurred might not be affected, and the cell might persist. Cells with mutations that result in heteroaggregate formation are more likely to be functionally impaired. Unless aggregation were required for its primary function, a protein would have been fine-tuned over evolutionary time *not* to interact with the many thousands of other diffusible protein species with which it had been traveling through the generations in the same cytosol. Such interaction might have impaired the function both of the protein and those with which it interacted. So organisms with mutations leading to interactions would have been negatively selected.

Heteroaggregate formation could, however, be of adaptive value if the primary mutation was in a gene controlling cell proliferation, which might result in a loss of control and hence cancer. In this case, a gene encoding, by chance, a normal protein that would coaggregate with a mutant cancer-causing protein (oncoprotein) would confer a selective advantage, over and above that conferred by virtue of the gene’s normal function (i.e. while still retaining its primary function, it would be *positively selected* for encoding an ‘immune receptor’ that would function as a detector of near-self oncogenic changes). Genes encoding products with such coaggregating functions would tend to make cancer a disease of post-reproductive life. This is a time when selective factors that tend to promote the number and reproductive health of descendents are less important; so cancer prevention would be less evolutionarily advantageous [6].

Protein ‘Immune Receptors’

From this we see that potential functions of a protein include (i) its primary function (e.g. enzyme activity), (ii) any rare secondary ‘moonlighting’ functions [7], (iii) its contribution to the total cytosolic aggregation pressure, (iv) its ability to form heteroaggregates with either mutated self-proteins (e.g. oncoproteins) or foreign proteins that may have been introduced by a virus. In the latter respect, cytosolic proteins can be regarded as ‘intracellular antibodies,’ or protein ‘immune receptors’ (Fig. 15.1).

The genes of a host cell and the genes of an invading virus differ in various ways that might assist discrimination between self and not-self. Within the species-limit, host cell self-genes *traveling together* through the generations should have had ample opportunity collectively to coevolve and fine-tune to each other. On the other hand, the goal of a virus is to multiply and spread, preferentially within the lifetime of its host. From this it might be thought that the cytosolic concentrations of virus gene products would be less fine-tuned than the cytosolic concentrations of host gene products. However, the high replication and mutation rates of viruses relative

to their hosts make it likely that viruses would be no less fine-tuned than their hosts. Indeed, host fine-tuning over evolutionary time would have tended to create a *uniform* intracellular environment, thus decreasing a major virus anxiety – how to anticipate the conditions it would find in its next host. Hence it is unlikely that the proteins of a ‘street smart’ intracellular pathogen would readily exceed the solubility limits imposed by host proteins in the crowded cytosol.

In their role as ‘immune-receptors,’ host cytosolic proteins should form a diverse antibody-like environment capable of forming heteroaggregates with virus proteins, which would accordingly register as not-self (Chapter 15). This primary self/not-self discrimination event, perhaps in an environment made permissive by dsRNA alarms (i.e. cells are alerted to become more conducive to registering not-self), should result in processing of the heteroaggregates by cytoplasmic particles known as proteosomes. Here proteins are fragmented into peptides, but are not further degraded to amino acids. Instead, the peptides are displayed at the cell surface by the peptide-display proteins (MHC proteins). This pMHC display is recognized by cytotoxic cells of the lymphoid system (T-lymphocytes), which destroy the virus-containing cell. The organism then loses the services of one of its cells. But most tissues can readily replace lost cells by the mitotic division of other cells that are not infected.

How diverse is the intracellular protein immune-receptor repertoire likely to be? At the very least it would include the many diffusible protein species normally present in the cytosol of a differentiated cell type. In addition, when a virus ‘tripped’ the self/not-self discrimination alarm, there might be translation into proteins of some of the RNA products of the ‘hidden transcriptome’ [8, 9]. Perhaps at the behest of something like the AIRE protein (Chapter 15), these proteins might include the products of tissue-specific genes *not* normally expressed in a host cell of a particular tissue type (i.e. a brain-specific gene might be abnormally expressed in an infected kidney cell).

Phenotypic Plasticity

It would be expected that within a kidney cell such abnormally expressed products would be present only at the very low concentrations needed for their role in identifying the virus proteins with which they react (so registering virus proteins as not-self and tripping appropriate alarms). But there would also be the possibility, in the case of proteins normally required at very low concentrations, of an unwelcome effect on the phenotype. For example, very low concentrations of critical regulatory proteins are synthesized at various unique time-points during embryogenesis, so bringing about developmental switches. Thus, developing embryos, within pregnant females who are under virus attack (or an equivalent stress), might produce a developmental switch protein at the wrong time (see rubella virus; Chapter 15). Offspring would then appear with mutant phenotypes resembling the mutant phenotypes observed among the offspring of pregnant females accidentally exposed to mutagens (e.g. X-rays) that had mutated embryo DNA.

However, if viable, the mutant offspring of a parent that had been under virus attack would not, in turn, be able to pass their mutant characters on to their offspring. In contrast, the mutant offspring of a parent that had been exposed to mutagens might be able to pass on their mutant characters if the mutation had affected developing germ-line cells. Richard Goldschmidt noted that mutant forms resulting from developmental stress would be classified as “phenocopies,” rather than “genocopies,” since their mutant characters would not be inherited (i.e. there would be no underlying causal genetic change) [10]. In other words, stressed organisms display “phenotypic plasticity” not “genotypic plasticity.”

Polymorphism Makes You Individual

High host polymorphism would make it difficult for viruses to anticipate the ‘immune receptor’ RNA and ‘immune receptor’ protein repertoires of future hosts. Furthermore, peptides generated from heteroaggregates of virus and host proteins, would *include* host protein-derived peptides. These self-antigens, as well as, *or instead of*, the antigens of the pathogen, would then serve as targets for attack by cytotoxic T-lymphocytes [11]. Accordingly, the variability (polymorphism) of intracellular proteins (i.e. differing from individual to individual of the host species) would tend to *individualize* the immune response to intracellular pathogens (and cancer cells). T-lymphocytes from one virus host (or one cancer subject) might not recognize cells of another host infected with the same virus (or afflicted with the same type of cancer). From this perspective, protein polymorphism is not neutral, as sometimes supposed, but serves to adapt potential host organisms as ‘moving targets,’ so *militating against pathogen preadaptation*.

Thus, the designation by the host of a virus protein as “not-self,” might involve both *quantitative* factors (i.e. homoaggregate formation if a virus protein’s concentration exceeds a solubility threshold, which the virus might easily avoid by mutation), and *qualitative* factors (i.e. the recruitment of various polymorphic host proteins into protein heteroaggregates, which the virus might not so easily avoid by mutation). However, the T-lymphocytes primed by specific peptide fragments from *self* antigens might, besides multiplying and attacking the virus-infected (or cancer) cell, also react against the same self antigens should they, perchance, be displayed by normal cells. In cancer patients this could result in *immunological diseases* of various tissues, other than the primary cancer tissue [12].

Such a “paraneoplastic disease” arises when melanoma tumors are attacked by cytotoxic T-lymphocytes that recognize peptide fragments from melanin pigment, a normal tissue-specific product. The lymphocytes can react against both the tumor cells and some normal melanin pigment-forming cells, creating white skin patches (vitiligo). Despite the activities of the AIRE protein in the thymus (Chapter 15), under normal circumstances many intracellular proteins (potential self antigens) are *not* displayed by cells, so there is no deletion of T-lymphocytes with the potential to react with these proteins when the T-lymphocyte repertoire is being purged and molded (see later).

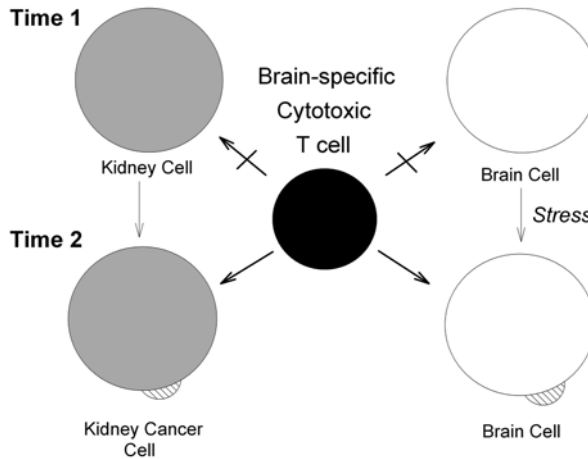


Fig. 16.4 Paraneoplastic disease. Normally a cytotoxic T-cell (centre) with a receptor recognizing a specific MHC-associated brain peptide (striped bump), will attack neither brain cells, nor cells of other tissues, such as kidney cells (Time 1). Later the kidney cells become kidney cancer cells. They may then display brain peptides which stimulate proliferation of T-cells (Time 2). These can attack both the stimulating cancer cells and stressed *normal* brain cells

Intriguingly, white skin patches are most likely to form where there has been local trauma to the skin [13]. Just as sound in the wood alerts sheep to move away (Chapter 15), so externally-inflicted stress (not-self) appears to alert skin cells, which are provoked to display fragments of self-antigens. If the concentration of the corresponding specific T-lymphocytes is sufficiently high, then the traumatized, but essentially normal, self-cells are attacked. Similarly, a minor knock on the head that might normally pass unnoticed, could lead to an immunological disease of the brain if, at that time, cytotoxic T-lymphocytes happened to be rejecting, say, incipient kidney cancer cells that were displaying fragments from a protein that was normally brain-specific (Fig. 16.4). Thus, paradoxically, the first symptom of a kidney cancer might be neurological impairment. Indeed, the T-cell attack might eliminate the provoking cancer, which might then never be detected. Instead the subject would have acquired an autoimmune disease of the brain!

Polymorphism Also Makes You Vulnerable

Polymorphism individualizes and keeps potential intruders guessing. So variation between allelic genes within a species (polymorphism) is generally of much adaptive value. But there is also a potential vulnerability. As diploid organisms, humans are heterozygous for many allelic pairs. Often only one allele is expressed in an individual cell – there is monoallelic expression from either the copy of maternal origin or the copy of paternal origin [14, 15]. But mass transcription as part of the ‘stress’

response to cellular invasion (Chapter 15) could include transcription from *both* parental alleles (biallelic expression). If the corresponding proteins were different from each other, then in some cases this polymorphism could result in their mutual aggregation and an unwanted presentation of self peptides to T cells. If this were a significant hazard then some method of prescreening would be advantageous.

The discovery that during T cell education the ‘promiscuous’ thymic expression (under control of the AIRE protein), of certain self-antigens specific for tissues other than the thymus, could be *biallelic* [16], suggests a central mechanism by which undesirable future chance interactions in the periphery, between the products of polymorphic allelic genes, could be eliminated. Thus AIRE-induced transcription would be needed for the display of self-peptides corresponding to genes whose allelic products happened to interact in a way that would generate aggregates for proteasome processing. Failure to eliminate T cells responding to peptides from such aggregates could contribute to the range of autoimmune diseases seen in AIRE mutants.

Under normal circumstances, having *centrally* purged the T cell repertoire of cells responding to self-pMHCs (negative repertoire selection), the remaining repertoire – the successful ‘auditioners’ – would be available in the periphery to seek out pMHC displays emanating from aggregates generated as part of responses to foreign invasion. If proteins from an intruder cannot be persuaded to self-aggregate, then the intracellular ‘antibody’ repertoire helps seek them out (normal-self recognizing not-self), and can generate a novel coaggregate, peptides from which (as pMHC) activate T cells.

This activation normally requires a *lower* pMHC concentration than negative selection. How would the thymus cells achieve a suitably high concentration of displayed pMHCs for negative selection? One solution would be for individual antigen-presenting cells (known as medullary epithelial cells; MEC) to transcribe only a small proportion of the total number of AIRE-responsive genes. Thus, if the entire biallelic gene set be represented as A-Z, one MEC might transcribe the bialleles ABC, and another the bialleles DEF, the choice being random. Thus, if A interacts with its fellow allele A, this will be picked up automatically; but if A interacts with Q (because of a failure of fine-tuning over evolutionary time), this will be evident only in a cell that, by chance, transcribes, say AQW at the behest of AIRE. Since there is less competition from other pMHCs, by restricting a MEC to displaying only part of the A-Z spectrum (mosaic expression pattern), the cell should be able to achieve the required high pMHC concentration needed for negative selection, at least to the extent of countermanding future inter-allele effects [17].

Death at Home or in Exile

Dividing cells are often seen when tissues are examined under the microscope. For adult organisms of relatively constant size, cell multiplication must be accompanied either by a corresponding number of cell deaths, or by the exile of superfluous cells beyond the body perimeter. Since exiled cells generally die, the options are bleak.

Die at home or die in exile! The only exceptions are the happy gametes that can find appropriate partners and so generate new individuals. If the balance (homeostasis) between cell multiplication and destruction is lost, cancer can result.

Death-style options are limited. When there is trauma to a tissue, cells may die by “necrosis,” a process that may involve activation of T-lymphocytes and the migration from dilated blood vessels of phagocytic cells (Greek: *phagein* = to eat). The region may become warm, tender and red (i.e. “inflamed”). Dying cells are ingested by phagocytic cells, which degrade their macromolecules.

However, usually cells are eliminated without inflammation. For example, dead skin cells are simply sloughed off into the environment. Each time you undress you discard not only clothes, but also around 400,000 skin cells [18]. Cells that cannot be discarded in this way invoke physiological auto-destructive mechanisms. Without fanfare, the cells self-digest, and their breakdown products are quietly ingested by neighboring cells (“apoptosis”). However, apoptosis is also a possible outcome of an intracellular self/not-self recognition event (see later).

The discarding of cells and/or their secretions into the environment may occur without an organism being aware of it (e.g. sloughing skin cells). However, an organism may be aware of a need to discard, and hence is able to consciously control it. Thus, you blow your nose, and cut your nails and hair, at times of your choosing. In your early years, your parents assist this. Even in later years, the result may be more satisfactory if another person is involved (e.g. chiropodist, hairdresser).

At adolescence gamete production begins. In early human communities the discardment of male gametes would usually have involved a sexual partner, who would have been unaware when she was discarding a female gamete. Today, by monitoring the small change in temperature that accompanies ovulation, a human female can know when she is discarding a gamete but, in the absence of medication (e.g. for contraception), she has no conscious control over the timing. In some species, the attention of the male (e.g. visual cues) provokes ovulation.

Selfish Genes and the Menopause

Human females also cannot consciously control the time of the discardment of the uterine cells that have proliferated in anticipation of the arrival of a fertilized ovum. In the absence of pregnancies (or medication), menstrual cycles continue until the menopause. Since natural selection generally favors those who produce most descendents, why is there a menopause? Biologists argue that human females will produce more descendents if, at around the age of 50, they discontinue gametogenesis and expend their energies in attending to the well-being of grandchildren. In other words, individuals with a ‘selfish’ gene that, directly or indirectly, causes arrest of gametogenesis in females, have tended, in the long term, to produce more descendents (who inherit that gene and thus the same tendency for females to discontinue gametogenesis) than individuals who do not have that gene.

This evolutionary trade-off is not so apparent in males, and whether there is a male menopause (e.g. whether being a good grandfather is of greater selective advantageous than continued procreation) is a subject of debate. The discardment of male gametes is usually under conscious control. In modern communities, prior to pair-bonding (legalized as marriage), human males usually discard gametes autonomously at a time of their own choosing. This may involve artificial visual cues (e.g. female images) rather than a partner. Thus, discarded male gametes usually do not prevent menstruation or expand populations. Like other exiled cells, discarded male gametes usually die. Whereas self-discharge of gametes (like menstruation) may have been rare in primitive communities, in modern communities (like other forms of contraception) it is the norm. Yet, over evolutionary time, the shaping of our sexual biology has been mainly influenced by the norms of primitive communities. Biologists argue that ‘selfish’ genes which, by some chain of events, cause us to frown on contraception (or abortion), must inevitably have increased in human populations. Thus, for most of us the crowded planet is now a more pressing reality than the crowded cytosol.

Molecular Chaperones

Hosts can generate complex multigenic systems for dealing with pathogens. However, pathogens, because of their need to replicate rapidly and disseminate, usually have smaller genomes and so are less able to encode complex systems to counter the increasingly sophisticated host-systems that can arise in an escalating arms race. Large viruses (e.g. with 200 kilobase genomes) have more countermeasures at their disposal than have small viruses (e.g. with 10 kilobase genomes). So the strategy of a large virus is usually different from that of a small virus.

The armamentarium of the host includes a set of proteins known as “molecular chaperones,” which include proteins induced by stresses such as virus infection or heat shock (Chapter 15). Some molecular chaperones play a normal role in cell operations, such as maintaining the structure of self-proteins in order to prevent inadvertent aggregation. This would permit fine-tuning of concentration to approach even closer to the threshold beyond which aggregation would occur. To prevent inadvertent aggregation there would also be a certain margin for error. Thus a self-protein would have to ‘stick its neck out,’ concentration-wise, before aggregating and triggering alarms that provoke apoptosis or peptide presentation to T lymphocytes.

Figure 16.5 summarizes some factors influencing the normal cytosolic concentration of a diffusible protein [19–23]. The leftward-pointing horizontal arrows symbolize the pressure, exerted by other cytosolic proteins, tending to reduce the solubility of the protein. If any particular cytosolic protein mutates in a way that does not affect its concentration (e.g. it may be impaired with regard to maintaining

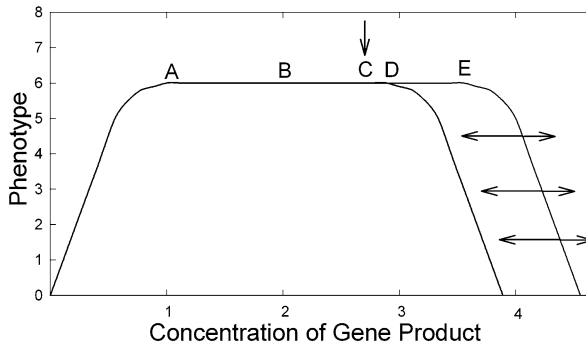


Fig. 16.5 Various factors affecting the concentration of a diffusible protein (e.g. enzyme) within the cytosol of a cell as revealed by a hypothetical dose-response curve. The curve shows some quantitative measure of phenotype (e.g. the color of a flower) as it is affected by increasing concentrations of the gene product that generates that phenotype. For example, flower color can depend on the rate of conversion of a colorless compound (enzyme substrate) to a colored pigment (enzyme product). This rate would be progressively increased by increasing the concentration of the enzyme (gene product) that catalyzes the conversion. If the concentration of the enzyme is directly related to the number of active gene copies (as often occurs), then the X-axis can be seen as providing an index of gene dosage (i.e. gene copy number). The phenotypic parameter (color measured on the Y-axis) increases with gene product concentration until point A when some other factor (e.g. substrate availability) becomes rate-limiting. The curve then plateaus. Increasing the amount of gene product (enzyme) now makes no difference to the color of the flower (no change in the value of Y).

In a diploid organism, B corresponds to the minimum gene dosage required to ensure that the phenotype would be unchanged in a heterozygote. The latter might have only one functional gene copy, and so there would be only half the concentration of enzyme (A) as exists in the homozygote (B). The phenotype would still correspond to the plateau of the curve. The color would be perceived as “dominant,” being no less in the heterozygote than in the homozygote (i.e. there would be “haplosufficiency”). E marks the concentration at which the gene product would still be soluble (i.e. the color would still be unchanged) *if no other proteins were present*. Above this concentration, enzyme molecules would tend to self-aggregate and the phenotypic parameter (Y-axis value) would decrease (i.e. color would fade).

D corresponds to the concentration at which aggregation would occur because of the presence of other cytosolic proteins that would collectively promote aggregation. The horizontal leftward-pointing arrows symbolize this aggregation pressure tending to push the descending limb of the dose-response curve to the left. Another factor, symbolized by the rightward-pointing horizontal arrows, would be the molecular chaperones (e.g. heat-shock proteins) that act to maintain protein conformation and thus decrease aggregation (i.e. tend to push the descending limb to the right).

Thus, it would seem that the concentration of a protein in cells of different members of a species could fluctuate between points B and D. It is likely, however, that over evolutionary time genes have ‘fine-tuned’ the concentrations of their products to a maximum consistent with avoiding self-aggregation. This point might correspond to C (marked by a vertical arrow), which is slightly to the left of D, thus providing a ‘margin of safety’ against inadvertent self-aggregation

its specific function, but not with regard to maintaining its concentration), then its ability to continue exerting this pressure is unaffected.

Molecular chaperones have the opposite effect to cytosolic proteins in general. The rightward-pointing arrows symbolize the pressure of molecular chaperones to increase the solubility of proteins. A particular type of molecular chaperone has a

subset of ‘client’ proteins. Since maintaining the solubility of proteins is a specific function of the molecular chaperones, when they mutate so that this function is compromised, then an important cellular defense against aggregation is removed. Individual protein species that might have just retained their solubility, now more readily cross the insolubility threshold. The proteins no longer have to ‘stick their necks out’ in order to aggregate.

Proteins that are perilously close to the insolubility threshold include mutant proteins which have sustained amino acid changes that affect their conformations, but not their specific functions. As long as they can, with the help of molecular chaperones, maintain their normal conformations, these proteins will maintain their distinctive primary functions. So no effects on the phenotypes of organisms with these mutations will be evident. However, when a molecular chaperone is mutated its client proteins now have to fend for themselves. Many proteins that are close to the insolubility threshold become more prone to aggregate. Mutant phenotypes that were previously hidden (latent) now emerge. They are “conditional mutants” – the condition being that their chaperone is not around to spruce them up. So solubilities (and hence functions) are no longer sustained.

Thus, a molecular chaperone can be seen as a mutational ‘buffer’ or ‘capacitor’ that allows organisms to survive certain types of mutation [24]. In the latter case an underlying hidden *genetic* change can be revealed by chaperone malfunction. This should not be confused with the “phenocopy” phenomenon, where an organism undergoing virus attack (or an equivalent stress) at a critical developmental stage might display mutant phenotypes, but there is *no* underlying genetic change; i.e. the product of a gene is, because of *environmental* factors, expressed at the wrong time or place, but the gene itself is unmutated [10].

On the other hand, viral attacks change the levels and modes of expression of heat-shock proteins. If some of these do not remain in chaperone-mode [25], then some of the observed phenotypes might not be phenocopies, but might reflect the exposure of unbuffered mutations. The expression of these mutations by future offspring would normally also be conditional on failure of chaperone function, but there are suggestions that sometimes mutant expression can switch from conditional to unconditional, so that the expression becomes a permanent characteristic of the line (a mysterious phenomenon known as “genetic assimilation”).

All this leads us to distinguish between cytosolic proteins in general, each of which has the potential to react weakly with erring members of a certain *small* subset of its fellow travelers [6], and molecular chaperones – ‘professional interactors’ – each of which has the potential to react strongly with erring members of a certain *large* subset of its fellow travelers.

Positive Repertoire Selection

Whatever the sophistication of the defense and attack systems of a host and its viruses, the usual initiating event is one of self/not-self discrimination, be it between two RNA species or between two protein species, or be it extracellular or intracellular (Chapter

15). Accordingly, a virus that could mutate to appear as self to its host should have an adaptive advantage. It would exploit the fact that during the development of lymphocytes, each specific for a particular antigen, negative selection of lymphocytes reactive with some self-antigens had generated gaps ('holes') in the repertoire. However, during development (e.g. creation of cytotoxic T-lymphocytes and antibody-forming B-lymphocytes) there is positive selection of those reactive with *near-self* antigens (positive repertoire selection) by mechanisms outlined elsewhere [17, 26–29]. The development of the repertoire of cytotoxic T-lymphocytes, for example, is strongly influenced by the polymorphic MHC proteins of the host. As a virus mutates progressively closer towards the self-antigens of a potential host, the chances that the host will be immunologically prepared become *greater* because its lymphocytes have been positively selected to react with near-self MHC proteins. In other words, the virus encounters stiffened host defenses. Thus, a host does not spread its resources too widely by trying to develop the potential to respond to the entire set of not-self antigens, formidable in range. Rather it anticipates viral strategy and focuses appropriately.

In this and the previous chapter we have considered immunology at the intracellular level in a way that is not found in immunology texts. For example, immunologists tend to use the term “altered-self,” rather than “near-self.” There is a fundamental difference between these usages. “Altered-self” implies a difference from self. “Near-self” emphasizes how close an entity can come to self, yet still be distinguishable from self. Immunologists also tend not to recognize the implications of the crowded cytosol for molecular interactions. But you should be warned that, while the association of peptides with MHC display proteins is well established, evidence on the underlying mechanism of the selective display is still fragmentary and controversial. The clonal selection theory in the 1950s was validated by demonstrations that individual lymphocytes each secreted a specific antibody (one cell, one antibody). But a subsequent immunological “thought collective,” ably described by Klaus Eichmann in 2008, led many to following false trails where “initial experiments yielded exciting results,” then the “evidence collapsed” and “interpretations became untenable” [30]. Immunologists remain wary [31].

Then why does a hypothetical mechanism have a place in an EB text? One reason, as has been stated, is that the full understanding of the genomes of a species requires an understanding of the genomes of the species with which it has coevolved. Species interactions involve processes of self/not-self discrimination, both extracellular and intracellular. Another reason is that understanding the role of intracellular protein concentration in self/not-self discrimination can make other evolutionary phenomena intelligible. It is in such terms that we will, in the next chapter, discuss “Muller’s paradox” and the mystery of sex chromosome dosage compensation.

Summary

The crowded cytosol is a special environment where weak interactions can be important. Here many macromolecules are close to the limits of their solubility, a condition conducive to weak, but specific, entropy-driven molecular interactions. In

addition to being under evolutionary constraint to preserve the functions of their own products, genes encoding specific cytosolic proteins are also under evolutionary constraint, both to support a pressure exerted collectively by proteins to drive other proteins from solution, and to maintain the solubilities of their own proteins in the face of that collective pressure. Thus, genes whose protein products occupy a common cytosol have co-evolved such that product concentrations are fine-tuned to a maximum consistent with avoiding self-aggregation. Cytosolic proteins collectively generate a pressure tending to drive proteins into aggregates. Each individual diffusible protein species both contributes to, and is influenced by, the pressure. Intracellular pathogens must fine-tune the concentrations of their own proteins to the solubility limits so imposed. Aggregates between viral proteins and normal host, antibody-like, ‘immune receptor’ proteins, provide a possible basis for intracellular self/not-self discrimination at the protein level. Molecular chaperones, including heat-shock proteins, modulate this process. In such terms, we can explain Goldschmidt’s phenocopy phenomenon, and paraneoplastic diseases. During their development there is positive selection of host lymphocytes reactive with near-self antigens (positive repertoire selection). This counters the tendency of pathogens to mutate towards self. Thus, hosts whose immunological forces are poised to attack near-self versions of not-self (a subset of not-self), rather than not-self per se (*the entire set* of not-self – formidable in range), are at a selective advantage.

References

1. Kipling R (1891) If. In: *Rudyard Kipling’s Verse. Inclusive Edition 1885–1918*. Copp Clark, Toronto (1919) pp 645
2. Fulton AB (1982) How crowded is the cytoplasm? *Cell* 30:345–347
3. Wainwright M (2003) Early history of microbiology. *Advances in Applied Microbiology* 52:333–355
4. Forsdyke DR (1995) Entropy-driven protein self-aggregation as the basis for self/not-self discrimination in the crowded cytosol. *Journal of Biological Systems* 3:273–287 [During their development from nucleated precursor cells, red blood cells lose their nuclei and become disc-shaped. The normal temperature of birds is higher than that of mammals, so entropy-driven reactions should be more evident. Avian red cells do not lose their nuclei, so are less disc-shaped and should less readily form rouleaux.]
5. Lauffer MA (1975) *Entropy-driven Processes in Biology*. Springer-Verlag, New York [In addition to entropy, “volume exclusion” may play a role in crowding phenomena; but see Breydo L, Sales AE, Frege T, Howell M, Zaslavsky BY, Uversky VN (2015) Effects of polymer hydrophobicity on protein structure and aggregation kinetics in crowded milieu. *Biochemistry* 54:2957–2966.]
6. Forsdyke DR (2001) Adaptive value of polymorphism in intracellular self/not-self discrimination. *Journal of Theoretical Biology* 210:425–434 [Other collective protein functions include redox buffering and control of ionic equilibrium across membranes (Donnan equilibrium).]
7. Jeffery CJ (2015) Why study moonlighting proteins? *Frontiers in Genetics* 6:211
8. Moreau-Aubry A, Le Guiner S, Labarrière N, Gesnel M-C, Jotereau F, Breathnach R (2000) A processed pseudogene codes for a new antigen recognized by a CD8+ T cell clone on melanoma. *Journal of Experimental Medicine* 191:1617–1623
9. Ingolia NT, Brar GA, Stern-Ginossar N, Harris MS, Talhouarne GJS, Jackson SE, Wills MR, Weissman JS (2014) Ribosome Profiling Reveals Pervasive Translation Outside of Annotated Protein-Coding Genes. *Cell Reports* 8:1365–1379

10. Goldschmidt R (1940) *The Material Basis of Evolution*. Yale University Press, New Haven, pp 266–271
11. Hickman HD, Luis AD, Bardet W, Buchli R, Battson CL, Shearer MH, Jackson KW, Kennedy RC, Hildebrand WH (2003) Class 1 presentation of host peptides following human immunodeficiency virus infection. *Journal of Immunology* 171:22–26
12. Blachère NE, Orange DE, Santomasso BD, Doerner J, Foo PK, Herre M, Fak J, Monette S, Gantman EC, Frank MO, Darnell RB. (2014) T cells targeting a neuronal paraneoplastic antigen mediate tumor rejection and trigger CNS autoimmunity with humoral activation. *European Journal of Immunology* 44:3240–3251
13. Lane C, Leitch J, Tan X, Hadjati J, Bramson JL, Wan Y (2004) Vaccination-induced autoimmune vitiligo is a consequence of secondary trauma to the skin. *Cancer Research* 64:1509–1514
14. Eckersley-Maslin MA, Spector DL (2014) Random monoallelic expression: regulating gene expression one allele at a time. *Trends in Genetics* 30:237–244
15. Borel C, et al. (2015) Biased allelic expression in human primary fibroblast single cells. *American Journal of Human Genetics* 96:70–80
16. Villaseñor J, Besse W, Benoist C, Mathis D (2008) Ectopic expression of peripheral-tissue antigens in the thymic epithelium: probabilistic, monoallelic, misinitiated. *Proceedings of the National Academy of Sciences USA* 105:15854–15859 [T cells must run the MEC ‘gauntlet’ before release from the thymus.]
17. Forsdyke DR (2015) Lymphocyte repertoire selection and intracellular self/not-self discrimination: historical overview. *Immunology and Cell Biology* 93:297–304
18. Heaman EA (2003) *St. Mary’s. The History of a London Teaching Hospital*. McGill-Queen’s University Press, Montreal, p 322
19. Shull GH (1909) The “presence and absence” hypothesis. *American Naturalist* 43:410–419
20. Forsdyke DR (1994) The heat-shock response and the molecular basis of genetic dominance. *Journal of Theoretical Biology* 167:1–5
21. Rajan RS, Illing ME, Bence NF, Kopito RR (2001) Specificity in intracellular protein aggregation and inclusion body formation. *Proceedings of the National Academy of Sciences USA* 98:13060–13065
22. Ciryam P, Kundra R, Morimoto RI, Dobson CM, Vendruscolo M (2015) Supersaturation is a major driving force for protein aggregation in neurodegenerative diseases. *Trends in Pharmacological Sciences* 36:72–77
23. Forsdyke DR (2012) Functional constraint and molecular evolution. *Encyclopedia of Life Sciences* John Wiley, Chichester
24. Sangster TA, Lindquist S, Queitsch C (2004) Under cover: causes, effects and implications of Hsp90-mediated genetic capacitance. *BioEssays* 26:348–362
25. Bersuker K, Hipp MS, Calamini B, Morimoto RI, Kopito RR (2013) Heat shock response activation exacerbates inclusion body formation in a cellular model of Huntington disease. *Journal of Biological Chemistry* 288:23633–23638
26. Forsdyke DR (1975) Further implication of a theory of immunity. *Journal of Theoretical Biology* 52:187–198 [This “affinity/avidity” model for positive repertoire selection is now generally accepted. An association between MHC proteins and antigens was implied, but it was a decade before the fine tailoring of antigens to create pMHC complexes came to light.]
27. Forsdyke DR (1991) Early evolution of MHC polymorphism. *Journal of Theoretical Biology* 150:451–456
28. Forsdyke DR (2005) “Altered-self” or “near-self” in the positive selection of lymphocyte repertoires. *Immunology Letters* 100:103–106
29. Vrisekoop N, Monteiro JP, Mandl JN, Germain RN (2014) Revisiting thymic positive selection and the mature T cell repertoire for antigen. *Immunity* 41:181–190
30. Eichmann K (2008) *The Network Collective. Rise and Fall of a Scientific Paradigm*. Birkhäuser, Basel

31. Forsdyke DR (2012) Immunology (1955–1975): the natural selection theory, the two signal hypothesis and positive repertoire selection. *Journal of the History of Biology* 45:139–161. [Many immunologists had also forgotten the early 20th century studies on innate immunity by Almroth Wright. (See: “Almroth Wright, opsonins, innate immunity and the lectin pathway of complement activation: a historical perspective.” (Forsdyke DR, 2016; *Microbes & Infection*, in press).]

Part VI
Sex and Error-Correction

Chapter 17

Rebooting the Genome

What makes hybrid male sterility of great current interest is the increasing evidence that the building blocks of this isolating barrier may be radically different from what we had come to believe. ... It is clear that a new paradigm is emerging, which will force us, first, to revise many conclusions ... that had gathered almost unanimous agreement, and, second, to try a completely different experimental approach.

Horacio Naveira and Xulio Maside (1998) [1]

With the notable exceptions of Butler and Miescher, in the nineteenth century the information concept as applied to biological molecules did not extend to information error and the need for its detection and correction. Miescher in 1892 thought that: “Sexuality is an arrangement for the correction of these unavoidable stereometric architectural defects in the structure of organized substances” (Chapter 3). While referring to “left handed coils” being “corrected by right-handed coils,” at that time he was unable to relate this to “nuclein,” a new substance he had discovered, later known as DNA. However, he appreciated that correction would require some sort of yardstick (i.e. “right hand coils”) to permit the fact of error in a molecule (i.e. “left hand coils”) to be detected, and then appropriately corrected.

If the various degrees of redundancy found among DNA molecules have a single explanation, it is error-detection and correction. This might explain another apparent redundancy, the fact that members of most biological species are either one of two sexes. Yes, this certainly makes our lives more interesting. But the biological advantage is not obvious. Imagine a world without males in which each woman, on average, was able to produce two offspring per generation asexually. Both of these would be female. In the first generation there would be two women. In the second generation there would be four women. In the third generation eight women, and so on. With a similar limit to offspring number, in a sexual world a woman would be likely to produce one male and one female per generation. Only the latter, on being fertilized by a male, would produce further offspring. In the first generation there would be two individuals, male and female. In the second generation, there would still be two, and so on to the third generation, etc. Since the winners in the struggle for existence are organisms that leave the most offspring, sex would seem very disadvantageous [2].

Redundancy

At the outset the two-fold parallel redundancy of naturally occurring DNA duplexes, which contain complementary ‘top’ and ‘bottom’ strands, was recognized as revealing not only how DNA was replicated, but also how it might be corrected. Changes in DNA, manifest as unusual bases (e.g. **U** instead of **T**), or damaged bases (e.g. cross-linked **Ts**), would have provided a selection pressure for the evolution of specific enzymes that could recognize and correct errors. Indeed, many such enzymes were discovered [3].

Sometimes, however, a normal base can change into another normal base (“base substitution”) so that an unusual base-pair results (e.g. **A** on one strand opposite a **C** on the other strand). Again, one can envisage a selection pressure for the evolution of enzymes that *recognize* base mispairing; but after the fact of recognition, how is it known which base to correct (Chapter 2)? Take the following duplex:



Consider a base change (transition mutation) from **T** to **C** (underlined):



The error might be *corrected* as:



Or *compounded* as:



There is only a 50% chance of actual correction. And there is another actor in the wings. While the ‘rapid response team’ of error-correcting enzymes is alert for base mispairs, and is quick to move into action, one can hear in the distance the whistle and thunder of an approaching DNA polymerase further up the line. If there is any delay, sequence 17.2 may be replicated before it can be corrected. Two new child duplexes (like 17.3 and 17.4) will be produced. One will not contain the error (unmutated duplex). One will contain the error (mutated duplex). In this case there is a 100% chance that there is no correction, at least in the short term. After replication there is no base mispairing in either duplex, so there is no indication, at the duplex level, as to which duplex is the mutant. On partitioning of chromosomes to child cells (segregation), one child cell might malfunction, and on this basis the mutant phenotype might be selected against. To this extent the mutation would be detected and corrected.

Assuming that the error-correcting team will usually arrive before suspicious DNA is replicated, the most obvious way to improve on the 50% correction success rate is to provide the team with another duplex to compare. If sequence 17.2 could be compared with sequence 17.1, the team would ‘know’ that the A-C mispairing in 17.2 should be corrected to an A-T Watson-Crick base-pair, rather than to a G-C Watson-Crick base-pair. Thus, a further level of redundancy is needed, either “in parallel” (i.e. another independent duplex), or “in series” (i.e. repetition further down the line).

Diploidy, in theory, should solve the problem. However, the system is not perfect as indicated by the need to race to correct errors before DNA polymerase can compound them and turn potentially correctable mutations into what may eventually become “accepted mutations.” The latter can be detected either experimentally, by those who sequence DNA molecules, or biologically, when there is an observable change in phenotype. But unless the environment is changing rapidly, most mutations are deleterious, so there may be no genotypes or phenotypes around for us to examine. Thus, what is commonly alluded to as *the* “mutation rate,” may be less than the actual rate.

In humans, if deleterious mutations affect only somatic tissues and occur after reproductive life, then from the viewpoint of the species there is no major problem (the benefits of infant care by grandparents notwithstanding). On the other hand, mutations that occur in germ-line tissue before the end of reproductive life might result in defects being passed to future generations. Meiosis appears as a special germ-line process with a low error rate that might both decrease germ-line transmission of mutations, and repair DNA damage. It is the nearest we have to ‘rebooting’ the genome.

Recombination Repair

For recognition and correction of a mutation there must be a template for comparison. Accordingly, you are diploid, and were generated from haploid paternal and maternal gametes that had themselves been subject to meiotic cleansing. You, in turn, can meiotically generate fresh gametes as a rich blend of your parental genomes, due to (i) the random assortment of (i.e. creation of various combinations of) 23 entire paternal and 23 entire maternal chromosomes (the only constraints being that 23 are allowed in a haploid gamete, and there must be one member of each homologous pair), (ii) the random assortment of (i.e. recombination among) parts of each chromosome pair, and (iii) recombination repair (Fig. 11.2). However, the effectiveness of the latter correction will depend on the templates. There are three template-critical alternative outcomes:

- (i) If your parents were closely related (e.g. incest), then it is likely that you were born with many identical mutations at identical positions in all chromosome pairs (i.e. you are homozygous for these mutations). So the fact of mutation may not be detectable. But the fact that you exist shows that these mutations can be non-lethal when homozygous.

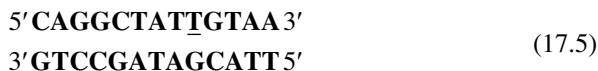
- (ii) If you are the result of out-breeding (i.e. your parents were not closely related), then there will be far fewer shared mutations in homologous chromosome pairs, so that many individual mutations can be identified. For example, one of your parents might have donated sequence 17.3 (unmutated) and your other parent might have donated sequence 17.4 (mutated). Thus, you would be heterozygous with respect to the one base pair difference between these sequences. One 'good' copy of a gene containing that sequence might suffice (i.e. you would be "haplosufficient" and therefore likely to be viable as an individual; Fig. 16.5). However, when the two sequences were compared at meiosis as candidates for promotion to gamete status, the fact of heterozygosity might be recognized. Recombination repair might convert one of the sequences to that of the other ("gene conversion;" Fig. 11.2). There would be a loss of heterozygosity and a gain of homozygosity. Without guidance, the conversion would have, on average, a 50% success rate (since which was the 'good' copy would not be known).
- (iii) Finally, if you are the result of extreme outbreeding the definition as "template" becomes spurious. In one position in the top strand, instead of either **A** (sequence 17.3), or **G** (sequence 17.4 containing an **A** to **G** transition mutation), there might be a **C** or **T** (i.e. a transversion mutation with the purine mutated to a pyrimidine). A few differences of this nature might be tolerated. But many such differences, be they transitions or transversions, will result in a failure of pairing, so that meiosis fails, no gametes are formed and, although you may not be aware of it, you are a sterile hybrid (a 'mule;' Chapters 9 and 10).

The second of the above three alternative outcomes, a consequence of parental out-breeding, appears the most advantageous. But if there is only a 50% gene conversion success rate with respect to base-pair heterozygosity at any *one* position, then the *overall* success rate for multiple heterozygous positions will remain at an average of 50%. Unless there is some guidance as to which parental contribution to your gametes is more advantageous *for each mutant position*, your average gamete will be no better off than you are.

However, 50% of your gametes will, by chance, emerge with more 'good' base pairs than you have. And 50% of your spouse's gametes will, likewise, be superior. Thus, if you have several children, some of these will be superior to you and your spouse, with respect to the function which Nature holds supreme, namely the ability to produce many, reproductively fit, children. Accordingly, over the generations, combinations of 'good' genes should be selected. This would be emphasized if the positive effects of 'good' genes were more than additive (i.e. genetic fitness would increase exponentially as the number of 'good' genes increased; "synergistic epistasis"). Similarly so, if the negative effects of 'bad' genes were more than additive, such genes would tend to be purged collectively from the population.

Strand Guidance

Depending on the type of mutation, sometimes there is guidance as to which copy of a gene is best. For example, there might be a transition micromutation from **C** to **T** in sequence 17.1, resulting in:



Of importance in this case is that the **C** was part of a **CG** dinucleotide sequence in the top strand, which must base pair with a complementary **CG** dinucleotide sequence in the bottom strand (recall that **CG** is read with 5' to 3' polarity, and the two strands are antiparallel; Fig. 2.4). Now for a bit of chemistry. The **C** in **CG** dinucleotides is sometimes changed (methylated), but the changed base – methylcytosine – is accepted in DNA. (So there are often five, not four, bases in the natural DNA alphabet; see Chapter 18). **C** is also sometimes chemically changed (deaminated) to the base **U**. The **U** is recognizably wrong, since **U** does not occur in DNA. But when a methylated **C** is deaminated the resulting base is methyluracil, otherwise known as thymine (**T**). This is, of course, a natural base.

When encountering a base mismatch in duplex DNA such that **T** in one strand pairs with **G** on the other, the rapid response team of error-correcting enzymes can check if the latter **G** has a **C** on its 5' side. Better still, it can determine if the **C** is methylated. It then has the information to deduce the high probability of a **CG** dinucleotide on the other strand, so that the error lies in the **T**, not in the **G**.

Nevertheless, the most secure strand guidance would be to have a sequence redundancy greater than diploid (Chapter 2). A diploid cell, prior to engaging in mitotic, or meiotic, cell division, replicates its DNA in a period referred to as the S-phase of the cell cycle. For a period after this (referred to as G2 phase of the cell cycle) the cell is actually tetraploid with four DNA duplexes (Fig. 17.1).

After cell division (named “mitosis” because the thread-like chromosomes become visible; Greek: *mitos* = thread), the cell will return to the diploid state. If a mutation is accepted and remains unrepaired in the period referred to as the G1-phase, it will be faithfully replicated in S-phase, and there will then be two identical mutant DNA duplexes to be repaired in the G2 phase. The relative proportions of mutated and unmutated duplexes will be unchanged. On the other hand, if the mutation occurs *after* DNA replication (i.e. in the G2 phase) then there will be three DNA duplexes against which the mutant duplex can be compared, so in this case there can be guidance. It is perhaps for this reason that some genes, chromosomes, and cells spend more time in the G2 phase of the cell cycle. Genes that are replicated early in S-phase remain longer in G2 phase and have less accepted mutations [4]. Indeed, meiotic oocytes remain suspended in G2 for many years prior to fertilization (see later). There is a ‘selfish gene’ perspective to this – ‘assassin genes’ overcoming others – that is not discussed here [5].

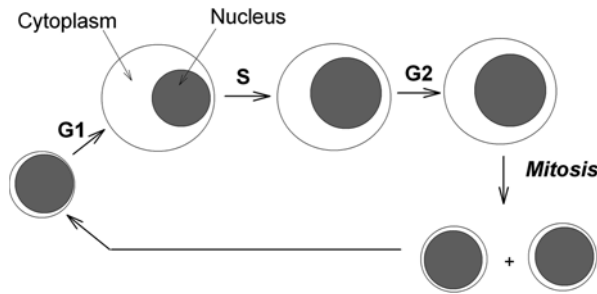


Fig. 17.1 Phases of the cell cycle. After normal cell division (mitosis), child cells (diploid with two copies of each chromosome duplex) may enter the ‘first gap phase’ (G1-phase), in which there is cytoplasmic growth, but little growth of the nucleus where the chromosomes are located. In S-phase, DNA is replicated and the nuclear volume approximately doubles in size. In the ‘second gap phase’ (G2-phase) the cell is tetraploid with four copies of each chromosome (i.e. four homologous DNA duplexes). Active genes (e.g. ‘house-keeping genes;’ see Chapter 18), and active chromosomes (e.g. the active X chromosome in human females), tend to be replicated *early* in S-phase. This means that they have shorter G1 phases, and longer G2 phases (i.e. they are longer in the tetraploid state), than inactive genes and chromosomes

Since mutation is essentially a random event, in a multicellular organism it should strike one duplex in one cell at one time. This means that a diploid cell, with an accepted somatic mutation in one of its two duplexes, could seek help from another diploid cell to verifying which of its duplexes was mutated. This would involve transfer of nucleic acid information between cells, a process that, if it occurred between single cells organisms, would be referred to as sexual. There is some evidence that such a process can occur between somatic cells in a multicellular organism [6].

In meiosis a single cell with two DNA duplexes (of paternal and maternal origin) first replicates to generate four DNA duplexes and then divides twice (producing four cells) with no intervening phase of DNA synthesis (Fig. 9.6). In human male meiosis four haploid gametes result. In human female meiosis one haploid gamete results and three are discarded. Thus, in female meiosis there may be a further opportunity for gamete quality control (see later).

Nevertheless, through the process of sexual reproduction, a given duplex finds itself, from generation to generation, in a succession of bodies each time in the company of a new homologous duplex. A mutant duplex (e.g. sequence 17.2) may never be in a situation where it can *simultaneously* be compared within a cell with three corresponding unmutated duplexes (e.g. sequence 17.1), but it can *sequentially* achieve this result over the generations. If in every generation duplex 17.2 repeatedly encounters duplex 17.1, and never another duplex 17.2, then even with a 50% probability of successful gene conversion on each occasion, a likely outcome will be elimination of the mutation. This process would be assisted if individuals with duplex 17.2 were negatively selected.

The process would also be assisted if, as pointed out by Virgil Reese in the USA [7], there were some way of *registering* an ambiguous region as *suspect* – perhaps using an epigenetic tag that could be transmitted through the germ line for a few

generations (see Chapter 18). Such tagged regions, perhaps as ‘recombination hotspots,’ would initiate the *directional* gene conversion process that, since it can favor the *non-initiating* DNA strand [8], would eliminate the tag and destroy the hotspots, which would then be succeeded by fresh recombination hotspots at other locations [9, 10].

DNA Damage

An ‘error’ in DNA must be understood in context. Of duplex sequences 17.3 and 17.4 above, one may have good and one may have bad consequences for an organism. On the other hand, sequence polymorphism is a natural phenomenon (see Part V). In some circumstances, neither sequence would be inappropriate. DNA damage, however, is always an error. As such it must be “corrected” or “repaired.” Strictly speaking these words have the same meaning but, although there is no generally agreed usage, in the literature incorrect base substitutions (micromutations) tend to be “corrected” and damage tends to be “repaired.” However, correction by recombination, whether it be of an incorrect base substitution, or of damage, has long been referred to as “recombination repair.” Repair of damage can be imperfect, resulting in mutation, or mutation can occur afresh without prior damage.

With damage there is likely to be little ambiguity regarding recognition, and if the damage affects just one strand then there is likely to be no problem with strand discrimination. So repair could be made on the basis of the information in the complementary sequence in a single duplex (i.e. repair could be effected in a haploid cell). If both strands are damaged (e.g. complementary base-pair damage), then repair must be made on the basis of information present in a homologous duplex that may be either in-series (i.e. sequence duplication) or in-parallel (i.e. diploidy).

One hypothesis of aging is that the selection pressure for efficient damage repair has not been high over evolutionary time because organisms have usually died early from environmental causes (e.g. hunger, predation, disease), rather than from DNA damage. To continue their line, individuals have had to reproduce *before* being struck by these apocalypses. So their offspring have tended to inherit traits that enabled them to deal with these, rather than with the ravages of time. Natural selection has favored traits that are manifest early in life. Butler gave us a glimpse of this [11]:

If heredity and memory are essentially the same, we should expect that no animal would develop new structures of importance after the age at which its species begins ordinarily to continue its race; for we cannot suppose offspring to remember anything that happens to the parent subsequent to the parent’s ceasing to contain the offspring within itself. From the average age, therefore, of reproduction, offspring should cease to have any further steady, continuous memory to fall back upon; what memory there is should be full of faults, and as such unreliable.

Thus, in modern populations where apocalypses have greatly decreased, the phenomenon of aging, possibly due to a failure to repair DNA damage, has emerged. The damage would affect the transcription and replication of DNA, with phenotypic consequences that lead to death [3].

Although the disposable soma (that's you and me) may not have highly efficient DNA damage repair systems, such efficient repair is necessary for gamete production in the gonads. Here recombination repair would seem to constitute the last court of appeal, with a sequence in one duplex coming to the rescue of the sequence in the homologous duplex. Here again, the templating quality of the donating duplex should be considered. If meiosis worked efficiently in your parents' gonads, you initially inherited undamaged parental DNA sequences. So most of the damages present in your gonads are likely to have been incurred during your lifetime, and are unlikely to be at the same positions in the paternal and maternal haploid contributions to your diploid genome. Thus, the fact that your parents might have been closely related might not be important as far as the repair of DNA damage is concerned. In this respect, parental outbreeding might not necessarily be beneficial. However, DNA damage might summate with the intergenomic differences accompanying extreme outbreeding, so that meiosis would more likely fail (i.e. producing hybrid sterility) than in a genome with less damage.

Note that we have been considering correction and repair at the level of classical Watson-Crick *duplexes*. Here a **G-T** mismatch would be recognized and corrected. Yet **G-T** pairings occur normally in extruded DNA stem-loop structures (Chapter 5). Correction of such weak pairings (i.e. formation of a normal **G-C** or **A-T** pair instead of a **G-T** pair) would increase stem-loop stability, but generate mutations in the corresponding classical duplexes. This might have phenotypic consequences. So it can be presumed that, under biological conditions, error-detecting and correcting enzymes confine their activities to classical duplexes. **G-T** pairings in stem-loops are ignored. Alternatively such **G-T** pairings are repaired with possible phenotypic consequences, or mutations in the neighborhood that change local stem-loop configurations to avoid **G-T** pairs are accepted [12].

Piano Tuning and Calibrated Guessing

From the studies of Wada and others (Chapter 11), individual genes emerge as distinct (**G+C**)% entities (microisochores), each with a distinct potential vibrational frequency. Hence, it seems not unreasonable to regard individual genes much as we regard the individual strings of a musical instrument. A gene with a mutated or damaged base is a gene 'out of tune' to the extent that the change may marginally change its vibrational frequency. We should consider the possibility that this has not escaped Nature's notice. Perhaps, there are 'tuning' enzymes that can check for changes in resonance, and hence supplement the methods of accurate detection of errors based on direct templated interactions between DNA sequences as has been considered here.

Guidance as to which strand is responsible for a mismatch can be *local*, as in the case of **CpG** sequences. Error-correcting enzymes could have evolved to utilize the regularity with which **CpGs** in the top strand match **CpGs** in the bottom strand, to 'guess' which strand had the incorrect base (see above). Furthermore, if it 'knew' it was in a protein-coding region, an enzyme might 'assume' purine-loading (Chapter 6) and,

when in doubt, hope to correct accurately with a purine in the top strand and with a pyrimidine in the bottom strand. Another enzyme might even be attuned to codon positions and make appropriate replacement guesses assuming the **RNY** rule (Chapter 7).

Such calibrated guessing should also be possible based on more *dispersed* genomic characteristics. In the absence of more precise information, an enzyme that could sense a high genomic (**G+C**)% [13] would 'guess' that a mismatched **A** or **T**, should best be replaced with a **G** or **C**. And there is the astonishing precision with which genomes obey Chargaff's PR2 (see for example the composition of vaccinia virus in Chapter 4). There may exist enzymes that can summate base frequencies and, assuming PR2 to apply, make appropriate corrections [14]. Beyond Chargaff, there are other mathematical regularities in base compositions that could play similar roles [15].

Spot Your Reprotype?

In general, outbreeding is beneficial producing "hybrid vigor." But extreme outbreeding could result in disparities in base composition leading to hybrid sterility [16]. From the time of Darwin [17], behavioral studies have suggested how prospective mating partners might have evolved genes inducing courtship displays that would subconsciously allow assessment of the likely vigor of the resulting children should courtship lead to mating. In this context, base compositional differences (i.e. differences in "reprotype;" Chapter 10) might seem to be below the threshold of detectability by prospective mating partners.

On the other hand, intragenomic conflicts between the pressures on the conventional Darwinian phenotype, and the pressures on the genome phenotype, are sometimes won by the latter (Chapters 12 and 13). Thus, the genome phenotype *can* affect the conventional phenotype. In this respect the arrow denoting classical Darwinian natural selection (from environment to phenotype to genotype) is, in principle, reversible. It is thus theoretically possible, although so far without experimental support, that an organism could, as part of its conventional phenotype, *display* its reprotype – its (**G+C**)% (i.e. signal its (**G+C**)% to prospective mates). The selective advantage of this (in terms of reproductive success) should be considerable, since extremes of inbreeding and outbreeding could be avoided. Thus, hybrid sterility may not be just something that is encountered on the roulette wheel of life. "Your face" may indeed be "your fortune," or rather, your children's fortune.

Haldane's Rule

The advantages of outbreeding argue strongly for the advantages of sexual reproduction, as opposed to asexual reproduction (essentially extreme inbreeding). Up to the hybrid sterility limit, the more disparate the genomes, the better. As discussed in

Part III, the cycle of reproduction among members of a biological species will proceed continuously, unless interrupted at some stage (Fig. 8.3). A likely initial stage for cycle interruption (i.e. a potential origin for speciation) is when disparate haploid parental genomes meet in diploid meiosis and cannot ‘agree’ to pair. Gamete production fails and the organism, a hybrid of paternal and maternal genomes, is sterile (Fig. 8.4). Turning a diploid into a tetraploid can ‘cure’ hybrid sterility since each chromosome at tetraploid meiosis has a pairing partner (Chapter 10).

Remarkably, diploid sterility is not gender neutral. As parental genome sequences get progressively disparate, the proportion of children that become sterile adults increases. But initially it is the sex that produces two types of gametes (the heterogametic sex) that is sterile. Among humans, males are the heterogametic sex, so boys are sterile and girls are fertile. This regularity was recognized by the British polymath J. B. S. Haldane [18]. Because of its applicability to a wide range of organisms (e.g. birds, insects, mammals, nematode worms, plants), the phenomenon has the potential to tell us much about sex and speciation [19]. However, “Haldane’s rule” is really two rules. Haldane’s rule for hybrid *sterility* should be clearly distinguished from Haldane’s rule for hybrid *inviability*. Diploid inviability is also not gender neutral. It is again the heterogametic sex that is preferentially affected.

A normal human male produces two types of spermatozoa in equal quantities – spermatozoa with an X-chromosome that will confer femaleness on children, and spermatozoa with a Y-chromosome that will confer maleness on children. A new organism is created when one of these spermatozoal types fertilizes a female gamete, which is always an X-bearing gamete. As discussed in Chapter 4, if X-bearing spermatozoa and Y-bearing spermatozoa are in equal quantities, there will be approximately equal chances of a boy (diploid sex chromosome status XY) or a girl (diploid sex chromosome status XX). But if the genes donated to the zygote by the paternal and maternal genomes cannot work together, then there may be a failure in development, so that at some stage zygote-derived cells fail to multiply and differentiate into a viable organism (Fig. 8.4). This “hybrid inviability” usually reflects incompatibilities between gene products and, for reasons discussed elsewhere, preferentially affects the heterogametic sex [20]. This means that boys are either not born, or fail to develop into mature adults. Since only adults produce gametes, *the question of the boy’s sterility does not arise*, except in the rare cases when incompatible gene products are *solely* concerned with gamete formation.

If a human female has a tendency towards hybrid sterility (i.e. the two parental genomes she carries are extremely disparate), then there will be decreased production of X-bearing ova. But, although it is likely that there will be fewer children, there is no bias towards boys or girls *by virtue of her sterility*. If a male has a tendency towards hybrid sterility (i.e. the two parental genomes he carries are extremely disparate), then there will be a decreased production of Y-bearing spermatozoa and X-bearing spermatozoa. But again, although it is likely that there will be fewer children, there will be no bias towards boys or girls *by virtue of his sterility*. All that Haldane’s rule for hybrid sterility implies is that *the tendency towards sterility is greater in males*. To understand this, we will first consider sex chromosomes and their role in the anti-recombination process that can lead to speciation.

Sex Chromosomes

In many species males and females are produced in equal quantities. Mendel himself noted that such equal quantities would be produced if a recessive homozygote were crossed with a heterozygote. Thus, if red is dominant to white, when a homozygous white (WW , producing one type of gamete, W and W) is crossed with a heterozygous red (RW , producing two types of gamete, R and W), on average, equal numbers of red and white offspring should be produced (RW , WW) (Fig. 8.1). If sex were similarly determined, this simple scheme would suggest that one sex be homozygous and the other sex be heterozygous for alleles of a particular gene.

As long as only *one* allele pair was required, there would be no reason to regard this process as different from any other genetic process. The chromosome pair containing the gene – let us call them X and Y chromosomes – would be equal in all respects, including size (“homomorphic”). One sex would be homozygous for the recessive allele (the “homogametic” sex) and the other sex would be heterozygous (the “heterogametic” sex). If there were recombination between the chromosome pair in the heterogametic sex, the gene might switch chromosomes, thus converting an X chromosome to a Y chromosome and the corresponding Y chromosome to an X chromosome. The *status quo* would be preserved and equal numbers of differentiated gametes would still be produced.

However, if the complexities of sexual differentiation were to require more than one gene – say genes X_1 and X_2 on the X chromosome, and the corresponding allelic genes Y_1 and Y_2 on the Y chromosome – the situation would get more complicated. Recombination (crossing-over between chromosomes at meiosis) might then separate the two genes to generate chromosomes (and hence gametes) with genes X_1 and Y_2 together, and genes Y_1 and X_2 together. Sexual differentiation would be impaired (Fig. 17.2). To prevent this, either the gene pairs – X_1 and X_2 , and Y_1 and Y_2 – would have to be closely linked on the X and Y chromosomes, respectively (so the chance of members of a pair being separated by meiotic recombination would be low), and/or one of the chromosomes would have to develop some local mechanism to *prevent* such recombination.

But if this *anti-recombination activity* could not easily be localized, then the activity might spread to involve other genes on the chromosomes; these genes might themselves play no role in sexual differentiation. This seems to be the situation that often prevails. The needs of the predominant function – sex determination – overrule, but do not necessarily eliminate, the functions of other genes on the same chromosome. The chromosomes are referred to as the ‘sex chromosomes,’ even though concerned with many functions not related to sex. However, we should not forget that anti-recombination is the essence of the initiation of speciation (Part III). The prevention of recombination *between* members of different incipient species, or of an incipient species and the parental species, can be fundamental for the speciation process (“reproductive isolation”), and can involve all the chromosomes, including the sex chromosomes [16].

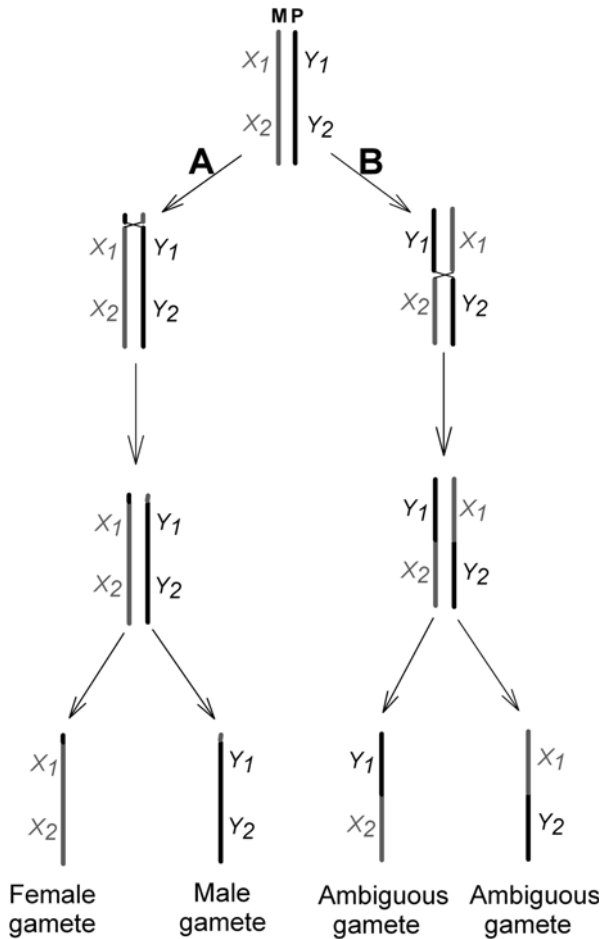


Fig. 17.2 Need for meiotic anti-recombination activity when multiple genes determine sex. In a human male (top), the maternally-donated sex-determining chromosome (M; gray) has a pair of sex-determining genes X_1 and X_2 , and the paternally-donated sex-determining chromosome (P; black) has a pair of sex-determining genes Y_1 and Y_2 . If path A is followed at meiosis, there is an exchange of chromosome segments (recombination) that does not disturb the linkage of each pair. The haploid female gamete (i.e. the gamete conferring the female state) has part of the original paternally-donated chromosome, but X_1 and X_2 remain linked (bottom). The haploid male gamete (i.e. the gamete conferring the male state) has part of the original maternally-donated chromosome, but Y_1 and Y_2 remain linked. If path B is followed at meiosis, the exchange of chromosome segments disturbs the linkage, so that each resulting haploid chromosome (at bottom) has both a female sex-determining gene and a male sex-determining gene

On the other hand, sex chromosomes exist *within* a species, and anti-recombination during meiosis is only beneficial to the extent that it prevents recombination between genes specifically concerned with sexual differentiation. When anti-recombination activity spread along sex chromosomes to genes not concerned with sexual differentiation, the latter lost the benefits of recombination.

To the extent that meiotic recombination allowed recombination repair, then, when such repair was lost, non-lethal mutations remained uncorrected. There was no longer prevention, *within* species, of changes in sex chromosomes, including micromutations (base substitutions, additions, or deletions) and macromutations (segment additions, deletions, transpositions and inversions). These changes are similar to those occurring *between* species as they progressively differentiate their chromosomes [21].

Thus, instead of remaining the same size (homomorphic), the sex chromosomes sometimes came to differ in size (heteromorphic). In some species (e.g. certain fish and amphibia) the sex chromosomes remained homomorphic and appeared to differ only in the genes affecting sexual differentiation. In other species (e.g. mammals, birds) the sex chromosomes became heteromorphic.

Sex and Speciation

Sexual and species differentiations can involve anatomical and physiological changes of similar orders. When classifying organisms into species, two forms initially considered as distinct species have sometimes been found to be merely the sexually differentiated forms of one species [22]. An extreme case of this was Darwin's realization that an apparent member of a parasite species found within the body of a female barnacle was actually the male form of that species ("complemental males") [23]. Bateson and Goldschmidt both drew attention to the possibility of a fundamental similarity between the evolutionary process that divides two groups within a species, such that they become two species, and the evolutionary process that divides two groups within a species, such that they become two sexes (Chapters 8 and 9). When the chromosomal basis of sexual differentiation became clearer, Bateson in 1922 commented on the chromosomally-borne "ingredients" responsible for this in plants [24]:

We have now to admit the further conception that between the male and female sides of the same plant these ingredients may be quite differently apportioned, and that the genetical composition of each may be so distinct that the systematist might, without extravagance, recognize them as distinct specifically [i.e. mistake the different sexes as distinct species]. If then our plant may ... give off two distinct forms, why is not that phenomenon [i.e. sexual differentiation] a true instance of Darwin's origin of species?

In 1940 Goldschmidt, who had postulated that the initiation of speciation requires a general change in a chromosome's "reaction system," noted [25]:

Another remarkable problem ought to be pointed out ..., though it seems at first sight to be rather remote from the problem of species formation. ... Sexual differences within a species may be of such a nature that, if found distributed among different organisms, they would provide a basis for classification into different species, families, or even higher categories. These differences frequently touch upon practically every single character of the organism, morphological and physiological. Two forms found in nature, which showed sexual differences of such degree ... would never be considered as belonging to the same species In the sexual differences we have, then, two completely different reaction systems in which the sum total of all the differences is determined by a single genetic differential... . The genetics of sex determination ought therefore to furnish information on how a completely different reaction system may be evolved.

Goldschmidt made the distinction between “reaction system” and genes explicit:

It is not this or that gene or array of genes which is acting to produce the extreme morphogenetic differences of the sexes, but rather the typical serial pattern within the X-chromosome, or definite parts of it. The chromosome as a whole is the agent, controlling whole reaction systems (as opposed to individual traits). The features which are assumed by many geneticists to prevent a scattering of individual sex genes by crossing over ... actually prevent major changes of the pattern within the chromosome as a whole. Once more I must emphasize that such a conception offers mental difficulties to those steeped in the classical theory of the gene.

Thus, species differentiation and sexual differentiation have an important feature in common. For their maintenance there must be an absence of meiotic recombination, either between one or more chromosome pairs in *both* sexes (species differentiation), or between regions of one chromosome pair in *one* sex (sexual differentiation). For speciation, chromosomes of members of a variant group must not recombine with those of the parental line, otherwise character differences due to multiple genes could blend, and the differentiation would be lost. For sexual differentiation, regions of the sex chromosomes must not recombine because the characters conferring sexual identity could blend, and sexual differentiation would then be lost. The sex chromosomes (e.g. X and Y in humans) must, between themselves, maintain something akin to the reproductive isolation that defines organisms as members of distinct species.

Haldane’s rule for hybrid sterility usually concerns crosses between members of different “races” (varieties, breeds, lines) *within* what appears as a single species (defined reproductively) [26]. We are accustomed to recognize as “races,” groups within a species the members of which show some conspicuous anatomical differentiation from the members of other groups. However, in principle, members of a group (“race”) within a species might *begin* to differentiate with respect to reproductive potential (i.e. no longer retaining full fertility when crossed with parental stock) *before* any anatomical or other physiological differences are evident (cryptic differentiation) [27, 28]. When fully differentiated in this way, the members of the group would be reproductively isolated, and so would constitute a new species, even if not anatomically or functionally distinguishable from members of the parental group.

Thus, Haldane’s rule for hybrid sterility has no obligatory anatomical or physiological correlates, and is a phenomenon of incipient speciation *alone*. It seems likely that the preferential sterility of the heterogametic sex can be regarded as a *step*, or *way station*, on the path towards the complete sterility associated with full species differentiation, a process generally accompanied by the anatomical and physiological differentiation of conventional phenotypic characters. Intriguingly, when trying to impress upon Darwin, that hybrid sterility could be of positive selective value (Chapter 9), Wallace (1868) had some feel for this: “The ... difficulty, of unequal sterility of reciprocal crosses, - seems none to me; for it is a step to more complete sterility, and as such would be useful and would be increased by selection” [29].

The sex chromosomes tend to progressively differentiate from each other, and in some species meiotic recombination remains possible only in small regions where there is sequence homology, which *includes* (G+C)% compatibility. In this respect the small regions behave like the autosomes, and are referred to as “pseudoautosomal”

regions. Other parts of sex chromosomes (the non-pseudoautosomal regions) appear to be kept from meiotic pairing and recombination by a failure of homology [30, 31]. However, recombination repair is still possible when there is *in-series* sequence redundancy (Chapter 2). Indeed, the human Y chromosome contains extensive palindromic duplications that should facilitate intra-chromosomal non-meiotic gene conversion [32, 33]. Simply stated, the Y non-pseudoautosomal region and the corresponding X chromosome cannot ‘have sex’ (recombine) *with each other*. The X chromosome can redeem itself *in future generations* (due to *in-parallel* sequence redundancy; diploidy) whenever it finds itself accompanied by another X chromosome (i.e. when in a female body). The Y chromosome redeems itself *at any time* in male bodies, using duplicated segments *within itself*. This strategy is also followed by certain organisms that have abandoned meiosis and are rich in long palindromes that incorporate many genes [34]. Indeed, so effective is the strategy in bdelloid rotifers that they can powerfully resist the mutagenic effects of radiation [35].

If the mechanisms of sexual and species differentiations were similar, then, since the opportunity for recombination between different sex chromosomes (e.g. X and Y) can occur only in the sex with both chromosomes (the heterogametic sex), by *preventing* meiotic recombination that sex could indeed be considered to have taken a *step* towards speciation. Whereas, for species differentiation, the homogametic sex (e.g. human females) would have to differentiate *both* sex chromosomes *and* autosomes, the heterogametic sex (e.g. human males) would have only the autosomes left to differentiate. By virtue of this head-start, among the progeny of crosses between an incipient species and its parental stock, *the heterogametic sex would be preferentially sterile* (i.e. Haldane’s rule for hybrid sterility).

General changes in the base composition [(G+C)%] of chromosomes – perhaps the modern equivalent of Goldschmidt’s changes in “reaction pattern” – are consistent with studies of Haldane’s rule in the fruit fly, where, like humans, the male is the heterogametic sex. Having attempted to locate and characterize chromosomal “factors” affecting hybrid sterility in fruit fly, Naveira and Maside concluded in 1998 [1]:

The total number of sterility factors [on chromosomes] must probably be numbered *at least* in the hundreds. The individual effect on fertility of any [one] of these factors is virtually undetectable, but can be accumulated to others. So, hybrid male sterility results from the [experimental] co-introgression [combining] of a minimum number of randomly dispersed factors (polygenic combination). The different factors linked to the X, on the other hand, and to the autosomes, on the other, are *interchangeable* [i.e. are equally effective] Recent experiments on the nature of these polygenes suggest that *the coding potential of their DNA may be irrelevant*.

In the latter respect it was further noted:

The effect detected after inserting non-coding DNA suggests that the coding potential of the introgressions ... might be ... irrelevant for hybrid male fertility. *It might be only a question of foreign DNA amount, ...*

In the general case, later-developing *secondary* mechanisms of isolation, reinforcing and/or substituting for a primary micromutation-dependent mechanism, would include chromosomal macromutations (likely to increase hybrid sterility), single or collective *genic* incompatibilities (likely to produce hybrid inviability), and mating incompatibilities (also likely to be of *genic* origin; Fig. 8.4).

Over evolutionary time, factors such as the transcription rates of genes on the Y-chromosome and the stabilities of their products (mRNAs and proteins) could have become fine-tuned to the needs of this *relatively stable* intracellular environment. A relatively constant concentration of each gene product could have become established (Fig. 17.3). On the other hand, an X-chromosome and its descendants alternate between male (M) and female cells (F). A typical path might be:

$$M \rightarrow F \rightarrow F \rightarrow M \rightarrow F \rightarrow M \rightarrow F \rightarrow \dots \quad (17.7)$$

Over evolutionary time it would be difficult to fine-tune, both for cells containing a second X-chromosome, and for cells containing a solitary X-chromosome. By inactivating one X whenever two are present, the intracellular environment would be stabilized and the collective fine-tuning of protein concentrations could continue. This would facilitate collective protein functions, such as intracellular discrimination between self and not-self.

In humans the ‘decision’ to dosage compensate the X-chromosome is a responsibility of females. Since mutations are generally loss-of-function, rather than gain-of-function, those affecting the dosage compensation system should selectively impair the ability of females to inactivate one of their X chromosomes, resulting in *increased* cytosolic aggregation pressure. If this relates to self/not-self discrimination (Chapter 16), then an immune response against self might be more easily

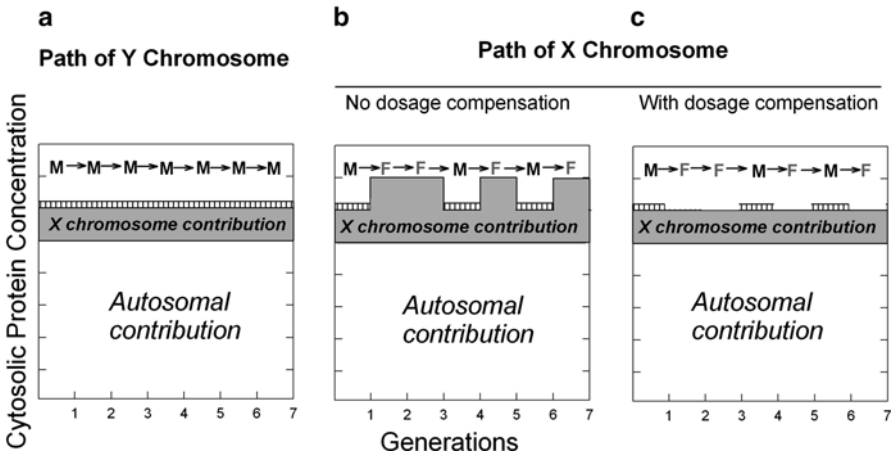


Fig. 17.3 Fine-tuning of protein concentration is not possible without X-chromosome dosage compensation. Passage of Y- or X-chromosomes through the generations occurs either in male (M) or female (F) cells. On average an X-chromosome exists for one third of its time in a male cell. Potential contributions of chromosome gene-products to the cytosolic protein concentration are shown for autosomes (white), X-chromosomes (gray), and Y-chromosomes (thin strip with vertical stripes). (a) The presence of the Y-chromosome defines an organism as male, and thus the Y-chromosome always exists in a male cell, in the presence of one X-chromosome. Fluctuation of total protein concentration is minimal. (b) The contribution of the Y-chromosome is very small so that in female generations, defined by the presence of two X-chromosomes, there would be a great fluctuation in total protein concentration in the absence of dosage compensation. (c) Halving the contribution of the X-chromosome (dosage compensation) in female generations makes the cytosolic protein concentration independent of the sex of the host cell

triggered, and females could be more prone to autoimmune disease than males. This is indeed found [41].

In fruit flies the single male X chromosome works twice as hard as each of the two female X chromosomes, but the ‘decision’ is also female. The freshly fertilized egg undergoes 13 rapid chromosomal divisions without cell division ("syncytial phase"). In this two hour period the solitary X chromosome in males doubles its transcription rate, but so do the two Xs in females. By the 14th division the male is happily set up with a concentration of X chromosome products that will be sustained for the rest of his life. But the female has an excessive concentration in each cell. To compensate, she could increase cell volume. But female cells are not bigger than male cells. Alternatively she could cut down the transcription rates of her X chromosomes. Indeed, by six hours post-fertilization the rate of each of her Xs is *half* that of the solitary X in males [42].

How is an entire human chromosome inactivated? While beyond our present scope, we may note that an important factor in the inactivation process is the RNA product of a gene, *Xist* ("X inactivation specific transcript"), which is the *only* gene that remains active on the inactive X chromosome in human females. *Xist* RNA coats the inactive X-chromosome, but not the active X-chromosome. *Xist* is a gene without a protein product. Its transcription generates a long RNA molecule that contains introns. The processing of the RNA involves excision of the introns to generate the mature *Xist* RNA, which executes the gene's functions.

The allelic copy of *Xist* on the active X chromosome ("sense") is kept inactive by virtue of a counter-transcript ("antisense") from a gene known as *Tsix*. The RNA products of *Xist* and *Tsix*, presumably recognize each other by "kissing" interactions that result in formation of a duplex in the region of complementarity. Inactivation of one X chromosome is accompanied by extensive DNA methylation (see Chapter 18).

The importance of constant concentrations of intracellular proteins is emphasized by studies showing that a balance must be maintained not only between sex chromosomes, but also between the expressions of sex chromosomes and autosomes [43]. Furthermore, some genes, by virtue of the concentration and/or size of their protein products, contribute more to cytosolic aggregation pressure than others. Indeed, depending on the organisms studied, some genes on sex chromosomes are found not to participate in dosage compensation, presumably because this failure has not been selectively important enough. For those that do, we should note that the dosage function may be no less important than their 'primary' functions (e.g. enzyme activity), and both functions are targets of evolutionary selective pressures.

Thus it is observed that, contrary to conventional wisdom, "the [primary] functional importance of a protein only has a weak impact on its evolutionary rate," which is "predominantly influenced by its expression level" [44]. Indeed, a negative correlation between expression level and protein evolutionary rate is often found. In other words, genes corresponding to highly expressed proteins have fewer accepted mutations – they are conserved.

This conservation affects both internal and surface segments of their protein products (Fig. 15.2). A role for the surface segments in maintaining solubility and homoaggregation-specificity, follows from evidence that conservation ensures

“misinteraction avoidance” [45], and that the higher “quinary” structure of proteins is important [46]. Tight control of protein concentrations is evident from reports that the concentrations of the mRNA transcripts that encode proteins vary *more* than those of the proteins themselves [47, 48].

Gain in Aggregation Pressure

Aggregation pressure might decrease (perhaps leading to immune impairment) or increase (perhaps leading to immune hyperactivity). Thus, the high female susceptibility to autoimmune diseases might be explained as a failure to inactivate an X, so resulting in a gain in aggregation pressure [41]. Large chromosomes carry more genes than small chromosomes. When uncompensated, extra copies of large chromosomes are generally lethal. Extra copies of small chromosomes may be compatible with life, but not necessarily a healthy life. Those with mongolism (Down syndrome) have three copies of chromosome 21 instead of the normal two. They are triploid for chromosome 21. This chromosome has 233 protein-encoding genes that include one encoding the AIRE protein (Chapter 15) and another encoding a protein (APP), which aggregates in the brains of people with Alzheimer’s disease (Chapter 14). In this condition there is sometimes a general, or tissue-specific (mosaic), duplication either of the APP gene alone or of one of the chromosomes carrying it [49, 50]. Thus, APP is heavily implicated in Alzheimer pathology.

Mongols are happy folk, recognizable by their characteristic smiling faces. But late in life this changes. Symptoms of Alzheimer’s disease emerge. Despite expressing all three copies of the AIRE protein gene, which should *protect* against autoimmunity, added to their afflictions is autoimmune disease. There might have been a gain in aggregation pressure due to increased concentrations of many of the 233 proteins. This might have moved the aggregation threshold to the left (Fig. 16.5), so enhancing anti-self immunity in multiple tissues and aggregating AIRE, so decreasing its thymic activity [51].

Thymus and Gonads as Auditioning Organs

AIRE promotes the ‘promiscuous’ expression of proteins in tissues where they are not usually expressed. Its own expression is largely tissue-specific, being especially prevalent in the thymus and gonads – organs where cell multiplication is high but few survive to ‘tell the tale.’

In general, the development of embryonic organ systems involves the differentiation of populations of cells that contribute to the function of the organ that contains them. Any cell death during embryonic development is part of normal organogenesis and usually involves small, specific, populations that are ‘programmed’ to die at distinct stages. Even in adult life, for most organs random cell

death (with replacement by other cells) appears to be part of the normal ‘wear and tear’ of existence, and its scale is relatively minor [52]. However, for the thymus and gonads, massive cell death is the norm. They are ‘seed organs’ in that the thymus seeds the body with T-lymphocytes, and the gonads seed gametes. Cell division within a seed organ produces a large, but non-uniform, cell population. Then ‘winners’ emerge, and the rest perish.

For T cells this is appropriately referred to as ‘auditioning,’ in that some survive the medullary epithelial cell (MEC) ‘gauntlet’ in the thymus and the rest die (Chapter 15). Perhaps gamete-forming cells, especially in the female, are also ‘auditioning,’ so there is an extra opportunity for gamete quality control. At birth in humans the ovary within a female foetus is packed with follicles, each with a G2-suspended oocyte capable of contributing to a new life. But during the potential mother’s lifetime oocyte number steeply declines and, over the years, very few emerge with the potential to encounter a sperm.

Given successive generations – A, B, C – and the imperative to continue the line, when considering evolutionary variables we usually focus on the immediately preceding generation – B in the case of C, and A in the case of B. Thus, failure of development of C, due to inability of the paternal and maternal genomes (B) to work together (hybrid inviability), is often due to one or more genic incompatibilities. Clearly the four grandparental genomes (A), based on the alleles expressed, did not have this problem, otherwise there would be no B parents to donate their genomes.

However, it is one thing to have each pair of grandparental genomes (A) working together. It is another to have the meiotic recombinational *selections* from those genomes (occurring in B gonads), working together. Some screening system would seem advantageous. The “audition” would check the compatibilities of the unique combinations of grandparental genomes that characterized each haploid B gamete. Perhaps the AIRE-induced co-expression of diverse genes, would allow detection of some degree of gene-product incompatibility that otherwise would not normally be tested in this location (the gonad). Instead the test would have to be post-fertilization, perhaps at an advanced stage of embryo development when the parents had already committed extensive resources to the pregnancy. Thus, this form of early pre-fertilization detection, albeit not so comprehensive, could be selectively advantageous.

Germ-Line Selection within Gonads

While further proliferation of pre-meiotic gamete-forming cells in female ovaries is largely arrested at birth, gametogenesis continues in mammalian testes throughout adult life. There are less rounds of cell division in oogenesis than in spermatogenesis. Whether or not the auditioning gamete-quality control mechanism exists, since DNA replication for cell division is mutagenic, with imperfect meiotic cleansing more mutations should be carried by emerging spermatozoa than by emerging ova. If a mutation in a spermatogonial stem cell accelerated its proliferation, then the

chance that an individual spermatozoon would carry that mutation would increase. In 1911 Bateson and Punnett suggested that such localized clonal proliferation would produce islands of mutant cells within a gonad. The mutant cells would be localized to particular gonadal segments [53]. In 1912 a German physician, Wilhelm Weinberg, suggested that it would take time for these pre-meiotic mutations to accumulate, so the last born children in a family would most likely be affected. He noted that the incidence of achondroplastic dwarfism correlated with the age of fathers, not of mothers [54]. We now know that such “paternal age effect” diseases result from dominant mutations that affect mitotic cell proliferation. In achondroplasia the mutation affects the sensitivity of a cell receptor for a growth factor. By some chain of events, this is manifest as dwarfism in offspring [55, 56].

Summary

The need for genomes to detect and correct errors has been a major force in evolution, driving DNA to ‘speak in palindromes’ and splitting members of a species into two sexes. Errors associated with DNA mutations or DNA damage may be imperfectly corrected in our bodies (“soma”), but should be better corrected in the germ line. To this end, meiotic recombination repair, in which maternal and paternal genomes are compared, is the last court of appeal. For this ‘rebooting’ of the genome, parental genomes must be neither too similar, nor too disparate. If too similar (inbreeding), differences will not emerge at meiosis in their child’s gonad. If too disparate, an exploratory speciation process (anti-recombination) may initiate, manifest as a healthy, but sterile, child (hybrid sterility). For the initiation of speciation, human females must: (i) differentiate their sex chromosomes (X and X), (ii) differentiate their non-sex chromosomes (22 autosomal pairs), and (iii) activate ‘check-points’ which respond to such differentiations by disrupting meiosis. In contrast human males, being already advanced in the first step due to differentiation of their sex chromosomes (X and Y), have to complete only the latter steps. Thus, the first sign of speciation (incipient speciation), manifest as hybrid sterility, is production of sterile males (Haldane’s rule). Long ago, anti-recombination activity prevented repair of Y-chromosomes, which degenerated, thus losing many X-chromosome equivalent genes. Human males now have potentially only one dose of many X-chromosome gene products, whereas females have potentially two doses. Dosage compensation in human females, leaving only one X-chromosome active, buffers fluctuations in intracellular protein concentrations between male and female generations. This permits a gene, independently of the sex which may harbor it, to fine-tune the concentration of its protein product to that of other proteins with which it has been traveling through the generations. In this way, collective protein functions, including intracellular self/not-self discrimination, are facilitated. Failure of this may predispose human females to autoimmune disease. By virtue of extensive palindromic sequences, Y chromosomes appear, like some rare ameiotic organisms, capable of in-series error-correction.

References

1. Naveira HF, Maside XR (1998) The genetics of hybrid male sterility in *Drosophila*. In: Howard DJ, Berlocher SH (eds) *Endless Forms and Speciation*. Oxford University Press, Oxford, pp 329–338
2. Delboeuf J (1877) Les mathématiques et le transformisme. Une loi mathématique applicable à la théorie du transformisme. *La Revue Scientifique* 29:669–679
3. Bernstein C, Bernstein H (1991) *Aging, Sex and DNA Repair*. Academic Press, San Diego, CA
4. Supek F, Lehner B (2015) Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* 521:81–84
5. Ridley M (2000) *Mendel's Demon. Gene Justice and the Complexity of Life*. Orion Books, London, pp 167–201
6. Medvinsky A, Smith A (2003) Fusion brings down barriers. *Nature* 422:823–825
7. Reese V (2002) Mutation repair: a proposed mechanism that would enable complex genomes to better resist mutational entropy and which suggests a novel function for meiosis. *The Human Behavior and Evolution Society 14th Annual Meeting*, Rutgers University. Abstracts of presentations to session on "New Developments in Biology," June 21, p. 40. [See also: "Meiotic pairing inadequacies at the levels of X chromosome, gene, or base: epigenetic tagging for transgenerational error-correction guided by a future homologous duplex." (Reese VR, Forsdyke DR, 2016; Biological Theory, in press.)]
8. McKee BD (1996) Meiotic recombination: a mechanism for tracking and eliminating mutations? *BioEssays* 18:411–419
9. Wahls WP, Davidson MK (2012) New paradigms for conserved, multifactorial, *cis*-acting regulation of meiotic recombination. *Nucleic Acids Research* 40:9983–9989
10. Baker CL, Kajita S, Walker M, Saxl RL, Raghupathy N, Choi K, Petkov PM, Piagen K (2015) PRDM9 drives evolutionary erosion of hotspots in *Mus musculus* through haplotype-specific initiation of meiotic recombination. *PLOS Genetics* 11:e1004916
11. Butler S (1914) *The Humour of Homer and Other Essays*. Kennerley, New York, pp 209–313
12. Noort V van, Worning P, Ussery DW, Rosche WA, Sinden RR (2003) Strand misalignments lead to quasipalindrome correction. *Trends in Genetics* 19:365–369
13. Navarre WW, McClelland M, Libby SJ, Fang FC (2007) Silencing of xenogeneic DNA by H-NS – facilitation of lateral gene transfer in bacteria by a defense system that recognizes foreign DNA. *Genes & Development* 21:1456–1471
14. Forsdyke DR, Zhang C, Wei J-F (2010) Chromosomes as interdependent accounting units. The assigned orientation of *C. elegans* chromosomes minimizes the total W-base Chargaff difference. *Journal of Biological Systems* 18:1–16
15. Sinclair R (2015) Necessary relations for nucleotide frequencies. *Journal of Theoretical Biology* 374:179–182
16. Forsdyke DR (2001) *The Origin of Species, Revisited*. McGill-Queen's University Press, Montreal
17. Darwin C (1871) *Descent of Man, and Selection in Relation to Sex*. Appleton, New York, pp 245–311
18. Haldane JBS (1922) Sex ratio and unidirectional sterility in hybrid animals. *Journal of Genetics* 12:101–109
19. Coyne JA (1992) Genetics and speciation. *Nature* 355:511–515
20. Forsdyke DR (1995) Fine tuning of intracellular protein concentrations, a collective protein function involved in aneuploid lethality, sex-determination and speciation. *Journal of Theoretical Biology* 172:335–345
21. Chandley AC, Jones RC, Dott HM, Allen WR, Short RV (1974) Meiosis in interspecific equine hybrids. 1. The male mule (*Equus asinus* X *E. caballus*) and hinny (*E. caballus* X *E. asinus*). *Cytogenetics & Cell Genetics* 13:330–341
22. Vries H de (1889) *Intracellular Pangenesis*. Open Court, Chicago, (1910) pp 18–19
23. Darwin C (1851) *A Monograph on the Subclass Cirripedia*, vol. 1. The Ray Society, London, pp 281–293

24. Bateson W (1922) Evolutionary faith and modern doubts. *Science* 55:55–61
25. Goldschmidt R (1940) *The Material Basis of Evolution*, Yale University Press, New Haven, pp 233–236
26. Forsdyke DR (2000) Haldane's rule: hybrid sterility affects the heterogametic sex first because sexual differentiation is on the path to species differentiation. *Journal of Theoretical Biology* 204:443–452
27. Romanes GJ (1886) Physiological selection: an additional suggestion on the origin of species. *Journal of the Linnean Society (Zoology)* 19:337–411
28. Romanes GJ (1897) *Darwin, and After Darwin: 3. Isolation and Physiological Selection*. Longmans Green, London
29. Wallace AR (1868) Letter to Darwin, 24th March. Available at <https://www.darwinproject.ac.uk/entry-6045>
30. Koller PC, Darlington CD (1934) The genetical and mechanical properties of the sex chromosomes. 1. *Rattus norvegicus*. *Journal of Genetics* 29:159–173. [In male meiosis, the Y chromosome is an inadequate pairing partner for the X chromosome with respect to error-correction. At that time the X chromosome's transcription products are not required and it is epigenetically silenced and cordoned off with the Y chromosome as the "XY body" - a process known as "meiotic sex chromosome inactivation" (MSCI). Thus, the X chromosome emerges without error-correction from male meiosis. It can only redeem itself in female generations, where there is a pairing partner. This appears to be an example of a more widely recognized phenomenon - "meiotic silencing of unpaired DNA" (MSUD; see Chapter 18), and is consistent with the shorter lifespans of the daughters of elderly fathers (see Gavrilov et al. 1997, "Mutation load and human longevity." *Mutation Research* 377:61–62). The latter is an example of a "paternal age effect."]
31. Montoya-Burgos JI, Boursot P, Galtier N (2003) Recombination explains isochores in mammalian genomes. *Trends in Genetics* 19:128–130
32. Willard HF (2003) Tales of the Y chromosome. *Nature* 423:810–813
33. Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G (2004) Inverted repeat structure of the human genome: the X chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Research* 14:1861–1869
34. Flot J-F, et al. (2013) Genomic evidence for ameiotic evolution in the bdelloid rotifer *Adineta vaga*. *Nature* 500:453–457
35. Gladyshev E, Meselson M (2008) Extreme resistance of bdelloid rotifers to ionizing radiation. *Proceedings of the National Academy of Sciences USA* 105:5139–5144
36. Gartler SM (2014) A brief history of dosage compensation. *Journal of Genetics* 93:591–595
37. Carrel L, Cottle AA, Goglin KC, Willard HF (1999) A first-generation X-inactivation profile of the human X chromosome. *Proceedings of the National Academy of Sciences USA* 96:14440–14444
38. Papadopoulos AST, Chester M, Ridout K, Filatov DA (2015) Rapid Y degeneration and dosage compensation in plant sex chromosomes. *Proceedings of the National Academy of Sciences USA* 112:13021–13026 [In some species MSCI extends transgenerationally, so that in daughters only the X chromosome donated by the mother is expressed and the X chromosome donated by the father remains epigenetically silenced (see: Engel 2015, "Imprinted X chromosome inactivation offers up a double dose of epigenetics." *Proceedings of the National Academy of Sciences USA* 112:14408–14409)]
39. Bridges CB (1922) The origin of variations in sexual and sex-limited characters. *American Naturalist* 56:51–63
40. Forsdyke DR (1994) Relationship of X chromosome dosage compensation to intracellular self/not-self discrimination: a resolution of Muller's paradox? *Journal of Theoretical Biology* 167:7–12
41. Forsdyke DR (2009) X chromosome reactivation perturbs intracellular self/not-self discrimination. *Immunology & Cell Biology* (2009) 87:525–528
42. Gladstein N, McKeon MN, Horabin JI (2010) Requirement of male-specific dosage compensation in *Drosophila* females—implications of early X chromosome gene expression. *PLoS Genetics* 6: e1001041

43. Forsdyke DR (2012) Ohno's hypothesis and Muller's paradox: sex chromosome dosage compensation may serve collective gene functions. *BioEssays* 34:930–933
44. Zhang J, Yang JR (2015) Determinants of the rate of protein sequence evolution. *Nature Reviews Genetics* 16:409–420
45. Yang J-R, Liao B-Y, Zhuang S-M, Zhang J (2012) Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proceedings of the National Academy of Sciences USA* 109:E831–E840
46. Monteith WB, Cohen RD, Smith AE, Guzman-Cisneros E, Pielak GJ (2015) Quinary structure modulates protein stability in cells. *Proceedings of the National Academy of Sciences USA* 112:1739–42
47. Laurent JM, et al. (2010) Protein abundances are more conserved than mRNA abundances across diverse taxa. *Proteomics* 10:4209–4212
48. Khan Z, Ford MJ, Cusanovich DA, Mitrano A, Pritchard JK, Gilad Y. (2013) Primate transcript and protein expression levels evolve under compensatory selection pressures. *Science* 342:1100–1104
49. Wiseman FK, Al-Janabi T, Hardy J, Karmiloff-Smith A, Nizetic D, Tybulewicz VLJ, Fisher EMC, Strydom A (2015) A genetic cause of Alzheimer disease: mechanistic insights from Down Syndrome. *Nature Reviews Neuroscience* 16:564–574
50. Jacobs PA (2014) An opportune life: 50 years in human cytogenetics. *Annual Review of Genomics & Human Genetics* 15:29–46
51. Gimenez-Barcons M, Casteras A, Armengol MdP, Porta E, Correa PA, Marin A, Pujol-Borrell R, Colobran R (2014) Autoimmune predisposition in Down syndrome may result from a partial central tolerance failure due to insufficient intrathymic expression of AIRE and Peripheral Antigens. *Journal of Immunology* 193:3872–3879
52. Danial NN, Korsmeyer SJ (2004) Cell death. Critical control points. *Cell* 116:205–219
53. Bateson W, Punnett RC (1911) On gametic series involving reduplication of certain terms. *Journal of Genetics* 1:293–302
54. Weinberg W (1912) Zur Vererbung des Zwergwuchses. *Archiv für Rassen- und Gesellschafts-Biologie* 9: 710–717
55. Goriely A, Wilkie AO (2012) Paternal age effect mutations and selfish spermatogonial selection: causes and consequences for human disease: *American Journal of Human Genetics* 90:175–200
56. Shinde DN, Elmer DP, Calabrese P, Boulanger J, Arnheim N, Tiemann-Boege I (2013) New evidence for positive selection helps explain the paternal age effect observed in achondroplasia. 22:4117–4126

Chapter 18

The Fifth Letter

Intriguingly, because methylated cytosine can preferentially mutate to thymine, there could arise a conceptually important ‘Lamarckian-like’ mechanism: genes that are not expressed in the germline (i.e. somatic tissue-specific genes) might at first be epigenetically switched-off by cytosine methylation, which could be later transformed into permanent suppression caused by mutational CpG depletion. This suggestion is consistent with a high CpG level of genes expressed in the germline. Thus, the epigenetic suppression of (somatic) tissue-specific genes in the germline could be transformed into the genetic (genome-wired) suppression.

Alexander Vinogradov (2005) [1]

In the biological sciences few generalizations are absolute and we have already noted strange bases in RNA in addition to the usual – **A**, **C**, **G** and **U** (Fig. 5.1). For many purposes DNA can be considered solely in terms of its four major bases – **A**, **C**, **G** and **T**. However, in written languages single letters are sometimes qualified with accents. The words *gène* and *gêne* have very different meanings in French. We should not be surprised to find that there are similar qualifications in the DNA language. The most evident of these is methylcytosine, where the base **C** acquires a chemical grouping (methyl) [2]. Thus, in many organisms DNA has five letters – **A**, **C**, **Me-C**, **G** and **T**. Apart from the four regular bases, there is a pattern of C-methylation at intervals along a DNA sequence.

Editing of DNA

DNA is synthesized as a string of the four base letters, but subsequently some of the **C**’s are modified so they become **Me-C**. The enzymes that bring this about (methyl transferases) can distinguish between **C**’s on the basis of their 3' (i.e. downstream) nearest neighbors. Thus, **C**’s in the **CG** dinucleotide can be modified, but **C**’s in the dinucleotides **CA**, **CC**, and **CT** are usually not. The nature of the 5' (i.e. upstream) base is usually irrelevant.

CG (often written as **CpG**) is a dinucleotide which, when considered at the level of a DNA duplex, has ‘palindromic’ characteristics to the extent that **CG** on the top

strand is matched by **CG** on the bottom strand (see sequence 17.1 in Chapter 17). It is a member of the set of four self-complementary dinucleotides (Table 4.1a). In some circumstances methylation of a **C** in the **CG** of one strand will only occur if the **C** in the **CG** in the complementary strand is *already* methylated. Thus, prior to replication of a DNA duplex (Fig. 2.3) the **CG**'s in the parental strands may already be symmetrically methylated. Immediately after replication each new duplex will then contain one parental strand with its original me-**C**'s, and one child strand without me-**C**'s. The methylating enzymes then methylate the child strands so that symmetric methylation is restored (Fig. 18.1). To this extent, the pattern of methylation

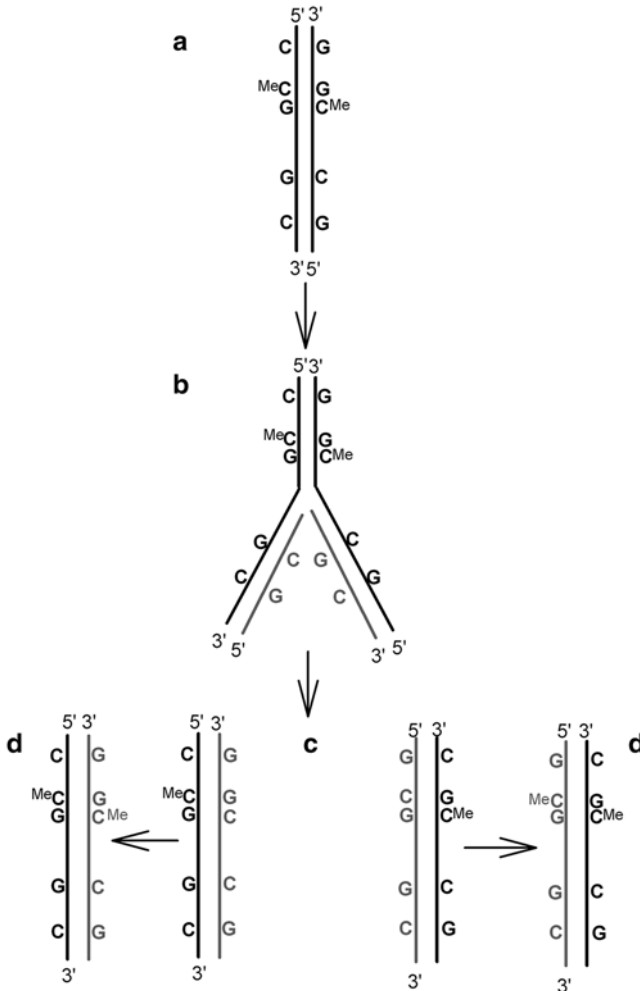
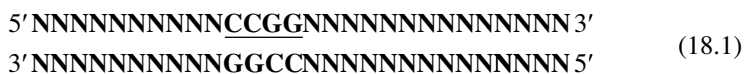


Fig. 18.1 The post-synthetic editing of DNA. In a parental DNA duplex, where only **C-G** Watson-Crick base-pairs are shown, Me-**C** is present in one complementary **CG** dinucleotide pair (a). The two strands of the parental duplex (black) are replicated (b) to produce child strands (gray). Initially, the child strands in the two freshly replicated duplexes (c) do not have methylated **CG** dinucleotides. These half-methylated new duplexes are subsequently fully methylated (d). Thus the symmetrical parental methylation pattern is inherited by child cells. The child ‘remembers’ the parental pattern

in a DNA sequence is inherited by a succession of cells – a form of inheritance sometimes referred to as “epigenetic” [3]. This word, however, has had a variety of historical usages, and must be understood in context.

Restriction Enzymes

Many bacteria edit their DNA to mark it as ‘self.’ The DNA of a virus that infects bacteria (bacteriophage) is recognized as ‘not-self’ because of the absence of the modification. Enzymes of bacterial origin are thus able to distinguish self from not-self. The enzymes cut the viral DNA duplex at, or near to, a specific recognition sequence (i.e. they are often sequence-specific endonucleases). This *restricts* the growth of the virus in a most definitive way, by cleaving its DNA into fragments. Thus, the “restriction enzymes,” like the editing enzymes, recognize specific base patterns, often palindrome-like, and generate “restriction fragments.” For example, a restriction enzyme named “HpaII” in the bacterium *Haemophilus parainfluenzae* recognizes the four base sequence **CCGG** in foreign duplex DNA. This sequence will already have been recognized in the bacterium’s DNA by its own editing enzymes and appropriately modified to declare it as “self” and therefore uncuttable. The following is unmodified, and is therefore declared “foreign” (not-self) and cuttable:



The DNA is cut asymmetrically to generate fragments:



Since the cutting enzyme (nuclease) attacks the nucleic acid internally, it is an “endonucleases,” rather than an “exonuclease” (enzymes which degrade nucleic acids from the ends; Latin, *exo* = outer; *endo* = inner). Some restriction enzymes attack a foreign DNA in such a way that sets of restriction fragments of uniform size are generated. The sets can be separated as individual bands in a gel, which can be stained to allow the bands to be visualized. The pattern of bands is generally unique for a particular DNA sequence, so “restriction digests” of a DNA sample allow “restriction-site mapping” of the DNA. From such maps, and the sequences of individual, overlapping, fragments, large genome sequences can be assembled. Genetic engineering began in the 1970s with the transfer of the fragments between organisms.

As you might expect, bacteriophages do not idly wait for prospective hosts to demolish their DNA. Their defense strategies include the encoding of “antirestriction enzymes” that can inactivate host restriction enzymes – yes, this seems to be another example of a biological arms race (Chapter 13). Whether their hosts have yet evolved “anti-antirestriction enzymes” remains to be seen!

Bacterial restriction enzymes are co-inherited with bacterial editing enzymes specific for the same sequence, and the corresponding genes are co-localized in the genome as a “restriction-modification gene complex.” This constitutes a prototypic package for intracellular self/not-self discrimination. Co-localization (linkage) of the genes would be expected, since this decreases the possibility of their separation by recombination, and facilitates coordinate regulation. A mutation in the restriction enzyme gene could impair a bacteria’s ability to repel a virus. On the other hand, a mutation in the editing enzyme gene might result in a failure of modification (a lack of site-specific DNA methylation). The mutant bacterium’s DNA would then become vulnerable to its own restriction enzymes.

This attack on self-DNA might not occur immediately, since the DNA would already have been edited at the time the mutation occurred, and the existing editing enzymes might still be operative, although subject to degradation and turnover (Chapter 2). After one round (cycle) of replication the DNA duplexes would still retain one parental methylated strand, and this might suffice to protect the DNA. But after two rounds of replication, some child cells would have completely unmodified DNA duplexes. By that time, the pre-existing editing enzymes would probably have been so degraded, or so diluted by cell growth, that their concentrations would have become insufficient to maintain DNA modifications. In this circumstance the microbial equivalent of autoimmune disease could arise. However, many bacteria also have the CRISPR system (Chapter 15), which endows them with the equivalent of the immunological memory observed in eukaryotes. Thus, they might escape to allow their descendents to fight another day.

CpG Islands

Regulation of the transcription of a eukaryotic gene is often brought about by the binding of various proteins (“transcription factors”) near to the 5' end of the gene, where there is a site (“promoter”) to which RNA polymerase molecules bind. Transcription factors generally provide for short-term, on-off, control of transcription – either activating or inactivating transcription. A more definitive stamp of inactivation involves the methylation of CpG dinucleotides in the promoter region. This can ‘lock-in’ (but not ‘hard wire’) the inactivation, sometimes for many cell generations [1].

A dramatic example of this somatic lock-in process involves the inactivation of one of the female X-chromosomes, which has been designated for inactivation by *Xist* gene RNAs acting on the chromosome from which they were derived (Chapter 17). In this case there is a general spread of methylation to the promoter regions of all genes of the chromosome, except that of the *Xist* gene itself. In the other X-chromosome (active) the allelic *Xist* gene is itself inactivated, a process requiring another gene (*Tsix*) and methylation of the *Xist* gene promoter.

‘Unwanted’ genetic elements, which may ‘intrude’ into DNA sequences (e.g. retroviruses, or gene duplicates), are sometimes inactivated by CpG methylation and/or

by mutation [4]. However, methylation of **C** means that, if the Me-**C** is subsequently deaminated, there will be a mutation to **T**, which is a normal DNA base. This **T** will now exist in duplex DNA attempting to engage in Watson-Crick pairing with the previous pairing partner of the **C** – namely a **G** [5]. The **T-G** mispairing (e.g. sequence 17.5 in Chapter 17) sometimes creates a problem for the repair enzymes, which may not know which is the correct base. So a region where CpG dinucleotides are methylated is especially vulnerable to mutation (i.e. it is a mutational ‘hotspot’).

When such a mutation occurs in the DNA of somatic (non-germ line) cells, it is not passed on to children. When it occurs in the DNA of germ line cells, gametes are affected and so it can be passed on to children, who can then pass it on to their children, etc.. As a consequence, sequences that are *inactive* (and hence methylated) in the germ line (e.g. genes that are inactive in the gonad, but are later required for a function *specific* to a particular somatic tissue), have tended to lose **CG** dinucleotides. The sequences then either change function [1] or maintain the inactive germ-line state by means other than **CG** methylation (i.e. they are now found to be deficient in **CG** dinucleotides and inactive, and it can be presumed that they once sustained the inactive state by methylation).

On the other hand, sequences that are *active* in the germ line (often containing ‘housekeeping’ genes required for *general* cell function), do not have methylated **CG** dinucleotides, and so are less subject to mutation. These sequences retain their **CG** dinucleotides, which may be recognized as “CpG islands.” An example is human *GOS2*, a putative tumor-suppressor gene. Transcription of the normally unmethylated gene is inhibited by methylation in cancer patients [6]. In Fig. 15.3 the gene is shown with a major peak of **CG** dinucleotides. Careful inspection also reveals minor peaks, some associated with *Alu* elements [7]. This suggests some *Alu* expression in the germ line and, thus, a ‘housekeeping’ function for *Alu* elements. Depending on the cut-off criteria for defining peaks as “islands,” in human chromosomes there are *far more* **CG** islands than can be accounted for by conventional genes, such as *GOS2* (‘macro-islands’), or by *Alu* elements (‘micro-islands’). This points to the ‘hidden transcriptome,’ perhaps with a ‘housekeeping’ role in intracellular defense (Chapter 15).

CpG Suppression

CG dinucleotides are 2-mers that, like k-mers in general, have frequencies characteristic of the species (Chapter 4). In many eukaryotes, the tendency of Me-**CG** to mutate to **TG** (Chapter 17) is an example of a random mutational bias that arises intrinsically from the underlying chemistry of DNA – in this case the instability of an amino group in **C** that, wherever the **C** be located in DNA, readily deaminates so **C** becomes **U**. We have noted that this is favored by the alkaline conditions within mitochondria (Chapter 6). The deamination can also be a targeted, physiological, enzyme-mediated process. Mutagenic cytosine deaminases contribute to the somatic generation of antibody diversity (extracellular *adaptive* immunity) [8], and can also operate at the RNA level (see later).

When it is Me-C that is deaminated there is suppression of **CG** frequencies and an elevation of **TG** frequencies. If uncorrected, **TG** on one strand will persist with the complement **CA** on the opposite strand (Table 4.1b). Thus, a species with **CG** suppression sometimes shows an elevation of the frequencies both of **TG** and **CA** [9]. When there is **CG** suppression, then the frequencies of k-mers containing **CG** (e.g. **CGN** and **NCG**, where **N** is any of the four bases) may also be suppressed (Fig. 4.2b).

CG suppression is particularly evident in primates, but is not evident in the gut bacterium *E. coli* [10], which protects its DNA against its own restriction endonucleases by selective editings that include methylation of **CG**. Relative to the DNA of the somatic cells of its primate hosts, *E. coli* DNA has more **CG** dinucleotides that, in general, are less methylated. Hosts that could recognize this difference might exploit it. Indeed, DNA with these characteristics (as might be released from dying bacteria) can be ‘read’ as a not-self signal by an unwilling host (another example of *innate* immunity) [11]. Thus, species differences in **CG** frequencies are of much interest with respect to immunological defenses.

When comparing **CG** dinucleotide frequencies in different species (Fig. 18.2) it is usual to ‘correct’ for base composition. By virtue of its base composition alone, a high **(G+C)%** genome would be expected to have many **CG** dinucleotides

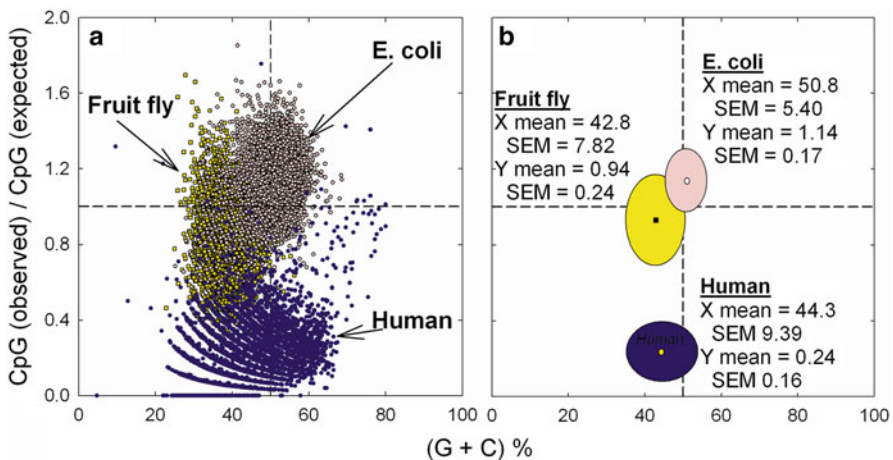


Fig. 18.2 CpG suppression. 5000 randomly selected 500 base segments from three genomes (the bacterium *E. coli*, the fruit fly *Drosophila melanogaster*, and the primate *Homo sapiens*) were each scored both for the ratio of observed to expected frequencies of the **CG** dinucleotide (scoring 1.0 if there is neither suppression nor enhancement), and for base composition, **(G+C)%**. For each organism in (a) the data points form distinct clusters, the fruit fly points (yellow squares) being partly hidden by *E. coli* points (pink circles) and human points (blue circles). The scattered points are summarized in (b) as a single point (defined by a mean X-axis value and a mean Y-axis value) surrounded by an oval that provides a visual measure of the scatter about the central point (SEMs; standard errors of the mean with respect to the mean X-axis and mean Y-axis values). Note that most *E. coli* segments show slight enhancement, whereas most segments from its human host show profound suppression. Some segments from the fruit fly genome show enhancement and some show suppression. Data for preparing this figure were kindly provided by Daiya Takai and Peter Jones

(i.e. if the sequence were shuffled, more **C**'s would be found with a downstream **G** on their 3' side than in the case of a low (**G+C**)% genome). Thus, **CG** frequency can be considered as having base *composition*-dependent and base *order*-dependent components (Chapter 5). Knowing base composition, an “expected” **CG** frequency can be calculated for a given DNA (i.e. by multiplying the independent probabilities of **C** and **G**). Thus, **CG** frequency can be presented as a ratio of “observed” **CG**'s (actual number of **CG** dinucleotides counted) to “expected” **CG**'s. In some species Nature has been adjusting base order to enhance **CG** frequency in most DNA segments. In some other species Nature has been adjusting base order to suppress **CG** frequency in most DNA segments. In other species some DNA segments show enhancement and others show suppression.

Methylation Differences in Twins

Normally a fertilized ovum develops into one individual. But sometimes it divides to produce two child cells that each develop into an individual. Two individuals with close genetic identity are then born. Such monozygotic twins are used in attempts to distinguish the relative roles of inherited and environmental factors (nature versus nurture). Character differences between monozygotic twins become more evident as they age and, accompanying this, there are increasing differences in patterns of methylation and gene expression. Whether these differences are random (i.e. the two individuals drift apart as the time from their divergence from a single egg increases), or are the result of differing inputs from the environment, can be determined by comparing twins that have been raised together, with twins that were separated to different environments at an early age.

Usually environment has a significant input into methylation as so assessed [12], and this seems to extend to behavior. Mother rats through nurturing behavior produce stress-resistant female offspring that themselves become nurturing mothers and, through this behavior, transmit the nurturing trait to their offspring. Ian Weaver and his colleagues [13] showed that maternal nurturing caused a transcription factor (NGF1A) to bind DNA receptors that respond to stress hormones (steroids) in the brain (hippocampus) of their newborn. This resulted in decreased DNA methylation of the receptors and a permanent change (into adulthood) in the response to stressful stimuli.

Imprinting

A major theme of this book has been conflict. And wherever we look for signs of conflict we usually find them. Even during pregnancy a human mother and her child are in potential conflict. The best diet for the child is cannibalistic – it should, literally, eat its mother – for she is likely to have an endogenous balance of nutrients that

would best promote *human* fetal development. Also she lacks the toxins found in exogenous foodstuffs, to which the embryo might be particularly susceptible at certain critical developmental stages [14]. It appears that embryos manipulate mothers both to avoid certain foods (to an extent that can cause maternal vomiting and loss of weight), and to increase the blood supply to the placenta (to an extent that, in extreme form, can cause death from high blood pressure). Even without a fetus, bizarre patterns of gene expression induced by hormones in adolescent females direct them to avoid exogenous food (anorexia nervosa).

However, it would seem that the genomes of male and female gametes, having united to form a diploid zygote, would then have a common interest and would cooperate unconditionally. Nature (i.e. the ‘hand’ of evolution) should not have ‘intentionally’ designed a conflict between them at the level of their embryo. The better the genomes cooperated, surely the more likely would be the zygote to complete its development and, in turn, produce healthy haploid gametes for the next generation? The phenomenon of imprinting that occurs in flowering plants and mammals tells us that this is not necessarily so [15].

By checking the methylation-status of the promoter region of a gene we can determine whether it is likely to be active or inactive. In the case of most genes, if they are needed for development then normally both the paternally-derived copy (paternal allele) and the maternally-derived copy (maternal allele) will be *unmethylated* (i.e. active). If the genes are not needed for development, then normally both copies (alleles) will be *methylated* (i.e. inactive).

In the case of some genes, however, one allele is always turned on (unmethylated) and one allele is always turned off (methylated). This determination (the ‘decision’ to methylate or not) was made in the gamete prior to fertilization (i.e. it was a determination made *from the perspective of each individual parent*). The differential methylation of certain allelic genes in the haploid genomes of male and female gametes is referred to as “gametic imprinting.”

One might dismiss this as trivial. Perhaps the developing embryo can manage quite well with one dose of a gene (i.e. it is haplosufficient). It does not need the other allele to be expressed. However, when the nature of the genes themselves was examined, it was found that genes that might *promote* embryonic development resulting in a high birth weight tended to be unmethylated (i.e. active) in the genome contributed by the father, but they tended to be methylated (i.e. inactive) in the genome contributed by the mother [16].

An ingenious explanation for this was that, in the long term, paternal and maternal interests in their children, biologically, are *not* the same. For a human mother, each pregnancy is a nine-month commitment to one child, during which time she can produce no further children. During her reproductive life there is a ceiling on the number of children she can produce. If her ‘selfish genes’ wish to maximize their representation in future generations they must accept mutations that ensure that the birth-weight of the first child is not so great as to exhaust the mother’s resources, so that she is less able to produce future children. Mutations that, by some chain of events, ensure during development of the maternal gamete that there will be future inactivation of genes whose overexpression in the embryo might increase birth-weight, would tend to have been accepted into human DNA.

For a human father there is no such ceiling on the number of children he can produce. Over evolutionary time, males who can fertilize the most females have tended to pass on more genes to future generations. Biologically, they have no necessary interest in those females' abilities to produce further children, which, indeed, might be fathered by other males (i.e. biologically, it is argued that male monogamy is less 'natural' than male polygamy). Thus, responding to an evolutionary pressure to maximize the representation of a father's selfish genes in future generations, humans must have accepted mutations in genes that, by some chain of events, ensure that in male gametes a state of activation is imposed on genes that will promote the development of a large embryo (thus giving the newborn a better start in life) [17, 18].

Transgenerational Inheritance

Because imprinting is imposed at the gamete level and inherited in the cells of the offspring, some of which will become the gametes of the next generation, patterns of methylation are usually erased (cancelled) and re-established each generation. The methyl group may be removed directly from **C** in DNA; alternatively Me-C may be excised from DNA and replaced with a fresh **C** by an appropriate repair enzyme [19].

But sometimes demethylation does not occur. Patterns of methylation can then be inherited unchanged over many generations – a form of transgenerational inheritance differing from the regular DNA-dependent form [1]. Indeed, the story of the fifth letter can be dated, like so many in genetics, back to William Bateson. In 1915 he and Caroline Pellew investigated the “rogue” phenomenon in peas [20]. From time to time lines of culinary peas “throw rogues.” The appearance of these weed-like individual pea plants does not follow the usual Mendelian pattern of inheritance. “Rogues” when crossed with normal plants always yield “rogue” offspring, and the non-rogue character is permanently lost. We can now understand this in terms of the sudden imposition of a pattern of methylation that is difficult to erase. This may involve a class of siRNAs (Chapter 15) known as piRNAs [21].

There is also growing evidence for transgenerational inheritance in animals (people feel uncomfortable calling it Lamarckian inheritance [1, 22]). For example, data suggesting that nutritional deprivation in childhood of grandfathers, increases the life-span of their grandsons, but not of their granddaughters, seems best interpreted in ‘epigenomic’ terms [23]. Also, as mentioned in Chapter 17, when it is known that a segment of DNA is in *need* of repair, but not *how* it should be repaired, then the segment might be epigenetically marked as “suspect.” This information would be employed in a subsequent generation to effect accurate repair when more information became available. Perhaps analogously, when an entire gene fails to find a meiotic pairing partner it can be epigenetically tagged to prevent its expression in offspring. However, should a pairing partner be found in a future generation, then the tag can be removed and function restored [24].

A problem here is that Nature does not have the foresight to ‘design’ a marking system the benefits of which may only become evident generations later.

So the marking would probably be an adaptation of a previously evolved marking system that served some other purpose. This is the “water-dipper” analogy of Samuel Butler [25]:

I have gone out sketching and forgotten my water-dipper; among my traps I always find something that will do, for example, the top of my tin case (for holding pencils). This is how organs come to change their uses and hence their forms, or at any rate partly how.

The same point was made by Stephen Jay Gould and Richard Lewontin in terms of “spandrels” [26].

Whereas often the epigenetic mark is DNA methylation, a mark can also be made on the proteins (histones) that accompany DNA. One of these (histone 3) is particularly adept at transmitting marks by accommodating variations in the pattern of methylation of its lysine residues [27]. Thus the configuration of chromatin – largely composed of DNA and histones – may be transiently modified. Such modifications may be referred to as “epigenomic.”

Editing of RNA

Editing at the DNA level is potentially transferable to the offspring of a cell and, in the case of germ-line cells, of an organism. But editing at the RNA level is likely to be confined to the cell of origin, with spread to other cells requiring special provocation [28]. One reason for RNA editing is to attend to the structural needs of the RNA transcript of a DNA that has different structural needs (Chapter 5). Thus, by post-synthetic editing, strange bases may be introduced into tRNA (Fig. 5.1). More general, but sometimes highly specific, editing, is by deamination either of **C** (**C** to **U** editing), or of **A** (**A** to **I** editing). Both affect base pairing (**U** pairs with **A**, and **I** (inosine) pairs with **C**). In the case of protein-encoding transcripts (mRNAs) the editing has the potential to change amino acid sequences, but most editing in humans affects non-protein-encoding transcripts, especially those from the “hidden transcriptome” – e.g. *Alu* repetitive elements. Such transcription can accompany an all-hands-on-deck emergency (Fig. 15.4), so care must be taken, when preparing RNA, to control for the stress of the extraction procedure [29, 30]. The extent to which RNA editing is a stress response needs to be known.

A parallel can be drawn between deaminations in DNA that expand the scope of extracellular immune diversity [8], and deaminations in RNA. *Alu* elements are already differentiated into long-lived subsets at the DNA level, but editing can add a further short-lived differentiation at the RNA level. This could expand the scope of intracellular immune diversity. Thus, following the transcription of *Alu* elements in response to viral infection, RNA editing would expand the scope of diversity, but at the risk of creating reactivity with a self-component. There would be specific base pairing with pathogen nucleic acid (Fig. 15.5) and, hopefully, vigilant RNA editors would prevent self-sequences forming dsRNA, so being mistaken as not-self [31]. When there is a cellular emergency there can also be a decline in the fidelity of mRNA translation [32] and the diversity of intracellular proteins (potential immune receptors) expands – again at the risk of creating self-reactivity.

Summary

Methylcytosine, the fifth base letter in DNA, is generated by modifying a **C** in a **CG** dinucleotide that has already been formed during DNA synthesis as one of the usual string of four base letters – **A**, **C**, **G** and **T**. The pattern of methylation is retained when cells replicate and may persist for many generations, both of cells (epigenetic inheritance) and of organisms (transgenerational epigenetic inheritance). Usually, methylation patterns are imposed and then cancelled within one generational cycle. In a bacterium, methylation is an example of an editing that marks bacterial DNA as “self” and hence prevents it being attacked by its own restriction endonucleases. The methylation of a eukaryotic cell’s own DNA serves to ‘lock in’ the inactivation of either a single gene, or a set of genes in the form of an entire chromosome (e.g. the inactivation of one X-chromosome in human females). This involves methylation of **CG** dinucleotides in regions where RNA polymerase binds to DNA. Genes that are *active* in the germ-line (mainly ‘house-keeping’ genes) have less methylation of their **CG** dinucleotides, so if the **C** deaminates the base **U** is formed, which is readily recognized as inappropriate for DNA and is corrected. Thus, an unmethylated **CG** dinucleotide will tend to persist. Accordingly, **CG** dinucleotides are preserved in eukaryotes as the “CpG islands” (macro and mini), which accompany the promoters both of house-keeping genes and of the putative ‘genes’ that compose the ‘hidden transcriptome.’ However, genes that are *inactive* in the germ line (mainly tissue-specific genes) have methylated **CG** dinucleotides. If the **C** deaminates then the base **T** is formed. This mutation is not readily recognized as inappropriate and correction is less likely. The gene may then become active in the germ-line. Since most mutations are disadvantageous, organisms with the mutation may die or produce fewer descendants. Accordingly, the genes evolve to maintain the desired inactive state with fewer **CG** dinucleotides. Hence, the dinucleotide **CG** has come to be under-represented in many eukaryotic genomes. As such, **CG** can serve as a quantitative self-marker, so permitting recognition of the DNAs of microbial invaders (not-self) that have higher levels of methylation - an example of innate immunity.

References

1. Vinogradov AE (2005) Dualism of gene GC content and CpG pattern in regard to expression in the human genome: magnitude versus breadth. *Trends in Genetics* 21:633–639
2. Johnson TB, Coghill RD (1925) The discovery of 5-methyl-cytosine in tuberculinic acid, the nucleic acid of the tubercle bacillus. *Journal of the American Chemical Society* 47:2838–2844
3. Bird A (2002) DNA methylation patterns and epigenetic memory. *Genes & Development* 16:6–21
4. Galagan JE, Selker EU (2004) RIP: the evolutionary cost of genome defence. *Trends in Genetics* 20:417–423
5. Semenov DA (2009) From the wobble to reliable hypothesis. *arXiv:0910.5371v1*. [G-T pairing in DNA is likely to be weaker than G-U pairing in RNA because, when U is methylated (to become T) it inclines to the “keto,” rather than the “enol,” configuration and hydrogen bonding is weaker.]

6. Yamada T, Park CS, Shen Y, Rabin KR, Lacorazza HD (2014) G0S2 inhibits the proliferation of K562 cells by interacting with nucleolin in the cytosol. *Leukemia Research* 38:210–217
7. Brohede J, Rand KN (2006) Evolutionary evidence suggests that CpG island-associated Alus are frequently unmethylated in human germline. *Human Genetics* 119, 457–458
8. Neuberger MS (2008) Antibody diversification by somatic mutation: from Burnet onwards. *Immunology & Cell Biology* 86:124–132
9. Simmen MW (2008) Genome-scale relationships between cytosine methylation and dinucleotide abundances in animals. *Genomics* 92:33–40
10. Takai D, Jones PA (2002) Comprehensive analysis of CpG islands in human chromosomes 21 and 22. *Proceedings of the National Academy of Sciences USA* 99: 3740–3745
11. Krieg AM (2002) CpG motifs in bacterial DNA and their immune effects. *Annual Reviews of Immunology* 20:709–760
12. Fraga MF, et al. (2005) Epigenetic differences arise during the lifetime of monozygotic twins. *Proceedings of the National Academy of Sciences USA* 102:10604–10609
13. Weaver ICG, D'Alessio AC, Brown SE, Hellstrom IC, Dymov S, Sharma S, Szyf M, Meaney MJ (2007) The transcription factor nerve growth factor-inducible protein A mediates epigenetic programming: altering epigenetic marks by immediate-early genes. *Journal of Neuroscience* 27:1756–1768 [Alternative names for the transcription factor NGF1A (rat), include EGR1, KROX24 and ZIF268 (mouse version), and G0S30 (human version)]
14. Flaxman SM, Sherman PW (2000) Morning sickness: a mechanism for protecting mother and embryo. *Quarterly Review of Biology* 75:113–148
15. Meyenn F von, Reik W (2015) Forget the parents: epigenetic reprogramming in human germ cells. *Cell* 161:1248–1251
16. Holmgren C, Kanduri C, Dell G, Ward A, Mukhopadhyaya R, Kanduri M, Lobanenko V, Ohlsson R (2001) CpG methylation regulates the *Igf2/H19* insulator. *Current Biology* 11:1128–1130
17. Willson MF, Burley B (1983) Mate choice in plants: tactics, mechanisms and consequences. *Monographs in Population Biology* 19. Princeton University Press, Princeton, NJ
18. Haig D (2010) Transfers and transitions: Parent-offspring conflict, genomic imprinting, and the evolution of human life history. *Proceedings of the National Academy of Sciences USA* 107:supplement 1, 1731–1735
19. Morales-Ruiz T, Ortega-Galisteo AP, Ponferrada-Marin MI, Martinez-Macias MI, Ariza RR, Roldán-Arjona T (2006) *DEMETER* and *REPRESSOR OF SILENCING 1* encode 5-methylcytosine DNA glycosylases. *Proceedings of the National Academy of Sciences USA* 103:6853–6858
20. Bateson W, Pellow C (1915) On the genetics of “rogues” among culinary peas (*Pisum sativum*). *Journal of Genetics* 5:15–36
21. Suter CM, Martin DIK (2010) Paramutation: the tip of an epigenetic iceberg? *Trends in Genetics* 26:9–14
22. Koonin EV (2014) Calorie restriction à Lamarck. *Cell* 158:237–238
23. Pembrey ME, et al. (2006) Sex-specific, male-line transgenerational responses in humans. *European Journal of Human Genetics* 14:159–166
24. Leopold LE, Heestand BN, Seonga S, Shtessel L, Ahmeda S (2015) Lack of pairing during meiosis triggers multigenerational transgene silencing in *Caenorhabditis elegans*. *Proceedings of the National Academy of Sciences, USA* 112:E2667–E2676. [This transgenerational MSUD is also found in mice and involves proteins that are also known to participate in MSCI. In both processes, conditional on the presence of a pairing partner, the silenced state can be reversed, so facilitating error-correction. Thus error-correction can be delayed chronologically. A DNA sequence marked as ‘suspect’ in one generation can be repaired in a future generation when a suitable pairing partner becomes available (see Chapter 17).]
25. Butler S (1926) In: Jones HF, Bartholomew AT (eds) *The Shrewsbury Edition of the Works of Samuel Butler*. Cape, London, volume 20, p.13
26. Gould SJ (1993) Fulfilling the spandrels of world and mind. In: Selzer J (ed) *Understanding Scientific Prose*. University of Wisconsin Press, Madison, WI, pp 310–336

27. Cedar H, Bergman Y (2009) Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics* 10:295–304
28. Molnar A, Melnyk CW, Bassett A, Hardcastle TJ, Dunn R, Baulcombe DC (2010) Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. *Science* 328:872–875
29. Baechler EC, et al. (2004) Expression levels of many genes in human peripheral blood cells are highly sensitive to *ex vivo* incubation. *Genes & Immunity* 5:347–353
30. Carmi S, Borukhov I, Levanon EY (2011) Identification of widespread ultra-edited human RNAs. *PLOS Genetics* 7:e1002317
31. Liddicoat BJ, Piskol R, Chalk AM, Ramaswami G, Higuchi M, Hartner JC, Li JB, Seeburg PH, Walkley CR (2015) RNA editing by ADAR1 prevents MDA5 sensing of endogenous dsRNA as nonself. *Science* 349:1115–1120. [The dsRNA alarm must be alerted, directly or indirectly, by foreign RNA (see Chapter 15): “Thus, we hypothesized that ADAR1 editing remodeled the secondary structure of endogenous RNA to abrogate the formation of long matched dsRNAs.” It is concluded that “A-to-I editing of endogenous dsRNA is the essential function of ADAR1” (Adenosine deaminase acting on RNA 1), so “preventing the activation of the cytosolic dsRNA response by endogenous transcripts.”]
32. Netzer N, et al. (2009) Innate immune and chemically triggered oxidative stress modifies translational fidelity. *Nature* 462:522–526

Part VII

Information and Mind

Chapter 19

Memory: What Is Arranged and Where?

You, your joys and your sorrows, your memories and your ambitions, your sense of personal identity and free will, are in fact no more than the behaviour of a vast assembly of nerve cells and their associated molecules.

Francis Crick 1994 [1]

How we ourselves acquire, store and transmit mental information – how our minds process (input, manipulate and output) information – would seem to be ultimate questions in bioinformatics. Having solved the corresponding questions on genetic information in the 50s and 60s, Crick moved to brain research in the 1970s. One would have thought the two rules for his success in genetics would still apply: seek the simplest experimental system; find the store first and then work out how it fills and empties. For his experiments Crick had chosen the smallest known biological forms – the viruses that infect bacteria. And he went directly to the genetic store – DNA. When its structure was known, solutions to the input (replication and mutation) and output (transcription and translation) problems, fell neatly into place. But neither rule was to be followed in his brain research.

The Organ *Par Excellence*

The DNA of bacterial viruses is like our own, and is replicated and transcribed in similar fashion. So experiments with viruses could often be extrapolated directly to humans. But, although bacteria display aspects of learning and memory [2], there has long been doubt that data from bacteria or other organisms, simple or complex, can be extrapolated to the human brain. While in their famous 1858 Linnaean Society addresses Charles Darwin and Alfred Wallace agreed on the power of natural selection, Wallace came to draw a line at the human brain and favored explanations that pointed to “an unseen universe” [3]. Likewise, the Oxford philologist Max Müller pointed to language as a unique human attribute that “is our rubicon and no brute will dare to cross it” [4]. For him language represented a difference, not in

degree, but in kind. There was a sharp evolutionary discontinuity – in modern terms “the cognitive equivalent of the Big Bang” [5]. Wallace and Müller offered explanations that we would today label as spiritual, religious or creationist. Thus, “the origins of *H. sapiens* acquires the status of mystery” [6].

In the first edition of *On the Origin of Species* Darwin excluded from discussion the origins both of “life itself” and of “the primary mental powers.” He considered that, *in all respects*, there was evolutionary continuity between man and animals [7]: While “it may be urged that, as man differs so greatly in his mental power from all other animals, there must be some error in this conclusion,” I hold that “there is no fundamental difference between man and the higher mammals in their mental faculties.”

Crick adopted this view [1]. Yet, most of Nature’s evolutionary experiments, if successful, are copied. Birds have wings, but so do bats and insects. Fish have fins, but so do whales (mammals). However, our adaptation, the human brain, appears quantally different. Its evolution by natural selection accelerated to generate a profound discontinuity from antecedent organisms probably because of intra-species competition (human versus human), rather than because of inter-species competition. Other species did not (have not yet) get into the act.

Some argue that there are other examples of quantal evolutionary differences – the neck of the giraffe for instance; but even here there are intermediary forms. The gerenuks (“giraffe antelopes” in the Somali language) act like medium size giraffes by combining a long neck with bipedal browsing. Nevertheless, the neck is the organ *par excellence* of the giraffe, so if one is interested in the neck one should study it in giraffes, not in humans; likewise the brain is the organ *par excellence* of humans. Crick studied the brain in humans when feasible; otherwise, he studied their nearest monkey relatives.

Not DNA

We saw in Chapter 1 that Butler in the 1870s thought that mental and hereditary information would be stored in the same way [8]. Yet when Crick started to work on the brain in the 1970s there were few clues as to the form long-term memory might take. Crick could not go directly to a ‘memory molecule’ and try to work out its structure as he had done with DNA in the 1950s. Instead, he allowed himself to be diverted into neuroanatomy and neurophysiology, studying input and output from the memory store – fascinating subjects with a wealth of novel gee-whiz technologies – but not the store itself.

Since DNA is an excellent long term information store, Crick naturally conjectured that the brain might also store its information – at least, information relating to long-term memory – in this way [9]. So if the store were intracellular, brain cells would be expected to contain DNA in excess of that needed for their normal function. But human brain cells have the same quantity of DNA as other cells of the human body and, indeed, of cells of mammals in general (which include rodents and

other primates). We saw in Chapter 15 that conventional genes make up less than 2% of human DNA and the remaining 98% likely plays important roles. Thus, unless DNA can store information in some way other than has been discussed here, there is no DNA to spare for the memory function [10].

Some have argued for a role of DNA methylation in long term memory [11, 12]. And there are various differences between the nucleic acids of the brain and of other organs, such as enhanced editing of the RNA transcripts of *Alu* repetitive elements [13–15]. However, these intriguing observations could relate more to the brain's special status as an 'immunologically privileged' organ, than to long term memory [16].

On balance, the brain's nucleic acid memory 'cupboard' appears bare. What are the alternatives? We here focus on stable *long term* memory that is distinct from labile *short term* memory, which is disruptable by trauma or electric shock. Long term memory, be it chemical (e.g. an ordered sequence of units; *digital information*), or physical (e.g. vibrations; *analog information*), could be stored in some macromolecule other than DNA, or in some unknown submolecular form [17–19]. The brain is rich in lipids, but biochemical analyses have revealed no candidate memory molecule among lipids. Among the other main biopolymers (RNA, protein, carbohydrate; Fig. 1.4), a case has been made for protein [20]. Thus, we might seek a novel, brain-specific, protein – let's call it 'brainin' – with properties that fit the bill. But the candidate protein turns out to be the familiar tubulin, which is abundantly present in most tissues and has well-characterized roles. There seems to be little tubulin to spare for the memory function [21].

Nor is there yet any evidence for submolecular forms. Physicists tell us that, according to the doctrines of quantum mechanics, the wave function that describes the likely position of a fundamental particle, collapses ("decoheres") when it is observed. To some, the latter implies a conscious observer and some kind of "entanglement" with human consciousness. This extrapolates to "an independent mental modality of the universe that can explain the personal flow of thoughts and, in particular, the experience of self-consciousness" [22]. However, others hold that "no human observer is necessary for wave-function collapse. In a sense it is the environment that performs the observations and measurements" [23]. While there are now imaginative attempts to interpret DNA structure in quantum terms [24], and the case is far from being closed (as we shall see in Chapter 20), for Crick in the 1970s the cupboard also appeared bare regarding non-DNA forms of memory. Rather than go back to first principles and think the whole subject out afresh, he adopted the conventional wisdom. Following the general theme of this book, we will begin by tracing historically how this wisdom acquired its conventionality.

The Conventional Wisdom

In the nineteenth century hereditary information was localized to the cell nucleus (Chapter 1). By ablating parts of the nucleus (e.g. deleting a chromosome) some of that information could be selectively destroyed. In the twentieth century ablation

of parts of the human brain in the course of tumor surgery produced no such clear cut results. Indeed, exhaustive ablation studies in rodents and primates failed to locate brain memory (“engrams”). In 1960 neurophysiologist Karl Lashley concluded [25]:

The engram of a new association, far from consisting of a single bond or neuron connection, is probably a reorganization of a vast system of associations involving the interrelations of hundred of thousands or millions of neurons. ... The conclusion is ... supported by electrical studies, that all the cells of the brain are constantly active and are participating, by a sort of algebraic summation in every activity. There are no special cells reserved for special memories.

This was consistent with long term memory being based on structural changes in the synapses that interconnect neurons – the “synaptic plasticity” hypothesis advanced by Hebb in 1949 [26]. It was this network (“connectionist”) viewpoint that Crick adopted [27]. But, having offered a prize for the best idea on what to do with the college chapels (considered redundant), in the 1990s he concluded his text *The Astonishing Hypothesis* [1] by admitting that in the future “some view closer to the religious one may become more plausible.” Yet there was also the possibility that the facts would “support a new, alternative way of looking at the mind-brain problem that is significantly different from the rather crude materialistic view many neuroscientists hold today.”

Thus, after two decades of brain research Crick’s hopes, albeit unrealized, were as high as in his earlier halcyon days of DNA research. In 1966 in his controversial *Of Molecules and Men* [28], he had attacked both those attempting to explain life as something apart from the laws of physics and chemistry (“vitalism”), and those who believed they could safely remain on one side of the ‘two cultures’ (science versus the arts) divide:

It can be confidently stated that our present knowledge of the brain is so primitive ... that when we do have fuller knowledge our whole picture of ourselves is bound to change radically. Much that is now culturally acceptable will then seem to be nonsense. People with training in the arts still feel that in spite of the alterations made in their lives by technology ... modern science has little to do with what concerns them most deeply. As far as today’s science is concerned this is partly true, but tomorrow’s science is going to knock their culture right out from under them.

Tomorrow

Well ‘tomorrow’ is here. We are in the 21st century. We know enough about Crick to know that the above remark was deeply felt. He was not alone. The anthropologist Gregory Bateson, atheist son of atheist father (William), struggled to articulate something similar in the 1970s [29]:

There is something that human religions are trying to get at that matters. And they get at some of it in many different ways which include vast amounts of nonsense, but we perhaps do not yet have a better way at getting at it, whatever *it* is.

Neurophysiologist Eric Kandel may have thought likewise in 2006 [30]:

In the study of memory storage we are now at the foothills of a great mountain range. ... To cross the threshold from where we are to where we want to be, major conceptual shifts must take place.

Of course, remarks of eminent thinkers prove little. But they do encourage us to lift our eyes to new horizons and to be prepared to listen to those who may offer a fresh perspective. Where and to whom should we, as evolutionary bioinformaticists, be looking and listening, with the aim, not “to knock their culture right out from under them” but, quite simply, to better apply our skills to the understanding of brain function? The answers to be suggested here are (i) that we should look for metaphors to new developments in computer technology – in particular, the internet, and (ii) that we should listen carefully to those who are trying to relate these developments to our own cerebral functioning.

Savants and Information Measurement

The prodigious memory of a village parson was celebrated by the poet Oliver Goldsmith in 1770 [31]:

In arguing, too, the parson own'd his skill,
For, ev'n though vanquish'd, he could argue still;
While words of learned length and thundering sound
Amazed the gazing rustics rang'd around;
And still they gaz'd, and still the wonder grew,
That one small head could carry all he knew.

But the greater stores of information displayed by those with savant syndrome (see Chapter 20) appear to have similar cranial confinements. On the assumption that parsons and savants store their information similarly, then theories of storage cannot exclude the high bar set by the latter.

When contemplating potential storage sites for an object, we take into account whether it is likely to be needed at short notice, its size, stability, and potential compressibility. We know that human long term memory can be stable and is often accessible at short notice, but fitting it into the cranial ‘suitcase’ faces the twin problems that we know neither what form (or forms) it can take, nor its magnitude, save that it is very great and, in those with savant syndrome, amazingly so.

If memory is “a phenomenon of arrangement,” then we need to know what is arranged and where the arrangement is located. Knowing the magnitude of the arrangement might help. As a frame of reference we can take the information content of human DNA as 12×10^9 bits (Chapter 2). Apart from his other memories, a savant described by Treffert and Christensen had perfect recall of 9000 books [32]. If each book had 100 pages each containing 1000 letters then, making the simplifying assumption that each letter is equally probable and, being drawn from a 26 letter alphabet, is worth 5 bits, his memory for books is at least 4.5×10^9 bits. The DNA of just *one* cell, *if it had no other functions*, would be sufficient to encode this.

In 1979 from various assumptions the physicist John von Neumann estimated a lifetime memory accumulation of 2.8×10^{20} bits [33], which, from the above assumptions, could be held in 2.3×10^{10} cells. Assuming 5×10^5 cells per cc of brain [34], this would require a brain volume of 5×10^4 cc (i.e. the volume of a cube with 37 cm sides). Discounting the amount of detail involved in the storage of pictorial information, others have arrived at an estimate much lower than that of von Neumann [35]. However, a recent study concludes that “visual long-term memory has a massive storage capacity for object details” [36]. Nevertheless, given the uncertainty of the assumptions, the idea of cranial DNA storage of long term memory would seem not too far out. But, as discussed above, when we look in the brain for the extra DNA this implies, the cupboard is bare.

It seems obvious that long-term memory is held *within the brain*. This assumption is so widespread that anyone thinking otherwise is deemed cranky, perhaps rightly so. The unconventional, and highly uncomfortable, alternatives are that the repository is elsewhere within the body, or that it resides outside the body (extra-corporeally). The former is unlikely since the functions of other body organs are well understood. Remarkably, the latter has long been on the table [8]. Since it requires *ad hoc* postulates for which there is barely a scrap of evidence, in this chapter we prepare the ground. We further explore scientific aspects of information storage and note that storage capacity can appear normal in some with abnormally small heads, while some with normally sized heads appear to have excessive capacity. In Chapter 20 we consider exceptional humans with scarcely any brain tissue who appear to function well. Somehow the information bar has not been lowered.

Internal Signal Detection and Emission

Despite their apparent solidarity, most objects around us are highly porous. Our mobile phones detect and transmit radio waves even when we are underground. So it does not require the assurance of physicists for us to believe that we are surrounded by moving forms of energy/mass (wave/particles), some of which we can detect biologically (e.g. photons), and some of which react only weakly, if at all, with cells [37]. Indeed, as far as we know, some wave/particles are not detected biologically [38]. The most sophisticated instruments detect certain forms, such as neutrinos, only with difficulty. Every second, enormous numbers of neutrinos pass, apparently without influence, through the earth and through our bodies. Other postulated forms are either undetectable, or their detectability is much debated [39, 40].

The first organisms to evolve probably did not detect photons. They existed in a sea of photons but were blind to them. At some point, through mutation, some molecule within an organism happened to acquire the ability to respond to photons. We surmise that, through natural selection, that organism prospered and its descendents further evolved their photon-detection abilities. So today we see organisms with external organs (eyes) specialized for this task. Likewise, we see ears specialized for sound waves.

The size of such organs often reflects the importance of the corresponding modality for survival. But wave/particle detection does not necessarily correlate with an

external organ. If a signal is not greatly impeded by biological tissue, as is the case with magnetic fields, then receptors can be located *internally*. A wide variety of species, including mammals and birds, orientate using the earth's weak magnetic field [37]. Certain 'magnetotactic' bacteria can concentrate particles of iron oxide – minicompasses – within themselves [41, 42]. With magnetic fields there is little delay between detection and an interpretation that can lead to an adaptive response.

In principle, an organism that can receive a signal from its environment has the potential to evolve to transmit that type (or another type) of signal back to the environment. A firefly detects photons with its eyes, but it can also transmit photons to external eyes. The photons are meant to attract another individual of the same species but of the opposite sex. Males and females of a given firefly species intercommunicate by means of specific movements as they emit light. Hence a firefly of one species will not attempt to mate with a firefly of another species. There is discrimination between individuals within a species in that an individual emitting stronger signals will be preferred. However, a firefly cannot focus its signal on a particular individual.

From this we can conclude that, (i) if at some point in time it became adaptively advantageous for an organism to detect a form of wave/particle, and (ii) if such detection were within the realm of physical possibility, and (iii) if the appropriate mutations occurred, then we might later find in its descendents an external *or internal* structure that would mediate such detection. Furthermore, the organism, by virtue of the same or a different structure, might act as a source of wave/particles that might be transmitted with some degree of specificity to a given target.

Cloud Computing

As indicated by Wallace and Müller, the transmission of information between early hominoids, first orally (spoken language) or gesturally (sign language), and then in written form, is regarded as a major evolutionary event [43]. To corporeal transmission from generation to generation through heredity, was added extra-corporeal transmission through language written on some medium permitting long-term storage, either locally or at a remote site. In the twentieth century, to writings preserved on stone and paper were added those on computer hard-drives [44].

When personal computers appeared in the 1980s they were essentially stand-alone, with their own software and data-storage (memory). It was easy to relate this, metaphorically, to the perception that an individual human brain is a stand-alone entity, with its own software and memory. But concerning "generalized cognition-space," Lenneberg in 1965 has already cast doubt on this [45]:

There is, however, another line of argument that induces many scholars to suspect a close relationship between brain size and intelligence. It is based on purely logical considerations; in fact, the reasoning underlying it is by analogy. The capacities of an electronic computer or desk calculator are directly related to the number of its constituent elements. This engenders the belief that an increase in the number of units in the brain has a similar consequence. However, evidence on this is surprisingly poor. ... Although it is entirely possible that the emergence of language and intelligence are historically related to the increase in size of the brain, the case is certainly not yet irrefutably proven.

In 1972 computer scientist John McCarthy noted there was no intrinsic need for a hard-drive to be near the computer it served [46]. With the advent of the internet this has become reality. Even though the Internet emerged in the 1990s, it has taken two decades for such ‘cloud computing’ to become established. With respect to long-term memory, a stand-alone computer can be regarded merely as a terminal for manipulating data, and one retrieves from, and store files at, some remote location. Local data storage needs can be low.

We can imagine that a biological system – a brain – might store information similarly. We can think of a brain as both a perceptual, and a transmitting, organ. There would need to be a two-way transaction, from the brain to the extra-corporeal information store, and back from that store to the brain’s ‘retina.’ There would have to be an appropriate form of ‘thought ray’ perhaps required to vastly exceed the speed of light (as can occur in through “entanglement” in quantum systems; [47]). Furthermore, one would transmit and ‘see’ with one’s brain only personal messages. Transmission would need to be highly focused – target specific.

From this perspective, the human brain would need to be large, not because of the stable information it contained (long-term memory), but because of the complexity of its operations and its short-term memory needs. To the extent that they could accommodate more complexity, the quality we label as ‘intelligence’ would be greater in larger heads than small, though the latter might have stored, and have access to, no less long-term memory. Thus small heads in which complexity and short-term memory could somehow have escaped impairment would, in theory, be as ‘intelligent’ as large (see Chapter 20).

The Physics of Space

There are imaginative attempts to relate such considerations to the workings of individual brains. The brain is seen as a receptor/transmitter of some form of electromagnetic wave/particle for which no obvious external structure (e.g. an eye) would be needed. Incredible as this may seem, such a scheme, based on the idea of the universe as a holographic information storage device [48] and invoking the “spooky” physical principle known as “non-locality,” has been given “a possible ‘hardware’ implementation” by computer scientist Simon Berkovich [49], and cited with approval by a leading neurophysiologist [50].

Initial, short-term “*synaptic consolidation*” of brain information, dependent on protein synthesis, occurs in a few hours (short-term memory). For long-term memory, later (or in parallel) “*system consolidation*,” independent of protein synthesis, takes weeks or months for completion. The latter could correspond to the period of transmission of information to the postulated remote store. In that recall would appear to be almost instantaneous, under this hypothesis the brain’s power as a receiver would appear even more amazing than its power as a transmitter.

The so-called “dark matter particles” may be relevant. Much of the mass of the ‘universe’ is dark matter that does not absorb light. Light waves from distant parts

pass through dark matter but are hardly affected by it. We know that dark matter exists because, by virtue of its great mass, it slightly affects the movement of visible objects. Whatever dark matter is, it does not seem to consist of the same building blocks as visible matter (protons, neutrons, electrons; [38]). Currently it is hoped that “dark matter particles” will, like neutrinos, be detected by devices hidden deep in the earth to screen out other signals. Indeed, using a device containing ultra-pure sodium iodide crystals, which should emit photons when struck by dark matter particles, Rita Bernabei and her colleagues have claimed positive results [39].

Irrespective of whether the putative ‘thought ray’ wave/particle does relate to dark matter, the point is that there are wave/particles ‘out there’ that we are just beginning to understand. Paralleling the known detection of photons by a complex of coenzyme (retinol) and protein (rhodopsin), it is conceivable that something akin to sodium iodide (acting as a coenzyme) might be bound in a specific configuration by a protein (enzyme) so as to screen out other signals and detect only one type of wave/particle. Just as the retinol-rhodopsin system evolved, so other systems conferring adaptive advantages could evolve.

If we are to take Wallace seriously, this adaptive advantage could have arisen just prior to the emergence of *Homo sapiens*. Here bioinformatic analyses already appear to have shed light [51]. The human genome has “human accelerated regions” (HARs) that were conserved between species, but underwent rapid changes (positive selection) at about the time of the emergence of the primate line leading to *Homo sapiens*. One of these regions (HAR1) is actively transcribed in the brain in the second trimester of embryonic development. Nevertheless, language acquisition appears late both in anthropoid evolution and in human development, seeming to dictate a period of post-natal brain expansion and continuing fetus-like helplessness that contrasts dramatically with the newborn state of our nearest anthropoid relatives [43].

Syntax

As for personifying messages – associating them with a distant address – one possibility would be some form of global position monitoring, since each individual occupies a distinct (albeit moving) position in time and space [52]. Another possibility would be a barcode labeling system (Chapter 12) [53]. Whatever the system, the stored information, wherever it may be, should be *structured* in some way – i.e. it should have a syntax. Presuming this syntax to have evolved before the evolution of our spoken or gestural languages, then the “deep structure” of this *preexisting* syntax could have guided the evolution of the later syntaxes [54, 55]. Butler long ago noted that words are but the means we employ to externalize something much deeper [56]:

To think of a thing they [words] must be got rid of: they are the clothes that thoughts wear – only the clothes. I say this over and over again, for there is nothing of more importance. Other men’s words will stop you at the beginning of an investigation. A man may play with words all his life, arranging them and rearranging them like dominoes. If I could *think* to you without words you would understand me better.

All thought is information. And to exist, information, albeit itself deemed abstract, *must* have a material or physical basis from which, under appropriate conditions, it can be transmuted to other material or physical bases (Chapter 3). The principle of information being a “phenomenon of arrangement” (i.e. some geometric arrangement) would apply to human memory. Something, somehow, would be arranged in its own particular form according to a, perhaps universal, syntax. Like language, the laterality of brain function seems also to be a uniquely human attribute [6]. The transmutation of ideas (held in ‘neural word’ form; “mentalese”) into conventional words, appears as a function of the left side of the brain (Broca’s area). Information for the ideas themselves could reside, at least in the short term, on the right side.

Argument from Incredulity

The “gazing rustics rang’d around” the parson found it incredible that “one small head could carry all he knew.” But we may be sure that none really doubted that it did. Not so the Arabic physician Avicenna (980-1037), who declared the brain cupboard to be bare. In the words of the philosopher William Hamilton, he “denied to the human mind the conservation of its acquired knowledge” and explained “the process of recollection” as “an irradiation of divine light, through which the recovered cognition is infused into the intellect”. Historian David Murray has referred to Avicenna’s “agent intellect” which “acted to transmute sensory knowledge into a kind of knowledge that was not ‘material’ – that is, not associated with any body part” [8].

Perhaps Avicenna had observed that, as we discuss in the next chapter, some of those with decreased brain tissue volume can have normal, or even advanced, ‘intelligence’ (to which memory must make a considerable contribution) [6]. Although one savant described by Treffert and Christensen had a large head with a strange internal anatomy [32], many savants are not exceptional in this respect. Thus, there is no necessary correlation of brain size with memory. Avicenna was probably influenced by the culture of Arabic psychological science which attempted to apply “the laws of physics to the soul-body complex” [57]. Following the death of a body its extra-corporeal long-term memory might persist for some time – the physical equivalent of the term “soul” [58]. Thus, its soul would not *leave* a body at death since it had never been in the body in the first place. Rather, death would leave the soul ‘stranded’ (for some this word would exchange for ‘liberated’).

Nevertheless, the notion that our mental information is stored *outside* the brain is even more incredible than that it is stored inside. Our credulity is further stretched by the need for *ad hoc* postulates – e.g. ultra-fast ‘thought rays’ [47]. Even if we disregard the religious baggage associated with the phrase “divine light,” the arguments of the professionals [48, 49], and discussions of paranormal phenomena by others [59], sound like science fiction, and are unlikely to convince the readers of this book. *Neither do they its author.*

The problem is compounded by activities of the “intelligent design” movement which may interpret the holographic universe idea as showing that it was right all along. But the present discussion should lend no comfort here. It is simply stated that we must shift our preoccupation with the biosphere to appreciate that life evolved as the result of interactions both terrestrial and extra-terrestrial. Our ability to respond to photons, most of which are generated extra-terrestrially, is an obvious example. Whatever is ‘up there’ and whatever is ‘down here,’ they *coevolved* by a process that can be construed as, in its essentials, Darwinian. At the same time we have to admit that the notion of an extra-corporeal base for long-term memory serves a deep-felt human need to suppose that something of ourselves, and of our friends and relatives, will persist after our deaths. So we must proceed with caution.

For John Wheeler (1911-2008), who coined the term “black hole,” there were three fundamentals – energy, matter and information – of which information, he suspected, was the most fundamental. But we should dispel the impression that Biology is supplicant to Particle Physics, through which it hopes to trace the postulated extra-corporeal information store. Noting that it was Galvani’s dancing frogs’ legs that brought electricity to Physics, the supplicancy can easily be reversed. Although Physics currently looks to the Large Hadron Collider to further its knowledge, it may be that studies of human memory will illuminate Physics even more!

Summary

Despite fascinating studies on the input (acquisition) and output (recall) of human mental information, we know not how or where that information is stored for prolonged periods (long term memory). Storage in digital or analog forms could be chemical (e.g. like DNA) or physical (e.g. vibrations), and either molecular or sub-molecular. In the absence of evidence for localized or delocalized storage in our brains, the brain ‘cupboard’ must be deemed bare. Alternative locations are elsewhere in our body (corporeal), or outside our body (extra-corporeal). Evidence on location is quite fragmentary and usually of negative nature: 1. Since brain cells have the same amount of DNA as other tissues, the storage form is unlikely to be DNA, and no comparable macromolecule has yet emerged. 2. Surgical interventions that delete parts of the brain do not remove specific information, the storage of which appears delocalized. 3. Brain size is not necessarily greater in those with phenomenal memories (savant syndrome). 4. Some of those with greatly decreased brain tissue volume can display normal or even advanced intelligence (see next chapter). These observations are consistent with extra-corporeal location. Drawing parallels from the internet – with desktop computers storing their information in a remote location (“cloud computing”) – researchers from various disciplines are looking to physics for support of this strange idea.

References

1. Crick F (1994) *The Astonishing Hypothesis. The Scientific Search for the Soul*. Simon & Schuster, London
2. Morimoto BH, Koshland DE (1991) Short-term and long-term memory in single cells. *FASEB Journal* 5:2061–2067
3. Wallace AR (1889) *Darwinism. An Exposition of the Theory of Natural Selection with Some of Its Applications*. Humboldt, New York, pp 318–321
4. Knoll E (1986) The science of language and the evolution of mind. Max Müller's quarrel with Darwinism. *Journal of the History of Behavioral Sciences* 22:3–22
5. Bates E, Thal DJ, Marchman VA (1991) Symbols and syntax: a Darwinian approach to language development. In: Krasgenor NA, Rumbaugh DM, Schiefelbusch RL, Studdert-Kennedy M (eds) *Biological and Behavioral Determinants of Language Development*, pp 741–742
6. Crow T (2002) Introduction. In: *The Speciation of Modern Homo Sapiens*. Oxford University Press, Oxford, pp 1–20
7. Jacyna S (2009) The most important of all organs: Darwin on the brain. *Brain* 132: 3481–3487
8. Forsdyke DR (2009) Samuel Butler and human long term memory: is the cupboard bare? *Journal of Theoretical Biology* 258:156–164
9. Crick F (1984) Memory and molecular turnover. *Nature* 312:101
10. Kaminsky ZA, Pependikyte V, Assadzadeh A, Petronis A (2005) Search for somatic DNA variation in the brain: investigation of the serotonin 2A receptor. *Mammalian Genome* 16:587–593
11. Holliday R. (1999) Is there an epigenetic component in long-term memory? *Journal of Theoretical Biology* 200:339–341
12. Miller CA, et al. (2010) Cortical DNA methylation maintains remote memory. *Nature Neuroscience* 13:664–666 [Following a learning experience in rats there are rapid changes in the methylation of the CpG islands of genes in the prefrontal cortex.]
13. Paz-Yaacov N, Levanon EY, Nevo E, Kinar Y, Harmelin A, Jacob-Hirsch J, Amariglio N, Eisenberg E, Rechavi G (2010) Adenine-to-inosine RNA editing shapes transcriptome diversity in primates. *Proceedings of the National Academy of Sciences USA* 107:12174–12179
14. Alon S, Garrett SC, Levanon EY, Olson S, Graveley BR, Rosenthal JJC, Eisenberg E (2015) The majority of transcripts in the squid nervous system are extensively recoded by A-to-I edition. *eLife* 4:e05198
15. Luo C, Ecker JR (2015) Exceptional epigenetics in the brain. Non-CpG methylation modulates gene expression in the adult brain. *Nature* 348:1094–1095
16. Louveau A, Harris TH, Kipnis J (2015) Revisiting the mechanisms of CNS immune privilege. *Trends in Immunology* 36:569–577
17. Venema L (2008) The dreamweaver's abacus. *Nature* 452:803–805
18. Heerden PJ van (1963) Theory of optical information storage in solids. *Applied Optics* 2:393–400; (1970) Models of the brain. *Nature* 225:177–178
19. Pribram KH (1971) *Languages of the Brain. Experimental Paradoxes and Principles in Neuropsychology*. Prentice-Hall, Englewood Cliffs, pp 140–166
20. Hameroff S, Penrose R (2013) Consciousness in the universe. A review of the 'Orch OR' theory. *Physics of Life Reviews* 11(1):39–78
21. Georgiev D (2009) Remarks on the number of tubulin dimers per neuron and implications for Hameroff-Penrose Orch OR. *NeuroQuantology* 7(4):677–679. [Cytoplasm is generally rich in tubulin. Because their cytoplasm extends into axons and dendrons, neurons can have a larger cytoplasmic volume than other cell types, so can have a greater absolute content of tubulin.]
22. Korf J (2015) Quantum and multidimensional explanations in a neurobiological context of mind. *The Neuroscientist* 21(4): 345–355
23. Loewenstein WR (2013) *Physics in Mind. A Quantum View of the Brain*. Basic Books, New York, pp. 235–247

24. Brovarets OO, Hovorun DM (2014) Can tautomerization of the A+T Watson-Crick base pair via double proton transfer provoke point mutations during DNA replication? A comprehensive QM and QTAIM analysis. *Journal of Biomolecular Structure & Dynamics* 32:127–154
25. Lashley KS (1960) In search of the engram. In: Beach FA, Hebb DO, Morgan CT, Nissen HW (eds) *The Neuropsychology of Lashley*. McGraw-Hill, New York, pp 478–505
26. Hebb DO (1949) *Organization of Behaviour*. Wiley, New York
27. Crick F, Mitchison G (1995) REM sleep and neural nets. *Behavioural Brain Research* 69:147–155
28. Crick F (1966) *Of Molecules and Men*. University of Washington Press, Seattle
29. Bateson MC (2008) Angels fear revisited: Gregory Bateson's cybernetic theory of mind applied to religion-science debates. In: Hoffmeyer J (ed) *A Legacy for Living Systems. Gregory Bateson as Precursor to Biosemiotics*. Springer, New York, pp 15–25
30. Kandel ER. (2006) *In Search of Memory*. Norton, New York, p. 423 [Butler's eyebrows might have risen at the expression “memory storage,” since to his way of thinking storage was implicit in the term “memory.”]
31. Goldsmith O (1770) *The Deserted Village*. Griffin, London
32. Treffert DA, Christensen DD (2005) Inside the mind of a savant. *Scientific American* 293: 108–113
33. Neumann J von (1979) *The Computer and the Brain*. Yale University Press, New Haven
34. Heller IH, Elliot KAC (1954) Deoxyribonucleic acid content and cell density in brain and human brain tumors. *Canadian Journal of Biochemistry & Physiology* 32:584–592
35. Landauer TK (1986) How much do people remember? Some estimates of the quantity of learned information in long-term memory. *Cognitive Science* 10:477–493
36. Brady TF, Konkle T, Alvarez GA, Oliva A (2008) Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences USA* 105:14325–14329; Cunningham CA, Yassa MA, Egeth HE (2015) Massive memory revisited: limitations on storage capacity for object details in visual long-term memory. *Learning & Memory* 22:563–566
37. Maeda K, et al. (2008) Chemical compass model of avian magnetoreception. *Nature* 453:387–392
38. Wilczek F, Devine B (1987) *Longing for the Harmonies. Themes and Variations from Modern Physics*. Norton, New York, pp 177–195, 315–334
39. Bernabei R, et al. (2008) Investigating electron interacting dark matter. *Physical Review D* 77:023506
40. Stawarz L, Petrosian V, Blandford RD (2010) On the energy spectra of GeV/TeV cosmic ray leptons. *Astrophysical Journal* 710:236
41. Frankel RB, Blackmore RP (1989) Magnetite and magnetotaxis in microorganisms. *Bioelectromagnetics* 10:223–237
42. Størmer FC (2013) Is memory stored in brain neurons and is magnetite involved? *Medical Hypothesis* 81:1170
43. Noll H (2003) The digital origin of human language – a synthesis. *BioEssays* 25:489–500
44. Draaisma D (2000). *Metaphors of Memory. A History of Ideas about the Mind*. Cambridge University Press, Cambridge, p. 13, 155
45. Lenneberg EH (1965) *Biological Foundations of Language*. Wiley, New York, pp. 257–261
46. McCarthy J (1972) The home information terminal. In: *Man and Computer*. Proceedings of International Conference, Bordeaux 1970. Karger, Basel, pp 48–57
47. Rudolph TG (2008) The speed of instantly. *Nature* 454:831–832
48. Bohm D (1981) *Wholeness and the Implicate Order*. Routledge, Kegan, Paul, London, pp 171–213
49. Berkovich SY (1993) On the information processing capabilities of the brain: shifting the paradigm. *Nanobiology* 2:99–107
50. Pribram KH (1991) *Brain and Perception*. Lawrence Erlbaum, Hillsdale, pp xxii–xxiv, 277–278

51. Pollard KS, et al. (2006) An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–172
52. O'Keefe J, Burgess N (2005) Dual phase and rate coding in hippocampal place cells: theoretical significance and relationship to entorhinal grid cells. *Hippocampus* 15:853–866
53. Berkovich SY (2007) Ultimate irreversibility in the universe: continuous holographic recording of every event and biological memory as part of it. In: Berkovich S, Al Shargi HM (2010) *A Constructive Approach to Fundamental Science: Selective Writings*. University Publishers, San Diego, pp 6–56
54. Chomsky N (1965) *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA
55. Gómez DM, Berent I, Benavides-Varela S, Bion RAH, Cattarossi L, Nespor M, Mehler J (2014) Language universals at birth. *Proceedings of the National Academy of Sciences USA* 111:5837–5841
56. Butler S (1878) *Life and Habit*. Trübner, London, p. 84 [This is a quotation from a friend in New Zealand, probably Charles Paine Pauli.]
57. Martino C Di (2007) Memory and recollection in Ibn Sina's and Ibn Rushd's philosophical texts translated into Latin in the twelfth and thirteenth centuries. In: Lagerlund H (ed) *Forming the Mind. Essays on the Internal Senses and the Mind/Body Problem from Avicenna to the Medical Enlightenment*. Springer, Dordrecht, pp 17–26
58. Yrjönsuuri M (2007) The soul as an entity: Dante, Aquinas, and Olivi. In: Lagerlund H (ed) *Forming the Mind. Essays on the Internal Senses and the Mind/Body Problem from Avicenna to the Medical Enlightenment*. Springer, Dordrecht, pp 59–92
59. Talbot M (1991) *The Holographic Universe*. Harper Collins, New York

Chapter 20

Certainty Now Uncertain

*But what about such a proposition as “I know I have a brain?”
Can I doubt it? Grounds for doubt are lacking! Everything
speaks in its favour, nothing against it. Nevertheless it is
imaginable that my skull should turn out empty when it was
operated upon.*

Ludwig Wittgenstein [1]

Mother Hubbard spoke authoritatively when she declared the cupboard bare [2]. She knew the size and form of bones likely to satisfy her dog. She also knew the size of her cupboard. Furthermore, she knew that bones were stable, uncompressible and, most importantly, likely to be needed at short notice by her canine friend. We cannot speak on human long-term memory with equal authority. It often appears stable and can be called upon at short notice, but beyond these facts we cannot now go. We can, however, speak about the size of the metaphorical ‘cupboard’ where, conventional wisdom holds, long-term memory must lie. And if the cupboard seems inadequate, then we have identified a problem.

Allometry

When small organisms evolve into larger, there is usually a proportionate increase in all parts of their anatomy [3]. When head size increases, the neck has to be strengthened to support it. When this rule (“allometry”) is broken, it is inferred that natural selection has operated in some way to disproportionately change organ size. Thus the neck of the giraffe appears disproportionately enlarged. It did not need that neck just to support its head; something else was involved. Likewise, just as bigger frames and bigger brains go together *allometrically*, there is, *in addition*, a disproportionate increase in brain size in our species; we refer to ourselves as “*Homo sapiens*” and assume a positive association between the extra size and our “sapiens” [4].

But “allometric” is somewhat like the comfort words mentioned earlier (Chapter 15). It can be used to dismiss as “simple” the fact that whales have much larger brains than man [5]. Yet, given a compass and steering wheel, why would the control room of a transoceanic liner need to be proportionately larger than that of a diminutive tugboat. If the brain of man, appearing both as an information store and as

a cognitive device, satisfies his needs, it is hard to see why whales need so much more [6]. There is so much at stake, yet so much about the brain to be uncertain about. Even Wilder Penfield's claims to have localized specific memories to distinct parts of the brain, are now being questioned [7]. Thus, we should welcome evidence, however tentative, which brings a claim of certainty into question.

Physiologists can gain evidence on organ function by removal (ablation) of all or part of an organ. If the symptoms of the resulting change in function (a "syndrome") can be reversed by replacing the missing part, then the physiologist's case is strengthened. 'Experiments of Nature' – sometimes cruel ones – add extra support. The physiologist's removal of a pancreas is followed by an increase in blood sugar that can be reversed by administering a pancreatic protein – insulin. Nature's corresponding experiment, diabetes, can be traced back to primary changes in insulin-secreting cells in pancreatic islets. As suggested by savant syndrome (Chapter 19), a fresh look at some of Nature's other experiments now leaves us less confident in long-standing beliefs concerning brain function.

Certainty

The reliability of our mental information has long concerned philosophers. Some hold that it is only the uncertainty of our present knowledge that makes future prediction uncertain. We could predict the future precisely if we both had absolute knowledge of the past, and were smart enough. Neither of these conditions being met, the philosophers seek at least one uncontroversial axiom upon which to build their theoretical constructs. For Rene Descartes it was "I think therefore I am" (*cogito ergo sum*). For Ludwig Wittgenstein the issue was a life-long preoccupation and he left sufficient notes for his disciples to assemble an entire book on the matter – *On Certainty* [1].

A certainty that satisfied Wittgenstein in the early decades of the twentieth century was the soon-disproved impossibility of man landing on the moon. However, to many, his later certainty regarding his head's content of neural tissue remained (and still remains) highly plausible:

Now would it be correct to say: So far no one has opened my skull in order to see whether there is a brain inside; but everything speaks for, and nothing against, its being what they would find there?

This was more than an introspective I-have-a-brain-therefore-I-am. It also applied to others:

I ... am sure, that my friend hasn't sawdust in his ... head, even though I have no direct evidence of my senses to the contrary. I am sure, by reason of what has been said to me, or what I have read, and of my experience. To have doubts about it would seem to me madness.

However, some of Nature's ablation-style 'experiments' now suggest that even this certainty must topple [8, 9].

The Logic

Some years ago there was a TV series – “Candid Camera” – where unsuspecting people monitored by hidden cameras were put in preposterous positions. In one set-up, a car without an engine was discretely pushed into a gas station. An attendant emerged to fill it up. On being asked to check the oil, he duly opened the bonnet and began scratching his head. The hilarious altercation that followed, with the driver angrily remonstrating that the car would have needed an engine to get to the station, and the attendant declaring that one was not present, encapsulates the dilemma now facing neuroscientists worldwide. There are a few people on this planet - for most intents and purposes just like you and me – whose “bonnets” (their skulls), when “opened” (x-rayed), reveal, at best, only 5% of the expected amount of brain tissue!

The logic is elementary. A neuroscientist might find abundant neural tissue in a thousand skulls of normal individuals thus certifying the close correlation between presence of tissue and normality. But the entire house of cards tumbles to the ground when individual one thousand and one is normal, yet has no tissue! Admittedly, in practice the distinction is not so stark. As we shall see, individual one thousand and one is found to have approximately 5% the brain tissue volume that is characteristic of each of the one thousand others. But we suspect that Ludwig would be no less surprised by 5%, than by 0%.

Hydrocephalics

In the 1970s a British pediatrician, John Lorber, reexamined, in adult life, the brains of people who had been treated as children for hydrocephaly, or “water on the brain” [10]. In these children the circulation of fluid in the ventricles of the brain is blocked. Yet fluid continues to enter the brain, so the ventricles expand and brain volume increases. In early life the skull bones allow this expansion, and an enlarged head may be the first sign that all is not well. By inserting a small tube (shunt), surgeons are able to drain off the excessive fluid, so relieving the pressure. If successful, the ventricular expansion is reversed and head size returns to normal. When the operation is carried out early, before the skull bones begin to fuse, a child can expect a normal life. Taking advantage of new brain-scanning technology, Lorber anticipated that the adult brains of treated hydrocephalics would appear normal.

For an example of a normal brain, see the scans at the left in Figure 20.1 [11]. Here brain tissue, encased within the outer skull bones (white), is seen as the usual “gray matter” (gray) and “white matter” (pale gray), and there are small black central areas (labeled “LV”). These are the fluid-containing ventricles. On the far right are the corresponding scans of a hydrocephalic whose surgery had failed and who remained severely impaired neurologically. Almost the entire skull is filled with ventricular fluid (black) and there remains only a small surrounding rim of actual brain tissue. Given the extent of tissue loss, it is astonishing that he is alive. Yet 60

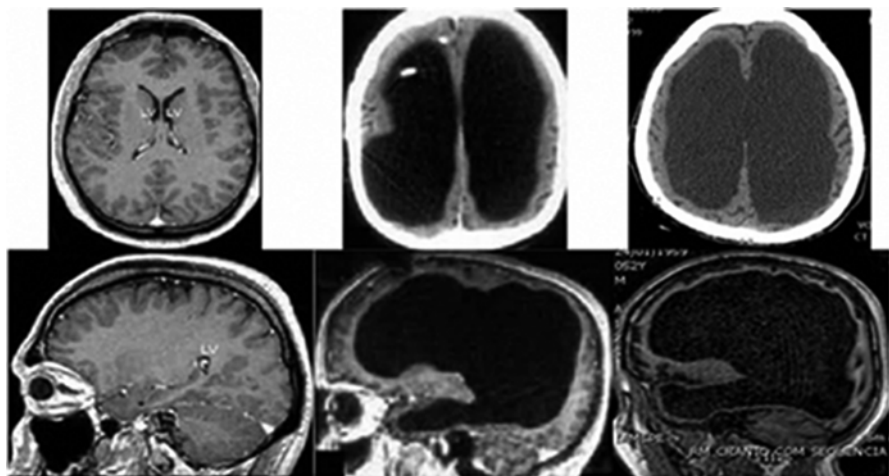


Fig. 20.1 Brain scans. Normal adult appearance (left), with “LV” referring to the small black fluid-filled ventricles. Enlarged ventricles (middle and right). The middle patient is clinically normal, whereas the right patient has had “deep cognitive and motor impairment since childhood.” Reproduced from [11] under Creative Commons Licence

of the 600 cases Lorber studied displayed such extreme brain scans. Ventricular fluid occupied at least 95% of cranial capacity! Relative to other primates, as brain size increases humans acquire proportionately more white matter [12]. Yet it is the white matter that appears most depleted in these cases.

All this is quite shocking. But now for what really rocked the neuroscientists. Half of Lorber’s 60 cases were of above normal intelligence (as determined by standard IQ tests). The central scans in the figure – virtually indistinguishable from the severely impaired case on the right – are representative of this group. And doubtless a candid camera would have caught Lorber’s jaw dropping when, among them, he found a student who was “socially completely normal” and had a first class honors degree in mathematics [10]:

Instead of the normal 4.5 centimetre thickness of brain tissue between the ventricles and the cortical surface, there was just a thin layer of mantle measuring a millimeter or so. The cranium is filled mainly with cerebrospinal fluid. ... I can’t say whether the mathematics student has a brain weighing 50 grams or 150 grams, but it’s clear that it is nowhere near the normal 1.5 kilograms.

Of course, when Lorber presented his findings at a conference in 1980 – under the title ‘Is Your Brain Really Necessary?’ – he met with much scepticism. Journalist Roger Lewin related how experts had pointed to difficulties in the interpretation of brain scans, and had even declared that “Lorber’s style is less scientific than it might be.” But the ripples cast on the neurobiological pond soon faded. Lorber disappeared from view. He died in 1996 at the age of 82. Nevertheless, his results were

considered dramatic enough to warrant a TV documentary. Barry Beyerstein of the Brain Behavior Laboratory at Simon Fraser University quipped [13]:

The telecast employs the ever-popular theme of a brave outsider struggling against a mulish establishment to suggest that, once again, the so-called 'experts' aren't as bright as they think they are. Along the way, the program encourages the misapprehension that there is a huge reserve of unnecessary brain mass that can be casually dispensed with.

Lorber Confirmed

Nearly three decades after Lorber's initial claim, no respectable neuroscientist had risen to his defense and Lorber was still an easy target. In January 2007 an eminent anthropologist set out to dissect Lorber's claim "part by part" [14]. He concluded that the story of the mathematics student was "quite obviously incredible" and that Lorber had really been on a well-intentioned publicity-binge "to show that hydrocephalus is a condition that can be successfully overcome."

A few months later, however, there appeared in *The Lancet* the first of two independent confirmations suggesting that Lorber should not have been so lightly dismissed. Under the title, "Brain of a White-Collar Worker," French neurologists reported a "massive ventricular enlargement" (like the central subject in the above figure) in the brain scan of a civil servant who was married with two children and had come to them with relatively mild neurological symptoms that responded to treatment [15]. And Brazilian neurosurgeons later reported a similar case [11]. We cannot easily shrug these cases off. Rather, we follow the advice of a wise geneticist – "treasure your exceptions" [16] – and consider their implications.

Savants and Microcephalics Don't Scale

Studies of rare individuals with exceptional memory (savant syndrome), and with exceptionally small heads (microcephalics), also argue against the scaling of human brain size with its presumed information-content. Those with savant syndrome tend to have normal-size heads. The presumed high information-content of their heads does not match their head size [17]. The syndrome appears to be inborn, but may sometimes appear in someone with normal memory, after trauma to the head [18]. It may sometimes take the form of an exceptional ability to recall details of one's past life [19]. On the other hand, while most microcephalics are intellectually impaired, there are a few where intelligence is normal; hence, long-term memory is likely to be normal. Again, presumed information-content does not match head size [8].

These arguments were supported, albeit indirectly, by some frank admissions of ignorance by neuroscientists at New York's Columbia University. Fusi and Abbott [20] called for "radical modification of the standard model of memory storage," and

Firestein echoed this in his book – *Ignorance - How It Drives Science* [21]. Furthermore, our brain’s storage capacity for visual detail was seen as enormously greater than previously estimated. This led others to challenge the standard “neural models of memory storage and retrieval” [22].

Why Doesn’t Size Matter?

Given the doubts of these neuroscientists and the growing appreciation that brain size does not scale with information quantity, the case for rechecking possible ways our brains might store their information gains in strength. Broadly, three hypotheses exist:

1. Information relating to long-term memory is held within the brain in some chemical or physical form consistent with current knowledge of brain chemistry and physiology. This ‘standard model’ has many versions that need not detain us [23]. However, we can note new work of Glanzman and colleagues [24] that casts further doubt on the ‘connectionist’ view that synapses are plastic to the extent that they can change (“strengthen”) in parallel with specific memory increase.
2. Information relating to long-term memory is held within the brain in some extremely minute, sub-atomic, form, as yet unknown to biochemists and physiologists. Those who have witnessed in recent decades the vast increase in the power of computers to store large quantities of information in progressively smaller spaces, should not be surprised if evidence for this alternative eventually emerges.
3. Information relating to long-term memory is held outside the brain. Since most non-neural tissues and organs appear unsuited for this task, this extrapolates to long-term memory being *outside* the body – extracorporeal! As noted in Chapter 19, this startling alternative has long been on the table.

Only with two and three, which seem so improbable, do we really confront the dilemma posed by the brain scans of hydrocephalics. Yet some neuroscientists still seek to reconcile the scans with standard models. Indeed, Lorber originally suggested that the “primitive deep structures that are relatively spared in hydrocephalus” may have allowed his subjects to live normal lives, so that “there must be a tremendous amount of redundancy or spare capacity in the brain” [10]. This implied that normally much of the brain is simply idling, ready to act as a backup should the need arise.

Somewhat more convincing was a “plasticity” explanation advanced by Bateson and Gluckman in 2011 [25]. In similar fashion, the Brazilian neurosurgeons invoked the “resilient adaptation of brain networks” associated with “the ability of neuronal tissue to reassume and reorganize its functions” [11]. These plasticity explanations imply that, in keeping with the sometimes amazing recoveries reported for severe brain injuries, an otherwise-occupied part of the brain can change to compensate for a defective part.

However, there must be *rules* for redundancy and plasticity. There *must be limits*. It is a matter of elementary logic that, at some stage of brain shrinkage, these explanations must fail. The drastic reduction in brain mass in the hydrocephalic cases seems to demand *unimaginable levels of redundancy and/or plasticity – superplasticity*. How much brain must be absent before we abandon these explanations and admit that the standard model, however incarnated, will not work?

The Search for the Soul

While the cloud computing metaphor (Chapter 19) is intriguing, it falls far short on evidence. However, the rare hydrocephalic cases described here suggest that we should be cautious when tempted to cast aside the astonishing idea of personal information – long-term memory – being stored elsewhere. There are those, like Berkovich, who urge us to lift our eyes to new horizons [26–28]. Although some may lack a formal training in neuroscience, and sometimes seem to overstate their case, we should listen carefully. After all, Nature is not obliged to conform to our preconceptions. And, as Sherlock Holmes once remarked: ‘*How often have I said to you that when you have eliminated the impossible, whatever remains, however improbable, must be the truth?*’ [29].

The importance of this extends beyond neuroscience and into the clinic. Perhaps in these terms we will someday be able to explain the strange ‘voices’ experienced by schizophrenics, and the bizarre condition described by the American Psychiatric Association (manual of psychiatric conditions; DSM-5) as “disassociative identity disorder” (otherwise “multiple personality syndrome”)? And, of course, when speaking of extracorporeal memory we enter the domain of ‘mind’ or ‘spirit,’ with corresponding metaphysical implications. In the words of Canadian philosopher, Ian Hacking [30], we begin to “secularize the soul.” Indeed, “the scientific search for the soul” was how Francis Crick in 1995 described his own brain studies [31].

Perhaps we should return to 1867 and harken to an exchange between two of Charles Darwin’s contemporaries, Robert Chambers and Alfred Russel Wallace: “My idea is that the term ‘supernatural’ is a gross mistake. We have only to enlarge our conceptions of the natural, and all will be alright” [32]. We chuckle on learning how spiritualists duped such characters. Yet a prominent modern sceptic has argued similarly [33]:

There is no such thing as the paranormal or the supernatural. There is just the normal, the natural, and mysteries we have yet to solve with normal and natural explanations. Until such time as we can provide natural explanations for apparently supernatural phenomena, we need do nothing with such stories because in science we will never be able to explain everything. There is always a residue of unexplained phenomena, and in science it is O.K. to simply say “I don’t know” and leave it at that. Unexplained does not equal supernatural.

Here unexplained phenomena are referred to as “a residue” – an interesting usage (Chapter 9), which, as so often in this book, can be traced back to nineteenth century sources (see Epilogue). Indeed, regarding extracorporeal memory

the possibility now emerges of at least some grains of truth amidst the dross (unexplained ‘residues’) that we poor creatures, imprisoned within the second decade of the twenty-first century, can comprehend no better than Romanes, imprisoned in the latter decades of the nineteenth, could understand “the missing five ounces” of the female brain that was deemed to correlate with a “marked inferiority of intellectual power.” Sadly shaking his head he predicted “with confidence” that it would “take many centuries for heredity” to allow females to “recover the ground which has been lost” [34]. Yet, Romanes, to his credit, anticipated what is now referred to as “theory of mind” (ToM) and showed how it might cast fresh light on things deemed metaphysical.

Theory of Mind

Wittgenstein was certain that his friend did not have sawdust in his head. It was easy to imagine himself ‘in his friends shoes’ because he believed (had a theory) that his friend’s mind was closely similar to his own. Since Wittgenstein was certain that he, himself, did not have his head filled with sawdust, it followed that neither did his friend’s.

Your knowledge of mind begins with your subjective understanding of your own thought processes. To this you attach the word ‘mind.’ This is the seemingly inescapable frame of reference through which you view both the world and the minds of others [35]. Since I share numerous attributes with you (e.g. we belong to the same species) it would seem reasonable to suppose that my thought processes operate in the same way as yours. Based on this ToM, you can make predictions about my behaviour that will generally prove correct. But there remains uncertainty since the way you view *the world* differs from the way you view *other minds*. Thus, the mind of another person – not directly accessible to you – is an *inferred* entity.

A friend of Romanes, the philosopher William Clark Clifford, coined the word “eject” for this inferred entity. This distinguished it from an “object” that could be observed directly by the senses. Thus a ToM, relating to “a subjective order in many respects analogous to my own,” can be referred to as an ejective hypothesis. So there are three ‘ivities:’ *subjectivity*, *ejectivity* and *objectivity*, of which the first is one’s own, introspective, personal frame of reference.

Clifford held that “recognition of a kindred consciousness in one’s fellow-beings – is clearly a condition of gregarious action among animals so highly developed as to be called conscious at all.” Romanes took this much further. Could we ejectively infer what an animal was thinking and so predict its behaviour? Conversely, could an animal ejectively infer what we think and so predict our behaviour? While experimental studies can date the emergence of human ejective powers during childhood, the design of such experiments in animals has proved problematic [36]. No less problematic are ejective predictions concerning what appear to be higher levels of complexity than either animals or ourselves.

The Social Organism

Like Clifford, Romanes located mind within the brain, but he disagreed with Clifford when it came to the scope of ejective thought being limited only to individual natural forms: “Because within the limits of human experience mind is only known as associated with brain, it clearly does not follow that mind cannot exist in any other mode.” Building on Clifford’s “kindred consciousness,” he pointed to society – the “social organism” – that “although composed of innumerable individual personalities, is, with regard to each of its constituent units, a part of the objective world – just as the human brain would be, were each of its constituent cells of a construction sufficiently complex to yield a separate personality.” And it was possible to discern in this social organism some form of subjectivity – a collective psychology or *Zeitgeist*:

Seeing that ideas are often, as it is said, “in the air” before they are condensed in the mind of individual genius, we habitually speak of the “*Zeitgeist*” as a kind of collective psychology, which is something other than the mere sum of all the individual minds of a generation. That is to say, we regard Society as an eject The ejective existence thus ascribed to Society serves as a stepping-stone to the yet more vague and general ascription of such existence to the Cosmos.

The World as Eject

Romanes proposed that the subjective individual can extrapolate ejectively from other individuals he/she knows, to the collective of individuals we deem as a society and, ultimately, to the inanimate world around us – universal mind. Yet, according to Romanes in a paper entitled “The world as an eject” [37], Clifford would exclude the possibility of such a cosmic eject “unless we can show in the disposition of the heavenly bodies some morphological resemblance to the structure of the human brain.” Romanes thought otherwise:

For aught that we can know to the contrary, not merely the highly specialized structure of the human brain, but even that of nervous matter in general, may only be one of a thousand possible ways in which the material and dynamic conditions required for the apparition of self-consciousness can be secured. To imagine that the human brain of necessity exhausts these possibilities is in the last degree absurd. ... To take the particular conditions under which alone subjectivity is known to occur upon a single planet as exhausting the possibilities of its occurrence elsewhere, is too flagrant a use of the method of simple enumeration to admit of a moment’s countenance. ... Therefore, ... the one exhibition of subjectivity furnished to human experience, is the proof thus furnished that subjectivity is possible under *some* conditions. ... The macrocosm does furnish amply sufficient opportunity ... for the presence of subjectivity, even if it be assumed that subjectivity can only be yielded by an order of complexity analogous to that of a nervous system.

What form might this cosmic subjectivity take? In the scheme of things, it is likely that you regard your mind as at a high level – much higher than, say, a stone,

or even your arm muscle. If there is something higher than your mind, it is an abstraction that you can imagine, if at all, only by extrapolation from mind. Considering “that the world ... is at least as susceptible of an ejective as it is of a objective interpretation,” then Romanes supposed that, “just as a mathematician is able to deal symbolically with space in n dimensions, while only able to conceive of space as limited to three dimensions, so I feel that I ought not to limit the abstract possibilities of mental being by what I may term the accidental conditions of my own being.” Indeed, borrowing words from the theological domain, concerning his “world eject” Romanes held [38]:

There is a mode of mind which is not restricted to brain, but co-extensive with motion, is con-substantial and co-eternal with all that was, and is, and is to come; have we not at least a suggestion, that high as the heavens are above the earth, so high above our thoughts may be the thoughts of such a mind as this?

Consciousness

Following this reasoning, perhaps consciousness is to an individual’s memory store, as the postulated social organism (*zeitgeist*) is to media (stone tablets, paper, internet). The quantity of information that can be held (paid attention to) *at any one time*, either by the individual mind (consciousness) or by that of the social organism, is small relative to that which is held, respectively, in individual memory, or in the media. Like a search-light, we dip into our personal memory stores, and perhaps society dips into its societal memory stores. When both ‘sleep’ (the search-light is off) the memory stores appear isolated, but our unconscious minds seem able to continually engage our personal memory stores (e.g. dreams and ‘eureka moments’ on waking). Could there be a societal equivalent of unconsciousness that can engage societal memory stores? Simply stated, if we were all – you, me and media-managers (e.g. librarians) – asleep, would the media remain passive?

Well, it appears to be adaptively advantageous for us to engage in ‘background’ (unconscious) mental activity, both when we are awake and when we are asleep (e.g. ‘eureka moments’). That is, natural selection seems to have favored those who evolved this characteristic. But what would be the equivalent of natural selection at the societal level? It depends on the media. A book on the shelf remains a book on the shelf until someone picks it up. But a book whose contents are ‘out in the cloud’ might be less restricted if it were organized along the panpsychic lines envisaged by some eminent thinkers in diverse disciplines. These include Romanes [34], Richard Dawkins [39], and Stuart Hameroff and Roger Penrose. The latter postulate “a connection between the brain’s biomolecular processes and the structure of the universe” [40].

The critical point – far from comprehended by the present author – seems to be that an attempted measurement that involves the consciousness of an observer (presumed human) causes the binary (wave/particle) coexistence (‘superposition’) of

information to ‘collapse’ (‘decohere’) into unitary particulate form (Chapter 19). Succinctly stated: “Quantum decoherence (i.e. the destruction or collapse of the wave function), is considered a mental quality” [41]. To further this line of thought, we are urged to view the “classical world” as a “first layer” of reality containing “objects that are directly observable, but no longer predictable” because they are driven by events in the unobservable “quantum world.” This constitutes a “second layer” corresponding to states that “obey simple laws” that are “precisely linear” and “possess a wealth of marvelous symmetries” [42].

A Hilbertian List

In general, a problem is first recognized, then addressed. At a conference in Paris in 1900 David Hilbert unveiled his famous list of unsolved problems that he thought should be high on the mathematical agenda in the century to come. Given the state of brain research, a comparable list for the 21st century may be timely [43]. Obviously, we must continue seeking to know whether the ‘cupboard’ really is bare, and continue to characterize genes that are brain-specific, especially those that emerged at the time of the divergence of man from other primates. The following list of topics, some touched upon above, is in no particular order.

1. Why is a newborn human child more like an embryo in its inability to move, eat and control its excretory functions, than related primates?
2. Why is the first memory we can recall often no earlier than the age of three?
3. Why does music charms us [44]?
4. Why do we dream?
5. Why do we laugh when two apparently unrelated items are suddenly brought into conjunction, as in a joke?
6. Why are we delighted when two words rhyme, as in a poem?
7. What is the basis of “savant syndrome” [17]?
8. In what way does someone we label a “genius” differ from someone we label as displaying “savant syndrome”?
9. Is multiple personality syndrome real? If so, what is its basis [30]?
10. Why do some brain diseases show seasonal fluctuations?
11. Do the birth dates of those with either impaired, or superlative, brain function, cluster to any particular season?
12. How does lithium work to relieve depression?
13. How does electroconvulsive therapy (ECT) work to relieve depression?
14. What is the basis of “blindsight” – the ability of some blind people to navigate an obstacle course [45]?
15. Why is the lateralization of brain function so marked in humans, more so in females than males [46]?
16. What are the ‘languages’ neuroscientists tap into when observing oscillatory spikes travelling along nerve fibers?

To close let us return to Francis Crick [47]:

When we understand the nervous system ... the impact will be overwhelming ... it will, for example, essentially make much of our literature unreadable, as much as alchemy is unreadable to modern chemists. I know many of you think that's unsound, I merely leave it with you as my personal opinion that when we understand in detail the nervous system we shall think so differently about ourselves, ... people in that time will hardly be able to understand much of our mental processes today.

And let's give Butler the final word on information superposition [48]:

In the end we shall be driven to admit the unity of the universe so completely as to be compelled to deny that there is either an external or an internal, but must see everything both as external and internal at one and the same time. ... This will knock our whole system over, but then every system has got to be knocked over by something.

Summary

The material bases of information—paper, computer discs—usually scale with information quantity. Large quantities of information usually require large material bases. Conventional wisdom has it that human long-term memory locates within brain tissue, and might be expected to scale with brain size which, in turn, depends on cranial capacity. Large memories, as in those with savant syndrome, should always require large heads. Small heads, as in microcephalics, should always scale with small memories. Neither of these predictions is invariably true. Brain tissue size also depends on ventricle size, which can remain large in some survivors of childhood hydrocephaly, occupying 95% of cranial volume. Yet some of these have normal or advanced intelligence, indicating little impairment of long-term memory. This paradox strongly challenges the scaling hypothesis. The scope of possible explanations (plasticity, redundancy) should not exclude extracorporeal information storage, for which 'cloud computing' appears an apt metaphor. It is intriguing to suppose that this view might relate to the cosmic level of subjectivity, deduced by Romanes from Clifford's "kindred consciousness," which inspires the modern "Theory of Mind" debate.

References

1. Wittgenstein L (1969) *On Certainty*. Anscombe GEM, von Wright GH (eds) Paul D, Anscombe GEM (trans). Basil Blackwell, Oxford, p. 37
2. Martin SC (1805) *The Comic Adventures of Old Mother Hubbard and Her Dog*. Harris, London.
3. Gould SJ (1966) Allometry and size in ontogeny and phylogeny. *Biological Reviews* 41:587–640
4. Jerison HJ (1985) Animal intelligence as encephalization. *Philosophical Transactions of the Royal Society London B* 308:21–35
5. Garn SM (1963) Culture and the direction of human evolution. *Human Biology* 35:221–236

6. Montgomery SH (2013) The evolutionary history of cetacean brain and body size. *Evolution* 67:3339–3353
7. Draaisma D (2015) *Forgetting: Myths, Perils and Compensations*. Yale University Press, New Haven
8. Forsdyke DR (2014) Long-term memory. Scaling of information to brain size. *Frontiers in Human Neuroscience* 8:397
9. Forsdyke DR (2015) Wittgenstein's certainty is uncertain: Brain scans of cured hydrocephalics challenge cherished assumptions. *Biological Theory* 10(4):336–342
10. Lewin R (1980) Is your brain really necessary? *Science* 210:1232–1234
11. de Oliveira MF, Pinto FCG, Nishikuni K et al (2012) Revisiting hydrocephalus as a model to study brain resilience. *Frontiers in Human Neuroscience* 5:181
12. Hofman MA (2014) Evolution of the human brain: when bigger is better. *Frontiers in Neuroanatomy* 8(15):1–12
13. Beyerstein BL (1999) Whence cometh the myth that we only use ten percent of our brains? In: Della Sala S (ed) *Mind Myths: Exploring Everyday Mysteries of the Mind and Brain*. Wiley, New York, pp 3–24
14. Hawks J (2007) Ninety percent of your brain is (not) useless. John Hawks Weblog. http://johnhawks.net/weblog/reviews/brain/development/ten_percent_brain_myth_2007.html
15. Feuillet L, Dufour H, Pelletier J (2007) Brain of a white-collar worker. *Lancet* 370:362
16. Cock AG, Forsdyke DR (2008). *Treasure Your Exceptions. The Science and Life of William Bateson*. Springer, New York
17. Treffert DA (2010) *Islands of Genius*. Jessica Kingsley, London
18. Treffert DA (2014) Accidental genius. *Scientific American* 311(2):52–57
19. McGaugh JL, LePort A (2014) Remembrance of all things past. *Scientific American* 310(2):41–45
20. Fusi S, Abbott LF (2007) Limits on the memory storage capacity of bounded synapses. *Nature Neuroscience* 10:485–492
21. Firestein S (2012) *Ignorance. How it Drives Science*. Oxford University Press, New York
22. Brady TF, Konkle T, Alvarez GA, Oliva A (2008) Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences USA* 105:14325–14329
23. Tsien RY (2013) Very long-term memories may be stored in the pattern of holes in the peri-neuronal net. *Proceedings of the National Academy of Sciences USA* 110:12456–12461
24. Chen S, Cai D, Pearce K, Sun PYW, Roberts AC, Glanzman DL (2014) Reinstatement of long-term memory following erasure of its behavioral and synaptic expression in *Aplysia*. *eLife* 3:e03896
25. Bateson P, Gluckman P (2011) *Plasticity, Robustness, Development and Evolution*. Cambridge University Press, Cambridge
26. Clark A (2008) *Supersizing the Mind: Embodiment, Action and Cognitive Extension*. Oxford University Press, Oxford
27. Majorek MB (2012) Does the brain cause conscious experience? *Journal of Consciousness Studies* 19:121–144
28. Tibbetts PE (2014) Where does cognition occur? In one's embodied/extended environment? *Quarterly Review of Biology* 89:359–368
29. Doyle AC (1890) *The Sign of Four*. Spencer Blackett, London
30. Hacking I (1995) *Rewriting the Soul. Multiple Personality and the Sciences of Memory*. Princeton University Press, Princeton
31. Crick F (1994) *The Astonishing Hypothesis. The Scientific Search for the Soul*. Simon & Schuster, London
32. Wallace AR (1905) *My Life. A Record of Events and Opinions*, vol 2. Chapman and Hall, London, pp 285–286
33. Shermer M (2015) Comment on article “When things happen that you can't explain.” *New York Times* March 5th

34. Forsdyke DR (2015) "A vehicle of symbols and nothing more." George Romanes, theory of mind, information, and Samuel Butler. *History of Psychiatry* 26:270–287
35. Tonini G, Koch C (2015) Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B* 370(1668):20140167
36. Penn DC, Povinelli DJ (2007) The lack of evidence that non-human animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B* 362:732–744
37. Romanes GJ (1886) The world as an eject. *Contemporary Review* 50:44–59
38. Romanes GJ (1885) Mind and motion. *Contemporary Review* 48: 74–94 [Religious texts have so shaped our language over millennia that we now use words – such as 'goodbye' (presumably 'God be with you') – with little thought for religious implications. Thus, it does not follow, since we borrow words from the domain deemed theological, that we necessarily subscribe to the beliefs of that domain.]
39. Dawkins R (1976) *The Selfish Gene*. Oxford University Press, New York, pp 205–206
40. Hameroff S, Penrose R (2013) Consciousness in the universe. A review of the 'Orch OR' theory. *Physics of Life Reviews* 11(1):39–78
41. Korf J (2015) Quantum and multidimensional explanations in a neurobiological context of mind. *The Neuroscientist* 21(4):345–355
42. Dyson, F (2012) Science on the Rampage. *New York Review of Books* 59(6):38–39
43. Adolphs R (2015) The unsolved problems in neuroscience. *Trends in Cognitive Sciences* 19:173–175
44. Ball P (2010) *The Music Instinct – How Music Works and Why We Cannot Do Without It*. Oxford University Press, Oxford. [In the sense that poetry can serve as an error-correcting device (Chapter 4), so it is possible that music likewise somehow serves the internal language of the brain (mentalese; see Chapter 19). And just as cell operations may stall so that error-correcting enzymes can operate (Chapter 17), so brains may be 'stalled' by headaches, so that errors in mentalese can be corrected.]
45. Schmid MC, et al., (2010) Blindsight depends on the lateral geniculate nucleus. *Nature* 466:373–375
46. Priddle TH, Crow TJ (2013) The protocadherin 11X/Y (*PCHD11X/Y*) gene pair as determinants of cerebral asymmetry in modern *Homo sapiens*. *Annals of the New York Academy of Sciences* 1288:36–47
47. Olby R (2009) *Francis Crick. Hunter of Life's Secrets*. Cold Spring Harbor Laboratory Press, New York, pp 315–338
48. Butler S (1965) *Ernest Pontifex or The Way of All Flesh*. Howard DF (ed) Methuen, London, p. 265

Epilogue: To Preserve is Not To Select

His nature was too large, too ready to conceive regions beyond his own experience, to rest at once in the easy explanation, 'madness,' whenever a consciousness showed some fullness and conviction where his own was blank.

George Eliot (1876) [1]

Librarians, booksellers and publishers like books that can be conveniently classified. Books on science go on one shelf. Books on science politics go on another [2, 3]. Most students encounter science first, and read books that give no inkling of the underlying politics. This has been described as “an enormous problem of which students, or people who look only at standard textbooks and sanitized histories of science, are usually unaware. ... You get the impression that the history of science is a totally progressive, orderly, logical development of ideas” [4].

To some extent the science literature has tended to become a pat-on-the-back literature. To pass the peer-review gate, authors are inclined to paint a rosy picture, avoid controversy, and positively emphasize the work of possible gatekeepers. An exception is where controversy is exploited in a holier-than-thou literature that attempts to slip flawed arguments by the reader, while criticizing easy targets (such as people who are no longer alive), and advising the reader to avoid unsavoury authors who try to slip flawed arguments by. Students will inevitably learn about science politics in their later years if they become engaged in one of the most exciting aspects of science – the construction of hypotheses based on prevailing knowledge, and the making of discoveries that extend or refute that knowledge.

Percepts

All scientific knowledge rests on hypotheses. Immanuel Kant argued that percepts without concepts are blind [5]. This point was also made by physiologist John Scott Haldane when writing in 1891 about his uncle, John Burdon Sanderson, a mentor to Romanes [6]:

[He] would say ... that he is very tolerant about theories -- [but] that what really tells is facts. But then what are facts that are essential? It's the theory that determines that. I would simply disregard as trivial and misleading heaps of things which he considers essential, and vice-versa. And even the simplest 'facts' are expressed - perceived - through theory.

We perceive facts in the context of theory. Theories themselves, until the evidence becomes overwhelming (e.g. the theory that the earth is not flat), are unsubstantiated theories, or mere hypotheses. In an ideal world, competing hypotheses would be dispassionately analyzed and the selection of one rather than another would be made with much diffidence. Darwin in 1868 set the standard [7]: "It is a relief to have some feasible explanation of the various facts, which can be given up as soon as any better hypothesis is found."

Torch Passed

But we do not live in an ideal world. In an ideal world, the ideas advanced by Gregor Mendel in 1865 would have been selected by scientists worldwide as a basis for further experimentation. By 1870 the ideas would have entered university-level biology curricula, and students at Cambridge, such as Romanes and Bateson, would have read Mendel along with their Darwin. Yet, as has been told many times, had Mendel never lived progress in genetics would not have been much affected [8]. Thankfully his work was preserved or we would not now know of it. But, alas, preservation does not necessarily lead to selection. At least, in the years before the independent discovery of Mendel's laws in 1900, his work was not castigated:

For historians there remains the baffling enigma of how such distinguished biologists as ...W. Bateson ... could rest satisfied with such a crassly inadequate theory. ... The irony with which we must now read W. Bateson's dismissal of Darwin is almost painful.

This remark in 1983 by Richard Dawkins [9], deservedly one of the most influential scientists of our times, is representative of the multiplicity of attacks on William Bateson that occurred both before and after his death in 1926 [10].

Many of the words Bateson used when establishing the fledgling science of genetics – homozygote, heterozygote, homeotic, epistatic – soon won wide acceptance. But his employment of a common word – residue – in this context (Chapters 9, 20), left people baffled. And while many biologists, to establish broad generalizations, would study regular members of a group of organisms, Bateson

sought out the oddities – a hand with six digits, a white poppy in a field of red – and repeatedly admonished his students to “treasure your exceptions” [11].

It is likely that his research strategy was developed during a friendship with the philosopher/mathematician Alfred North Whitehead, who was born in the same year as Bateson (1861) and whose years of work at Cambridge University (1870-1910) overlapped Bateson's. Indeed, Whitehead's wife helped Bateson in his quest to wed the fair Beatrice, and their young families grew up as neighbors in the village of Grantchester. Whitehead sought to distinguish the discrete (discontinuous) from what was continuous (a major Batesonian theme; Chapter 9), proposed a linkage between quantum theory and consciousness [12], and deeply studied the American philosopher/psychologist/physician William James (1842-1910). The following quotation from an essay published by James in 1890 (with my added bold emphasis) indicates an intriguing congruence of thoughts [13]:

"The great field for discoveries" said a scientific friend to me the other day, "is always the **Unclassified Residuum**." Round about the accredited and orderly facts of every science there ever floats a sort of dust-cloud of **exceptional** observations, or occurrences minute and irregular, and seldom met with, which it always proves less easy to attend to than to ignore. The ideal of every science is that of a closed and completed system of truth. The charm of most sciences to their more passive disciples consists in their appearing, in fact, to wear just this ideal form. Each one of our various *ologies* seems to offer a definite head of classification for every possible phenomenon of the sort which it professes to cover; and, so far from free is most men's fancy, that when a consistent and organized scheme of this sort has once been comprehended and assimilated, a different scheme is unimaginable. No alternative, whether to whole or parts, can any longer be conceived as possible. Phenomena unclassifiable within the system are therefore paradoxical absurdities, and must be held untrue. When, moreover, as so often happens, the reports of them are vague and indirect, when they come as mere marvels and oddities rather than as things of serious moment, one neglects or denies them with the best of scientific consciences. Only the born geniuses let themselves be worried and fascinated by these outstanding **exceptions**, and get no peace till they are brought within the fold. Your Galileos, Galvanis, Fresnels, Purkinjes, and Darwins are always getting confounded and troubled by insignificant things. *Anyone* will renovate his science who will steadily look after the irregular phenomena. And when the science is renewed, its new formulas often have more of the voice of the **exceptions** in them than of what were supposed to be the rules [James' italics].

It cannot be said that progress in genetics would not have been affected had Romanes and Bateson never lived. Their many contributions have been acknowledged and extended. However, Bateson's detractors repeatedly proclaimed that, through his refusal to accept the conventional genic wisdom, he had delayed progress in genetics. I have argued both here (Chapters 8, 9, 17, 18) and elsewhere [10, 11] that, to the contrary, Bateson's thinking was decades ahead of his contemporaries and of many who came after. It is his detractors who may have delayed progress. Nevertheless, the Batesonian torch was passed through the twentieth century by Richard Goldschmidt, by Gregory Bateson, by Michael White, and, with qualifications, by paleontologist Stephen Jay Gould.

Gould

In an article entitled “The Uses of Heresy,” Gould in 1982 introduced a reprint of Goldschmidt’s classic text *The Material Basis of Evolution*. Here Gould described “the counterattack” of the neo-Darwinians that included his Harvard colleague Ernst Mayr’s despair at the “total neglect” by Goldschmidt of the “overwhelming and convincing evidence” against his ideas [14]. In Gould’s opinion Goldschmidt “suffered the worst fate of all: to be ridiculed *and* unread” [Gould’s italics], although his “general vision” was held to be “uncannily correct (or at least highly fruitful at the moment),” and “interesting and coherent, even if unacceptable today.”

In 1980 Gould himself had come close to embracing the Batesonian-Goldschmidtian argument in an article in the journal *Paleobiology* entitled: “Is a new and general theory of evolution emerging?” [15]. But, after a two decade struggle with both the evolution establishment and cancer, Gould recanted, while still maintaining “a hierarchical theory of selection.” In his *The Structure of Evolutionary Thought*, published shortly before his death in 2002, he wrote [16]:

I do not, in fact and retrospect (but not in understatement), regard this 1980 paper as among the strongest ... that I have ever written ... I then read the literature on speciation as beginning to favor sympatric alternatives to allopatric orthodoxies at substantial relative frequency, and I predicted that views on this subject would change substantially, particularly towards favoring mechanisms that would be regarded as rapid even in microevolutionary time. I now believe that I was wrong in this prediction.

Generally courteous when responding to those who did not share his evolutionary views, Gould, like Romanes a century earlier, characterized his many opponents as “ultraDarwinian” fundamentalists, to be contrasted with his few “pluralist” supporters. There are remarkable parallels between Gould and Romanes [17, 18]. Both were of the establishment and center-stage. Steeped in the substance and history of evolutionary science, their writings were welcomed by the leading journals of their days. Both wrote prolifically for the general public, as well as for scientists. Both faced attacks from the very top of their profession – Romanes from Huxley, Wallace, Thiselton-Dyer and Lankester – Gould from Maynard Smith, Mayr and Dawkins. Sadly, both were stricken with cancer in their forties. A difference was that, with modern therapies, Gould survived another two decades, whereas Romanes died at age forty-six. Furthermore, Romanes never recanted and, I suspect, never would have, even if given extra time.

Gould was strongly tested in 1995 by John Maynard Smith, the “Dean of British ultra-Darwinians” [19]:

Gould occupies a rather curious position, particularly on his side of the Atlantic. Because of the excellence of his essays, he has come to be seen by non-biologists as the pre-eminent evolutionary theorist. In contrast, the evolutionary biologists with whom I have discussed his work tend to see him as a man whose ideas are so confused as to be hardly worth bothering with, but as one who should not be publicly criticized because he is at least on our side against the creationists. All this would not matter, were it not that he is giving non-biologists a largely false picture of the state of evolutionary theory.

In the USA Mayr protested that Gould and his allies “quite conspicuously misrepresent the views of [biology’s] leading spokesmen.” Other evolutionists were less restrained [20]:

Evolutionary biology is relevant to a large number of fields – medicine, neuroscience, psychology, psychiatry, cognitive science, molecular biology, etc. – that sometimes have an impact on human welfare. Many scientists in these fields look to Gould, as America’s most famous evolutionist, for reliable guidance on his field, and so the cumulative effect of Gould’s ‘steady misrepresentation’ has been to prevent the great majority of leading scientists in these disciplines from learning about, or profiting from, the rapid series of advances made in evolutionary biology over the last thirty years.

Long after his death Gould was still deemed a man “whose genius for getting things wrong matched the eloquence with which he did so” [21]. It seems not to have occurred to his critics that if Gould had, in fact, provided “reliable guidance,” then it will be *their* “steady misrepresentation” that will have negatively impacted human welfare. While some biohistorians did express dissatisfaction with the prevailing orthodoxy [22–24], following Gould’s demise it was left largely to scientists such as Patrick Bateson (a relative of William and Gregory) and myself, to bear the torch onward into the twenty first century [10, 11, 25, 26]. The task is now a little easier because, as has been shown here, we can flesh out the Batesonian-Goldschmidtian abstractions in both informatic and chemical terms.

Voting with Facts

The ground was well prepared when Watson and Crick presented their model for DNA structure in 1953. The comments of Muller and Wyatt (Chapter 2) show that the possibilities of a helix and base-pairing were “in the air.” The time was ripe. The model soon won wide acceptance. However, many new hypotheses are not so readily accepted.

Another man was convinced that he had the mathematical key to the universe which would supersede Newton, and regarded all known physicists as conspiring to stifle his discovery and keep the universe locked; another, that he had the metaphysical key, with just that hair’s-breadth of difference from the old wards which would make it fit exactly. Scattered here and there in every direction you might find a terrible person, with more or less power of speech, and with an eye either glittering or preternaturally dull, on the lookout for the man who must hear him; and in most cases he had volumes which it was difficult to get printed, or if printed to get read.

These words of Victorian novelist George Eliot, who knew Romanes, and almost certainly knew *of* Butler, were written three decades before Einstein superseded Newton with his special theory of relativity [1].

Acceptance of scientific ideas can be seen as a struggle between two voting democracies – the democracy of the people, and the democracy of the facts. A scientific theory comes to be ‘accepted’ when a major proportion of the ‘experts’ in the field agree that this should be so. This is a true democracy of the people in the sense

that, since all democracies contain non-experts, namely those who are insufficiently qualified to have a vote (e.g. children), then their vote is assigned implicitly to those who are deemed to be qualified.

However, often a theory appears to be consistent with certain facts and inconsistent with others. In the mind of the theoretician the facts *themselves* vote. Theoreticians embrace theories that are consistent with the most facts. Indeed, theories are usually elaborated in the first place to explain facts that seem discordant (i.e. are paradoxical). In many cases it is likely that a theoretician first finds an explanation for a few facts. As the theory develops, more and more facts fall into line (i.e. vote positively). Predictions are then made, and further facts are found that add to the positive vote. Thus, even before attempting publication or seeking experimental support, the theoretician is confronted with an array of voting facts. Even if the votes of the experts end up not supporting his/her theory, the theoretician can still affirm the probable correctness of the theory to the most significant critic – *his/her doubting self* – based on the voting of the facts. Thus, a disconnect can arise between the theoretician and his/her peers.

As described in Chapter 1, Romanes was unable to follow Butler's conceptual leap that heredity was due to stored information (memory). Today, this seems rather obvious. But, much in awe of Darwin, in 1881 Romanes, aged thirty-three and already a Fellow of the Royal Society, scathingly reviewed Butler's *Unconscious Memory* and scoffed at the exposure of the flaws in Darwin's theory [27]:

Now this view ... is interesting if advanced merely as an illustration; but to imagine that it maintains any truth of profound significance, or that it can possibly be fraught with any benefit to science, is simply absurd. The most cursory thought is sufficient to show

One reason for this, of course, is that each person on this planet is somatically different from every other person, and occupies a unique position in time and space. Each person has a unique perspective. Einstein worked in a Swiss patent office when time zones were being established and there was an enormous interest in accurate time-keeping. Switzerland was the center of the clock-making industry. It is likely that he reviewed many submissions on the synchronization of clocks, a major principle underlying relativity theory [28]. This focused his mind in a way denied to his peers in their academic ivory towers. It is to their credit that they did not bar publication of his work in 1905 and, within a decade, most came to recognize its worth. Yet, in all probability, Einstein would still have believed his theory correct, even if his peers had not accepted it.

Butler was also far from the ivory tower. By the time of his death in 1902, the Victorian establishment had scarcely moved an inch towards acknowledging his contribution. As expressed in one very long sentence, this annoyed, but did not overly disconcert [29]:

If I deserve to be remembered, it will be not so much for anything I have written, or for any new way of looking at old facts which I may have suggested, as for having shown that a man of no special ability, with no literary connections, not particularly laborious, fairly, but not supremely, accurate as far as he goes, and not travelling far either for his facts or from them, may yet, by being perfectly square, sticking to his point, not letting his temper run away with him, and biding his time, be a match for the most powerful literary and scientific coterie that England has ever known.

Yes, a powerful coterie, but at heart a kindly one, not malicious. George Bernard Shaw, who counted himself as among “the best brains” that supported Butler, put it this way [30]:

Darwin, a simple-souled naturalist with no comprehension of the abyss of moral horror that separated his little speciality of Natural Selection from Butler’s comprehensive philosophic conception of Evolution, may be pardoned for his foolish estimate of Butler as ‘a clever unscrupulous man,’ and for countenancing the belittling of him by Huxley and Romanes that now seems so ridiculous. They really did not know any better.

False Leads?

Quick acceptance of the double helix model soon led to prediction of the existence of mRNA. But here there was a prominent decoy – rRNA. This provided powerful ammunition to the skeptics. Crick noted [31]:

It is astonishing how one simply incorrect idea can envelop the subject in a dense fog ... the assumption that the ribosomal RNA was the messenger RNA. ... Only the gradual accumulation of experimental facts that appeared to contradict our basic idea could jolt us out of our preconception. Yet we were acutely aware that something was wrong and were continually trying to find out what it was. It was this dissatisfaction with our ideas that made it possible for us to spot where the mistake was. If we had not been so conscientious in dwelling on these contradictions we should never have seen the answer.

Judicial wisdom may not be displayed even by people of scientific eminence who, earlier in their careers, have themselves failed to convince skeptics. Several years after Einstein’s seminal paper on special relativity, Max Planck and other distinguished physicists, while embracing relativity, incorrectly cautioned that Einstein “may sometimes have missed the target in his speculations, as, for example, in his hypothesis of light quanta” (which, of course, was correct) [4].

Planck’s own ideas on the second law of thermodynamics had not been appreciated by Helmholtz, whose ideas, in turn, had also not been recognized. Indeed, Helmholtz made the fundamental point that “new ideas need the more time for gaining general assent the more really original they are” [32]. But he may have been thinking more of the originality of his *own* ideas than those of others!

The work of the seventeenth century polymath Robert Hooke was condemned to oblivion because of “the implacable enmity of Newton But deeper than this is perhaps ... that Hooke’s ideas ... were so radical as to be inassimilable, even unintelligible, to the accepted thinking of his time” [33]. In this way it is argued that peers are obliged to discard ‘premature’ claims, to avoid being overwhelmed by false leads [34].

So, if a theoretician’s peers do not accept, what is an impartial observer to do? In the words of Butler in 1890 [27]:

We want to know who is doing his best to help us, and who is only trying to make us help him, or bolster up the system in which his interests are invested. ... When we find a man concealing worse than nullity of meaning under sentences that sound plausibly enough, we should distrust him much as we should a fellow-traveller whom we caught trying to steal our watch.

Similarly, a century later Dawkins noted [35]:

It's tough on the reader. No doubt there exist thoughts so profound that most of us will not understand the language in which they are expressed. And no doubt there is also language designed to be unintelligible in order to conceal an absence of honest thought. But how are we to tell the difference? What if it really takes an expert eye to detect whether the emperor has clothes?

Butler had a solution [27]:

There is nothing that will throw more light upon these points than the way in which a man behaves towards those who have worked in the same field with himself ... than his style. A man's style, as Buffon long since said, is the man himself. By style, of course, I do not mean grammar or rhetoric, but that style of which Buffon again said that it is like happiness, and *vient de la douceur de l'âme*. ... We often cannot judge of the truth or falsehood of facts for ourselves, but we most of us know enough of human nature to be able to tell a good witness from a bad one.

More specifically, four questions need to be asked. Is the theoretician sane? Is the theoretician objective? Is the theoretician honest? Does the theoretician have a credible track record in the field? A negative answer to these questions will not disprove the theory. A positive answer, however, should serve to counter a negative vote from alleged experts.

For Bateson and Goldschmidt, the answers are quite positive. Their demeanors, relationships and writings attest to their sanity, objectivity and honesty, the latter being most evident towards the person they would most likely deceive, *themselves*. Like Romanes before them, having been immersed in evolutionary science for many decades, they were both very well prepared. Yet, like Romanes before them, they were denied the benefit of the doubt and were subjected to personal attacks. George Eliot wrote [1]:

Like Copernicus and Galileo, he was immovably convinced in the face of hissing incredulity; but so is the contriver of perpetual motion. We cannot fairly try the spirits by this sort of test. If we want to avoid giving the dose of hemlock or the sentence of banishment in the wrong case, nothing will do but a capacity to understand the subject-matter on which the immovable man is convinced, and fellowship with human travail, both near and afar, to hinder us from scanning any deep experience lightly.

People are not impartial. As suggested here and elsewhere [3], people act from a variety of motives other than that of seeking the truth. They will not diversify and hedge their bets – an admonition that even those in Wall Street are prone to forget. And sometimes, despite the best intentions, they just do not find time to do their homework. Writing in 1932 to the US geneticist Thomas Hunt Morgan, Ronald Fisher noted [36]:

One of the chief reasons why, in spite of raising so much dust, we are not making in this generation more rapid progress, is that we do not really give ourselves time to assimilate one another's ideas, so that all the difficult points, the things really worth thinking about, have to be thought out independently, with great variations in efficiency and success, some hundreds of times.

Marketing

While Madison Avenue-style blandishments convince that round pegs fit square holes, and square round, the bubble of complacency grows and grows. The conceptually challenging, seemingly abstract, arguments of Butler, Romanes, Bateson and Goldschmidt denied Darwinian dogma. For this they were dismissed by generations of biologists bewitched by genes and all that their marketers promised. With overwhelming rhetoric to the contrary, we were slow to appreciate the deep truths they were attempting to convey [10, 11, 17].

And so on into the twenty first century. Scientific understanding continues to advance at a pedestrian pace with much waste of time and energy, and of public and private funds. As indicated in the Prologue, commercial interests continue to exert an undue influence on what science is supported and published. A Canadian bioethicist comments [37]:

Conditions that are attached to government funding can affect the purposes and values upheld, especially when those conditions require academic-industrial partnerships for research to be eligible for funding, as in the case of the Canadian government's \$300 ... million investment in a series of genomics research centres (Genome Canada). Structuring funding in this way leaves out the funding of research that will not result in marketable products, and excludes those researchers who undertake it.

It should not be imagined that the situation will easily change. In 1931 the author of *Brave New World*, Aldous Huxley (a grandson of Thomas Huxley), wrote to Fisher [38]:

The really depressing thing about a situation such as you describe is that, the evil being of slow maturation and coming to no obvious crisis, there will never be anything in the nature of a panic. And as recent events only too clearly show, it is only in moments of panic that anything gets done. Foresight is one thing: but acting on foresight and getting large bodies of men and women to accept such action when they are in cold blood – these are very different matters.

Thus, the western democracies, driving to ever increasing prosperity with individuals liberated to indulge their interests and desires, appear somewhat like headless monsters – a situation infinitely preferable to the totalitarian alternative, *provided* there can be appropriate feedbacks from periphery to center so that things do not spiral out of control. Sometimes the 'bubbles' burst rapidly as when panics seize Wall Street. More often, there is "no obvious crisis" – people with AIDS, cancer, malaria or schizophrenia slip quietly away, one at a time. We all know the remedy. The bursting of such bubbles requires informed individuals with the courage peacefully to challenge the conventional wisdom, open-minded information gatekeepers, a free press, and an attentive and responsive executive that is attuned to history.

Joining the Dots

The “intelligence failure” of September 11th 2001 occurred while the first edition of this book was in preparation. It turned out that, prior to the event, there had been substantial clues, but the ‘Mendels’ and ‘Butlers’ in government agencies had been overlooked [10, 39]. The ‘authorities’ had failed to sift and weigh. They had failed to ‘join the dots.’ Likewise, as argued here, the biomedical research enterprise, while “concealing difficulties and overlooking ambiguities” has sometimes failed similarly. Sadly, there has been little recognition of this. Indeed, the very opposite, as the outcry of the complacent against Gould has revealed.

Hailing the sequencing of the human genome as again showing the vigor of free-market systems, commentators can contrast this with the attack on genetics by Lysenko who in the 1930s, with the support of Joseph Stalin, sent his competitors to their deaths [40]. Yet few are aware of the emergent ‘new Lysenkoism’ in the West that, while sparing their bodies, has forced academic suicide (i.e. loss of research funding, tenure, etc.) on those who could see beyond their noses and refused to embrace the prevailing orthodoxy [2, 3]. Is “suicide” too strong a word? Well, “It is suicidal” was how Marshall Nirenberg’s plan (ultimately successful) to decipher the genetic code was greeted by a colleague at the US National Institutes of Health in 1959 [41].

While Romanes, a friend of the editor of *Nature*, had no trouble getting his anonymous editorials and signed articles accepted [42], in the early decades of the twentieth century William Bateson battled the peer-review gatekeepers who controlled access to publication. He had to content himself with the less visible *Reports to the Evolution Committee of the Royal Society*. Butler had to self-publish most of his works on evolution, and negative reviews by the Victorian establishment ensured that he would suffer repeated financial losses. In his early days, Fisher submitted a paper on the probable error of the correlation coefficient to the journal *Biometrika*. The Editor, the biometrician Karl Pearson, rejected it since he was “compelled to exclude all that I think is erroneous on my own judgement, because I cannot afford controversy” [43].

Chargaff himself was subject to the “hissing incredulity” of the funding agencies. As an innovative and vigorous experimentalist (he lived to be 96), it is likely that knowledge of the base compositional regularities he described would by now have been much further advanced had he not been forced into retirement [2]. And, just as the victors in war write the history books, so it can be the successful marketers and their disciples who write the history of science. Like the hero in *Beau Geste* who propped up dead soldiers to suggest that the fortress was well defended [44], so the marketers – sometimes biology’s “leading spokesmen” – tend to prop up the careers of those disciples who will defend their views, even in the face of mounting evidence to the contrary. The “troubled life” of Niels Jerne is a recent example [45-47].

Cogito Ergo ...?

It is my hope that, having completed this book, you will have captured my enthusiasm for EB and my belief that studies in this area will greatly advance progress towards solutions to major problems of humankind. But this Epilogue implies caution. If you intend to engage in research, be prepared for difficulties. If you have a penchant for joining the dots, and find that you can see beyond your nose, then be prepared for a hidden life with quick passage to an academic tomb inscribed *cogito ergo non sum*.

Yet, there is a ray of hope suggesting that times may now be different from when Aldous Huxley penned those depressing words above. Through the Internet amazing new lines of communication have opened to all, and developments in search-engine design promise a sifting and manipulation of information that would have been unimaginable to our fathers. The headless monster (Adam Smith would have called it “the invisible hand”) of our biomedical/industrial enterprise, has generated more data than it can handle. This is preserved in open databases worldwide. Inspired programmers are writing ‘open source’ software packages that can be downloaded with a click. So, if we so decide, that which is preserved can be selected with ease. The playing field is being leveled in strange ways.

But ease of selection does not ensure that selection will be wise. Canadian chemist John Polanyi points out that although now, more than ever before, information is freely available to all, it “only *appears* to be freely available to all. In fact, it is available only to those who understand it’s meaning, and appreciate its worth” [3]. Hopefully, this book will help you understand its meaning and worth. Voltaire in 1770 wrote “*On dit que Dieu est toujours pour les gros bataillons*” [48]. Today, *les petits bataillons* may be similarly favored. Even so, the difficulties will be great, and you may try, and fail, as did George Eliot’s sad figure, Casaubon, in her novel *Middlemarch*. But, unlike Casaubon, through trying you may come to understand, and hence become able to communicate, the wisdom of others, so that the works of tomorrow’s Mendels, Herings, Butlers, Romanes, Batesons, Goldschmidts and Chargaffs will not go unrecognized. Let George Eliot close [49]:

Her full nature ... spent itself in channels which had no great name on the earth. But the effect of her being on those around her was incalculably diffusive: for the growing good of the world is partly dependent on unhistoric acts; and that things are not so ill with you and me as they might have been, is half owing to the number who live faithfully a hidden life, and rest in unvisited tombs.

References

1. Eliot G (1876) *Daniel Deronda*. William Blackman, London [Followed in 1879 by *Impressions of Theophrastus Such* (Blackwood, Edinburgh) this was one of Eliot’s last books. She worked on *Theophrastus Such* from June to November 1878 and corrected proofs in March 1879 when Butler was writing the preface of *Evolution, Old and New*. Eliot’s partner, George Lewes,

played a major role in establishing the Physiological Society. Romanes, who was Secretary to the Physiological Society and attacked Butler, was one of those privileged to attend Eliot's 'court' at The Priory on Sunday afternoons. Butler probably served as her model for one of the friends of Theophrastus who antagonized Grampus (Darwin), and spoke of humans being superseded by machines (Chapter 1); the latter is the theme of her chapter entitled "Shadows of the Coming Race."]

2. Chargaff E (1978) *Heraclitean Fire. Sketches from a Life before Nature*. Warner Books, New York [After being prematurely 'retired,' Chargaff obtained some laboratory space at the Roosevelt Hospital, New York, until 1992.]
3. Forsdyke DR (2000) *Tomorrow's Cures Today?* Harwood Academic, Amsterdam
4. Wilszek F, Devine B (1987) *Longing for the Harmonies*. Norton, New York, pp 111, 209
5. Kant I (1781) *The Critique of Pure Reason*. Guyer P, Wood AW (eds) Cambridge University Press, Cambridge (1998)
6. Haldane JS (1891) Letter to Louisa Trotter. 3 December. In: Romano T (2002) *Making Medicine Scientific. John Burdon Sanderson and the Culture of Victorian Science*. John Hopkins University Press, Baltimore (2002), p. 128 [JS Haldane's son, JBS Haldane, was named after his great uncle.]
7. Darwin C (1868) Letter to Alfred Wallace. 27 February. In: Darwin F, Seward AC (eds) *More Letters of Charles Darwin*. John Murray, London (1903), p 301
8. Olby RC (1966) *Origins of Mendelism*. Schocken Books, New York
9. Dawkins R (1983) Universal Darwinism. In: Bendall DS (ed) *Evolution from Molecules to Man*. Cambridge University Press, Cambridge, pp 403–425
10. Forsdyke DR (2001) *The Origin of Species, Revisited. A Victorian who Anticipated Modern Developments in Darwin's Theory*. McGill-Queen's University Press, Montreal
11. Cock AG, Forsdyke DR (2008) "Treasure Your Exceptions." *The Science and Life of William Bateson*. Springer, New York
12. Whitehead AN (1929) *Process and Reality. An Essay in Cosmology. Gifford Lectures Delivered in the University of Edinburgh During the Session 1927–1928*. Cambridge University Press, Cambridge
13. James W (1890) The hidden self. *Scribner's Magazine* 7:361–374. [Reproduced in a volume of collected essays in 1897.]
14. Gould SJ (1982) The uses of heresy. Forward to reprint of: Goldschmidt R (1940) *The Material Basis of Evolution*. Yale University Press, New Haven, pp xiii–xlii
15. Gould SJ (1980) Is a new and general theory of evolution emerging? *Paleobiology* 6:119–130
16. Gould SJ (2002) *The Structure of Evolutionary Thought*. Harvard University Press, Cambridge, MA, pp 1002–1003
17. Forsdyke DR (2004) Grant Allen, George Romanes, Stephen Jay Gould and the evolution establishments of their times. *Historic Kingston* 52:94–98
18. Schwartz JS (2010) *Darwin's Disciple, George John Romanes. A Life in Letters*. Lightning Rod Press, Diane Publishing Company, Darby, PA
19. Smith, JM (1995) Genes, memes and minds. *The New York Review of Books* 42(19):17–19 [The terms "ultra-Darwinian" and "neo-Darwinian" were used by Romanes to disparage Wallace's and Weismann's inflexible advocacy of the power of natural selection.]
20. Tooby J, Cosmides L (1997) [A letter to the editor of *The New York Review of Books* that was not accepted for publication: see <http://cogweb.ucla.edu/Debate/>]
21. Dawkins R (2015) *Brief Candle in the Dark: My Life in Science*. Bantam Books, New York
22. Adams MB (1990) La génétique des populations était-elle une génétique évolutive? In: Fischer J-L, Schneider WH (eds) *Histoire de la Génétique*, pp 153–171. ARPEM, Paris [Also: Adams MB. Little evolution, big evolution. Rethinking the history of population genetics. (unpublished personal communication, 2003).]

23. Provine WB (1992) Progress in evolution and the meaning of life. In: Waters CK, Helden A van (eds) *Julian Huxley, Biologist and Statesman of Science*. Rice University Press, Houston, pp 165–180
24. Provine WB (2014) *The ‘Random Genetic Drift’ Fallacy*. Create Space Independent Publishing Platform, Charleston, USA
25. Bateson P (2002) William Bateson: a biologist ahead of his time. *Journal of Genetics* 81:49–58
26. Huang S (2015) New thoughts on an old riddle: what determines genetic diversity within and between species. *arXiv* 1510.05918 [Huang notes that: “While population genetics may not be short of mathematics, the foundations for its mathematics are all man-made assumptions that are all self-evidently untrue or only true in a very limited sense.”]
27. Butler S (1914) *The Humour of Homer and Other Essays*. Kennerley, New York, pp 245–313 [The quotation is one of Romanes’ more moderate.]
28. Galison P (2003) *Einstein’s Clocks, Poincaré’s Maps: Empires of Time*. Norton, New York
29. Butler S (1985) *The Notebooks of Samuel Butler*. Jones HF (ed) Hogarth Press, London, pp 360–378
30. Shaw GB (1936) Samuel Butler. Introductory Essay. In: Butler S, *The Way of All Flesh*, Oxford University Press, London, pp v–xiii [Shaw added that “Butler and his clique belittled Grant Allen, one of the most amiably helpful men that ever lived, and one, moreover, who recognized Butler as a man of genius.” In turn, Butler heaped “scorn on Allen because he was not at once ready to declare that Butler was right about evolution, and Darwin a disingenuous sciolist.”]
31. Crick F (1988) *What Mad Pursuit. A Personal View of Scientific Discovery*. Basic Books, New York, p 140
32. Barber B (1961) Resistance by scientists to scientific discovery. *Science* 134:596–602
33. Sacks O (2001) *Uncle Tungsten*. Knopf, New York, pp 104–105
34. Hook EB (2002) *Prematurity in Scientific Discovery. On Resistance and Neglect*. University of California Press, Berkeley
35. Dawkins R (2003) *A Devil’s Chaplain: Reflections on Hope, Lies, Science and Love*. Houghton Mifflin, Boston, p 48
36. Fisher RA (1932) Letter to T. H. Morgan. In: Bennett JH (ed) *Natural Selection, Heredity and Eugenics. Including Selected Correspondence of R. A. Fisher with Leonard Darwin and Others*. Clarendon Press, Oxford (1983), p 239
37. Somerville MA (2002) A postmodern moral tale: the ethics of research relationships. *Nature Reviews Drug Discovery* 1:316–320
38. Huxley AL (1931) Letter to R. A. Fisher. In: Bennett JH (ed) *Natural Selection, Heredity and Eugenics. Including Selected Correspondence of R. A. Fisher with Leonard Darwin and Others*. Clarendon Press, Oxford (1983), p 220
39. Birchler JA (2015) Mendel, mechanism, models, marketing and more. *Cell* 163:9–11
40. Popovsky M (1984) *The Vavilov Affair*. The Shoe String Press, Hamden. [Stalin wanted quick results in agriculture, which the slow but sure methods of the new genetics could not meet. Enter Lysenko with a vernalization ‘cure’ that was easy to market. He seized the crown and ruled for decades: “There was something mystical about Lysenko’s career, for regardless of failures, Lysenko kept his position. Despite all the promises made by Lysenko’s ‘progressive biologists’ and ‘true Darwinists,’ harvest yields in the 1930s did not increase.”]
41. Nirenberg M (2004) Historical review: deciphering the genetic code – a personal account. *Trends in Biochemical Sciences* 29:46–54
42. Meadows AJ (1972) *Science and Controversy. A Biography of Sir Norman Lockyer*. MIT Press, Cambridge, pp 209–237
43. Bennett JH (1983) Notes. In: *Natural Selection, Heredity and Eugenics. Including Selected Correspondence of R. A. Fisher with Leonard Darwin and Others*. Clarendon Press, Oxford, p 118
44. Wren PC (1927) *Beau Geste*. Longmans, Toronto

45. Söderqvist T (2003) *Science as Autobiography. The Troubled Life of Niels Jerne*. Yale University Press, New Haven [This meticulous study sets a high historiographic standard.]
46. Eichmann K (2008) *The Network Collective: Rise and Fall of a Scientific Paradigm*. Birkhäuser, Basel [Supporting Söderqvist, one of Jerne's disciples reflects on the errors of his mentor.]
47. Forsdyke DR (2012) Immunology (1955-1975): The natural selection theory, the two signal hypothesis, and positive repertoire selection. *Journal of the History of Biology* 45:139–161 [A personal view of the Jerne era.]
48. Voltaire (1770) Letter to F. L. H. Leriche. 6th February. In: *The Complete Works of Voltaire*, Vol 120. The Voltaire Foundation, Banbury, 1975, p 18 [The full translation reads: "The number of wise men will always be small. It is true that it is increasing, but it is nothing compared with the number of fools and, although they say it is regrettable, God is always for the big battalions. It is necessary that honest people quietly stick together. There is no way their little force can attack the host of the closed-minded who occupy the high ground."]
49. Eliot G (1874) *Middlemarch. A Study of Provincial Life*. Haight GS (ed) Houghton Mifflin, Boston (1956) p 613

Appendix A: What the Graph Says

Since the probability of two bombs being on the same plane is very low, the wise statistician always packs a bomb in his/her luggage.

Anonymous (before the modern era of airport security)

Evolutionary bioinformatics (EB) requires the handling of large amounts of sequence data, sometimes from multiple biological species. This is greatly assisted by plotting the data points on a graph. The points are often widely scattered, and the volume of data – sometimes many thousands of points – can seem overwhelming. Fortunately, there are software packages (like Microsoft Excel) that will plot the points and draw the statistically best lines through them.

The graphs can sometimes be understood immediately without resort to numerical analysis. But confidence increases when visual interpretations are numerically confirmed. Thus, throughout this book graphical displays include values for slopes of lines, squares of the correlation coefficients (r^2), and probabilities (P). These capture key features of what might otherwise seem a hopeless jumble of points. Out of a cloud meaning emerges, sometimes discerned only as faint, yet statistically significant, signals.

Often lines are linear (rectilinear), but sometimes they are curved (curvilinear). Often rectilinear lines are tilted, but sometimes they are horizontal. Sometimes no linear interpretation is evident, but clusters of points can be demarcated from others by drawing circles around them. Without resort to a statistics text, the essence of what the various values tell us can be shown by constructing dummy graphs and then applying a standard software package.

Horizontal Lines

Table [A.1a](#) shows a spreadsheet with X-axis values increasing from 1 to 20. Four sets of Y-axis values were preselected as detailed below.

Table A.1 A spreadsheet of dummy data values (a) for generation of graphical plots that are either horizontal (columns 1 and 2), or sloping (columns 3 and 4). It is intended that, around the line that best fits the points (as determined by a standard computer program), there be either a *small* scatter of points (columns 1 and 3), or a *large* scatter of points (columns 2 and 4). Various statistical parameters are shown (b) below the related columns. Scatters around regression lines are measured as standard errors of the estimate (SEE) or standard errors of the mean (SEM)

(a)				
	Y-axis			
X-axis	Column 1	Column 2	Column 3	Column 4
1	6.0	6.0	6.0	6.0
2	5.9	5.0	6.9	6.0
3	6.1	7.0	8.1	9.0
4	5.8	4.0	8.8	7.0
5	6.2	8.0	10.2	12.0
6	6.0	6.0	11.0	11.0
7	5.9	5.0	11.9	11.0
8	5.8	4.0	12.8	11.0
9	6.1	7.0	14.1	15.0
10	6.2	8.0	15.2	17.0
11	6.1	7.0	16.1	17.0
12	6.2	8.0	17.2	19.0
13	5.9	5.0	17.9	17.0
14	5.8	4.0	18.8	17.0
15	6.0	6.0	20.0	20.0
16	6.1	7.0	21.1	22.0
17	5.9	5.0	21.9	21.0
18	6.0	6.0	23.0	23.0
19	6.0	6.0	24.0	24.0
20	6.0	6.0	25.0	25.0
(b)				
Mean	6.0	6.0	15.5	15.5
Standard error	0.133 [SEM]	1.333 [SEM]	0.133 [SEE]	1.333 [SEE]
r ²	<0.0001	<0.0001	0.999	0.951
Slope=a	0.003	0.003	1.003	1.003
P (slope probability)	0.954	0.954	<0.0001	<0.0001
y ₀	6.0	6.0	5.0	5.0

In Table A.1 Y-axis values for four plots of different sets of Y values against the X values are displayed in columns 1 to 4. Numbers were placed in the first two Y columns with the intention of plotting two horizontal lines that would cross the Y-axis at a value of 6 (i.e. Y₀, the value for Y when the X value was zero, would be 6).

For column 1, the numbers were allowed to deviate very little from the mean of 6. Thus, in the first row, 6 was entered. Then, in the second row, 5.9 was entered. Then, to compensate for this, so that an average of 6.0 was sustained, in the third row 6.1

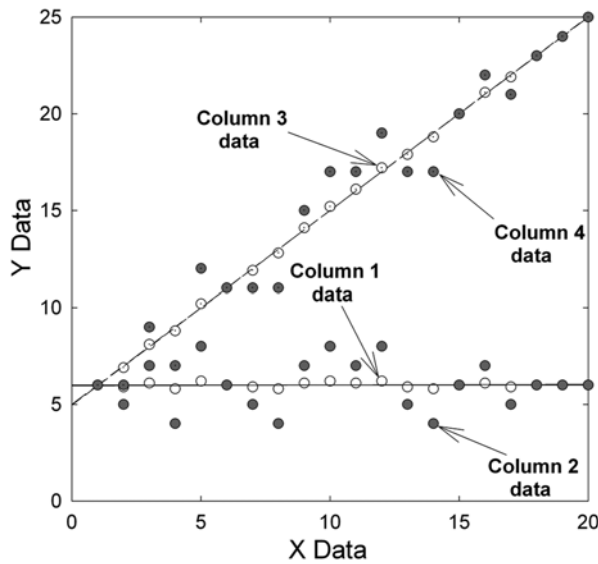


Fig. A.1 Computer-generated lines that best fit the plots of the data shown in Table A.1. Open circles (column 1 data); gray circles (column 2 data); open circles with central dot (column 3 data); gray circles with central dot (column 4 data). There are two horizontal lines for columns 1 and 2 and two sloping lines for columns 3 and 4, but the line pairs are so close that they cannot be distinguished (i.e. they overlap)

was entered. This was continued to generate 20 values for Y with minima no lower than 5.8 and maxima no higher than 6.2.

For column 2, the numbers were allowed to deviate further from the mean of 6. Thus in the second row 5 was entered. To compensate for this, so that an average of 6 was sustained, in the third row 7 was entered. This was continued so generating 20 values for Y with minima no lower than 4, and maxima no higher than 8.

The computer was then asked to plot the two sets of Y values against the set of X values. The resulting graph (Fig. A.1) shows two horizontal lines (“regression lines”), which are so close that they cannot be distinguished by eye. Values from column 1 (open circles) sit close to the lines. Values from column 2 (gray circles) are more widely scattered. The statistical read-out from the computer is shown below the corresponding columns in Table A.1. As expected, the averages (arithmetical means) of the two sets of twenty Y values are both 6. A measure of the degree of scatter about the regression lines is provided by the characteristic known as the standard error. As expected, the values in column 1 are less scattered about the regression line than the values in column 2 (0.133 versus 1.333).

The squares of the correlation coefficient (r^2 ; sometimes called the coefficient of determination) are essentially zero in both cases. This means that no part of the variation between the twenty data values for Y (the “dependent variable”) can be attributed to the changes in the twenty data values for X (the “independent variable”).

Both sets of numbers vary (in column 2 more than in column 1). *But, whatever caused that variation, it was not related to X* (as far as we can tell from the numbers at our disposal).

Being a near horizontal line, the slope value (change in Y value per unit X) is very close to zero ($a = 0.003$ in each case). That this slope value is *not* significantly different from zero is shown by the associated probability values (P), which normally range over a scale from zero to one. The P values are both 0.954 showing that, for the number of data points available, it is *most* probable that there is *no* significant difference from a slope value of zero. By convention, P values must be less than 0.05 to establish likely significance. Finally the values for the intercept on the Y-axis (Y_0 values) are both 6, as expected.

Tilted Lines

It was next intended to tilt the two horizontal lines so creating two new lines differing from the originals in slope and in intercept at the Y-axis (Y_0). Accordingly, values were generated for columns 3 and 4 by adding increasing numbers to the values in columns 1 and 2, respectively (Table A.1). Thus, zero (0) was added to the numbers in the first row (columns 1 and 2), so the numbers both remained at 6 (columns 3 and 4). Then, 1 was added to the numbers in the second row (columns 1 and 2) to generate values of 6.9 and 6.0 for columns 3 and 4, respectively. Then, 2 was added to the numbers in the third row to generate values of 8.1 and 9.0, respectively. Then, 3 was added to the numbers in the fourth row to generate values of 8.8 and 7.0, respectively. This was continued to fill columns 3 and 4, each with twenty numbers.

The resulting computer plots (Fig. A.1) show two new sloping lines (which are again indistinguishable from each other by eye). They retain the same scattering of points as in the horizontal plots from which they were derived. As expected, the mean values for the sets of numbers in columns 3 and 4 are greater than the corresponding sets in columns 1 and 2 (Table A.1b). But, since the scatters about the regression lines are unchanged, the standard errors are unchanged. The standard errors for columns 1 and 3 are identical. The standard errors for columns 2 and 4 are identical.

On the other hand, the r^2 values, which like P values range from zero to one, are 0.999 and 0.951, respectively. Thus, in the case of the plot with less scatter of the points around the regression line (Y values from column 3), much of the variation among these values can be explained on the basis of the X data values. Thus, the r^2 value of 0.999 means that essentially 100% of the variation among the twenty Y values correlates with the underlying variation among the twenty X values. In the case of the plot with more scatter of the points around the regression line (Y values from column 4), much of the variation between these data values can again be explained on the basis of the X data values. However, in this case the r^2 value is 0.951 showing that, while 95% of the variation among the twenty Y values correlates with the underlying variation in the twenty X values, 5% of the variation might relate to another factor, or other factors.

The slope values (both 1.003) are now no longer close to zero. Furthermore, the associated P values (<0.0001) affirm that the slope values are likely to be significantly different from the horizontal (i.e. there is only a very low probability that the slope values are *not* significantly different from zero). The Y_0 values show that both slopes extend back (extrapolate) to cross the Y axis when Y is equal to 5. Thus, all four lines intercept (have a common value) when X is 1 and Y is 6.

Curved Lines

In Figure A.1 the data points are economically fitted to straight lines, as is the case with many graphs in this book. Each slope can then be described mathematically by a single value (often designated as “a” in what is known as a “first order” relationships). However, sometimes the lines curve (e.g. Chapter 14). In these cases, description of the slopes requires further values (“a” and “b”) in what are known as a “second order” relationship). For dissection of more complicated relationships, procedures such as “principle component analysis” (Fig. 4.3) and “best subsets regression” are available in software packages (e.g. Matlab, Minitab). Table 10.1 is an example of the latter.

No Lines

Sometimes data points form a cluster and no linear relationship can be discerned. In this case the points can be represented as a single central point with some measure of their scatter around that point. The central point is determined by the two values for the means of the X -axis and Y -axis values. This point is surrounded by an oval, the limits of which are determined by the scatters of X and Y values along the corresponding axes. For this the corresponding standard errors provide a measure (e.g. Fig. 18.2). There are computer programs that can examine a large cluster of points and determine whether it can be divided into sub-clusters. Thus, although discriminated sets of points were plotted in Figure 18.2a, we could have plotted all the points without discrimination and then could have asked a computer program to attempt to independently cluster them.

A Caution

Huxley in 1864 noted “three classes of witnesses – liars, damned liars, and experts.” While retaining its truth, in modern time this has further mutated to three classes of lies – “lies, damned lies, and statistics.” There are many reasons to be cautious about statistics. My favorite statistician story is that of the statistician who, after war had ravaged Holland, noted a strong correlation between the subsequently rising

birthrate and the return of storks to nest on the housetops. However elegant the mathematics, if applied without an awareness of “the big picture,” statistics can be very misleading.

Of concern to the student of bioinformatics is that sequences are now available in GenBank because, at some point in time, people *decided* to sequence them. Sequencers being of the species *Homo sapiens*, they tended to prefer DNA either from *Homo sapiens*, or from species of economic importance to *Homo sapiens*, which includes species that are pathogenic for *Homo sapiens* and his dependent species (domesticated animals and cultivated plants). Thus, the sequences currently in databases are far from being a random sample. We must keep this in mind when selecting sequences for our studies, and when interpreting results.

Appendix B: Scoring Information Potential

A man who knows the price of everything and the value of nothing.

Oscar Wilde, Lady Windermere's Fan

Information scientists measure information as binary digits or “bits” (Chapter 2). The information value of each unit in a sequence segment depends on the number of types of possible unit that are present in the segment (e.g. base composition), not on the way the units are arranged (e.g. base order). According to their compositions, different segments can be scored for their respective potentials to carry order-dependent “primary” information. The same segments retain their potentials to carry composition-dependent “secondary” information. In spoken information the text can be regarded as primary, and the accent or dialect can be regarded as secondary (Chapter 3).

Two Units

If, in a hypothetical nucleic acid sequence that can contain only two types of base, the bases are distributed randomly but with equal probability (i.e. at a particular position the uncertainty as to which base is coming next is maximum), then each base adds 1 bit of potential base order-dependent information. Thus, $2 = 2^1$. The left 2 refers to the number of bases. The right 2 refers to the type of decisions (binary yes/no decisions; Chapter 2). The 1 bit on the right side of the equation can be referred to either as the “exponent” of the 2 on the right side of the equation, or as the “log,” or “logarithm,” of the 2 on the left side of the equation.

If only one of the two available bases is present in a segment (“window” in the sequence) there is no uncertainty as to which base is coming next, and each base adds 0 bits of potential base order-dependent information ($1 = 2^0$). So variation along a nucleic acid sequence, which contains only one or two types of base, can be

assessed in terms of its local content of potential base order-dependent information in sequence windows on a log scale scoring from 0 to 1 for each base (i.e. the number of bits in a window is equal to the logarithm to the base 2 of the number of available base types in that window). A ten base window consisting of only one base (e.g. **AAAAAAAAAA**) would score zero (10×0). A ten base window consisting of equal quantities of the two bases (e.g. **ATTATAATAT**) would score 10 (10×1). This score would not change if the order of the bases was varied.

Four Units

Along similar lines, if the four bases of a natural nucleic acid sequence are distributed randomly, but with equal probability, along a sequence (i.e. at a particular position the uncertainty as to which of the four bases is coming next is maximum), then each base adds 2 binary digits (bits) of potential base order-dependent information ($4 = 2 \times 2 = 2^2$). If only one base is present in a sequence segment, there is no uncertainty and each base adds 0 bits of potential base order-dependent information ($1 = 2^0$). Again, the left side of the equation refers to the number of bases, and the exponent of 2 on the right side refers to the number of binary (yes-no) decisions required to distinguish between those bases at a single position in the sequence segment.

Here, potential base order-dependent information content (the exponent) is on a scale from 0 to 2. A nucleic acid sequence containing one to four bases can be assessed in terms of its local content of potential base order-dependent information on a log scale scoring from 0 to 2 for each base (i.e. again, the number of bits is equal to the logarithm to the base 2 of the number of bases).

Twenty Units

A protein, with a structure even more elaborate than that of a nucleic acid, is hardly susceptible to analysis in these terms. Nevertheless, similar reasoning can be applied to protein sequences with one to twenty amino acids. Here the scale is from 0 ($1 = 2^0$) to 4.322 ($20 = 2^{4.322}$). Following the binary decision-making principle, one begins by dividing the pool of twenty amino acids into two *equal* groups. A hypothetical informed respondent, on being asked whether a certain pool contains the next amino acid, replies either positively or negatively, and one has reduced the number of possibilities to ten.

This ten is then divided into two equal groups of five, for a second round of interrogation. Third, fourth and, sometimes, fifth, interrogations follow. Since five amino acids do not divide into two equal groups, the number of interrogations is not unitary, but has to occur, on average, 4.322 times (i.e. on average, 4.322 binary (yes-no) decisions are needed to specify which of the twenty amino acids occupies the

next position in a protein sequence). The number of bits is equal to the logarithm to the base 2 of the number of types of amino acid in a window.

Meaning

The above calculations display a *quantitative* aspect of information. Given the number of each type of base in a DNA segment (its base *composition*) we can determine the segment's potential to carry different base order-dependent messages. Its potential is low if the segment contains only one of the four bases. Its potential is maximized if the segment contains equal proportions of the four bases (Chapter 14). However, of all the potential base order-dependent messages, only those of a small subset are likely to convey meaning (i.e. they convey information that may or may not be helpful to the receiver). The existence of a message does not imply that it has a meaning. We need to distinguish sense, which informs the receiver (helpfully or unhelpfully), from non-sense, which does not inform the receiver. Meaning requires the existence of a *code* through which some messages can convey meaning. The same code may be applied to other messages, but a meaning may not emerge. Sometimes meaning requires recognition of base composition alone (i.e. the code maps different meanings to different base compositions). However, often critical to meaning is base *order*. This is a *qualitative* aspect of potential information.

Take, for example, various three-unit message sequences composed of two unit types – say 9 and 1. At each position in each sequence there is either a 9 or a 1, so there are 3 binary decisions to be made. There are $2 \times 2 \times 2 = 2^3 = 8$ alternatives (i.e. the length of the sequence is equal to the logarithm of the number of possible messages). These eight messages are 999, 991, 919, 199, 911, 191, 119, and 111. The first and last contain only one unit type. Of the remaining six (underlined), three have two nines and a one, and three have one nine and two ones. Thus, there are four composition-dependent messages – three nines (999), two nines and a one (991, 919, 199), one nine and two ones (911, 191, 119), and three ones (111).

Which of the eight messages has order-dependent meaning depends on the existence of appropriate codes. Perhaps each message has a distinct meaning. Perhaps none has. 911 is a widely recognized 'universal' alarm call. For all I know, 191 may be the 'pin number' you enter when using your credit card. Thus, it is possible to distinguish *general* codes that are not individual-specific (i.e. they are observer-independent), and *local* codes that are individual specific (i.e. they are observer-dependent).

However, if you passed on your pin number to your children, and they passed on the same pin number to theirs, etc., then your pin number could become more general. All organisms in our planetary biosphere are considered to have evolved from a common ancestor. This seems to predict that all biological codes should be general, as is the case with the codes we have discovered so far (with a few minor wrinkles). So coding studies carried out on the bacterium *E. coli* are helpful in understanding coding in *Homo sapiens*.

A DNA sequence may be natural or artificial. A natural DNA sequence has evolved and so has distinguished itself from the large theoretical subset of DNA sequences of the same length and base composition. The entire sequence is likely to have base order-dependent biological meaning if the information channel through which it has ‘flowed’ through the generations has a limited carrying capacity. If carrying capacity is largely unlimited then the unlikely notion that genomes can carry many long meaningless messages for more than a few generations (Chapter 15) remains on the table.

Imagine that carrying capacity in an information channel is so limited that only three units out of a set of nines and ones can be accommodated. So we can send a three-unit alarm signal (911), which can be interpreted as “primary” order-dependent information and the appropriate code applied. The same three units can also be interpreted as “secondary” composition-dependent information and an appropriate code applied. In this case, 911 is one of the set with one nine and two ones (911, 191, 119). Each of these would have the capacity to convey the same composition-dependent message (whatever that might be).

Thus, if appropriate codes exist, 911 can simultaneously convey *two* messages. As the physicists would say, the two messages are “superimposed.” The order-dependent code (for “primary information”) allows the recipient to distinguish between a total of eight possible meanings. The composition-dependent code (for “secondary information”) allows the recipient to distinguish between a total of four possible meanings. The recipient would have to know from the context whether the encoder intended the message to be read as primary or secondary information, or both. In other words, unless both are intended, an act of cognition destroys the superimposition.

However, given that only three units can be accommodated, what if the encoder wanted to send both an alarm signal (911) and a composition-dependent signal that to encode required three ones (111)? There would then be a *conflict*. The encoder could either send 911 or 111, but not both. Furthermore, 911 would give the wrong information if decoded as secondary information. 111 would give the wrong information if decoded as primary information. The encoder would have to weigh the relative importance of the two messages and decide which to send, and in which context. This, in a nutshell, is the topic of this book.

Appendix C: Selection Level for PR2?

If you don't make mistakes, you're not working on hard enough problems. And that's a big mistake.

F. Wilczek. Longing For The Harmonies (1987)

PR1, in the hands of Watson and Crick, being so informative (Chapter 2), its life as 'a problem' was short-lived. With the resulting genic euphoria, it is understandable that PR2 (Chapter 4) disappeared from view, only being named as such in the 1990s. However, a more likely explanation is that PR2 was (and remains) a hard problem. Future historians will decide whether ignoring it has been a mistake.

PR2 is evident for all k-mers that are likely to exist in a genome of given size (Chapter 4). We can ask whether PR2 is a manifestation of a selective force that acted at a particular k-mer level. Did it act primarily at the single base level (i.e. did it act at the 1-mer – base composition – level) so that higher order k-mers were a consequence, or vice-versa? A pen-and-paper 'thought experiment' might guide us.

For simplicity, take just two bases, **A** and **T**, which pair with each other. Write out a sequence (a) of **As** and **Ts** (Table C.1). Although this sequence is for convenience written as a linear sequence, imagine it is circular with the last base (**T**) being connected to the first base (**A**). Does it obey PR2? Yes, there are 8 **As** and 8 **Ts**. Does the rule extend to complementary dinucleotides? Remembering that the sequence is circular, and that base pairs can overlap, score by writing dinucleotides below the sequence and then counting (b). Note that **AA** = 7 and its complement **TT** = 7; **AT** = 1 and its complement **AT** = 1; **TA** = 1 and its complement **TA** = 1.

The similar decomposition into trinucleotides (c) shows **AAA** = 6 and its complement **TTT** = 6; **AAT** = 1 and its complement **ATT** = 1; **TTA** = 1 and its complement **TAA** = 1. Note that the 16 base sequence could be repeatedly copied and the resulting copies joined end-to-end (concatenated) to produce one large circular single strand.

Whatever the extent of this replication and concatenation, the parity relationship would be retained at the levels both of single bases and of the above oligonucleotides. Thus, if this were the concatenated sequence of some organism, it would not be pos-

Table C.1 Demonstration that, if Chargaff's PR2 applies for complementary mononucleotides (single bases), it does not always extend to complementary dinucleotides, and higher order complementary oligonucleotides, when arranged in a circular single-strand sequence

(a)	A A A A A A A A T T T T T T T T
(b)	A A A A A A T T T T T A A A A A A T T T T T A A A A T T T T T T
(c)	A A A A A T T T T A A A A T T T T T A A A T T T T T A A A A T T T T A A A A A T T T A A A T T T
(d)	T C A C T A G G G A T A C A T T
(e)	T C C T G G A T C A T T C A T A G G T A A T A C A G G A A C T T
(f)	T C A G G G C A T C A C G G A A T T A C T G A T T T T C T A A T A T T C T A G T A C A G G A C A

sible to determine whether 'Nature' had first 'written' the sequence at the single base level, with parity at the oligonucleotide level being an automatic consequence, or the converse.

Sequence (d), however, also obeys PR2 ($A = 5$, $T = 5$, $G = 3$, $C = 3$). Here complementary dinucleotides, trinucleotides, etc., are not necessarily present in equal frequencies. For example, there is no reverse complement (**GT**) for the two **CAs**. The trinucleotide **TTT** is not matched by an **AAA**. The trinucleotide **GGG** is not matched by a **CCC**. Again, the 16 base sequence could be repeatedly copied and the replicates concatenated to produce one large circular single strand. Disparities at the oligonucleotide level would still be present. Parity at the level of single bases does not necessarily imply parity at the oligonucleotide level. Since biological sequences generally show parity at the oligonucleotide level, this suggests that they were initially 'written' at that level.

Appendix D: K-mer Probabilities and Genome Size

Whatever he [the natural scientist] does, there remains much darkness between the few points of light. Whether he emphasizes the light or dwells on the obscurities will depend upon his temperament, but even more upon the temper of the times and upon a form of ever-changing vogue or fashion which acts as a censor forbidding him to be ahead by more than one or two steps.

A Grammar of Biology, Erwin Chargaff (1977) [1]

People are often asking me what's going to happen next in science that's important, and of course, the whole point is that if it's important, it's something we didn't expect. All the really important things come as a big surprise. ... dark energy being the latest example.

Freeman Dyson (2015) [2]

Whatever the temper of the times, one way of tackling the apparently obscure is to go back to first principles [3]. Assuming that each of the four DNA bases has an equal chance of being represented at a position in a sequence, a hypothetical genome sequence only 4 bases in length would contain one of the four 1-mers, **A**, **C**, **G** and **T**. In a sequence 16 bases in length, each of these could be accompanied by one of the other bases to produce $4 \times 4 = 16$ 2-mers (see Fig. 4.1). Each of these 2-mers could be accompanied by one of the 1-mer bases to produce $16 \times 4 = 64$ 3-mers in a sequence 64 bases in length (see Fig. 4.2). Continuing this process, we can generate the data shown in Table D.1.

On a chance basis, each 1-mer base could appear only once in the 4^1 row, which allows only 4 bases. Similarly, each of the sixteen possible 2-mers could appear only once in the 4^2 row, which allows only 16 bases. Similarly, each of the sixty four possible 3-mers could appear only once in the 4^3 row, which allows only 64 bases. Thus, while we can bet that a randomly generated genome sequence of 4 bases (the 4^1 row) has a good chance of containing one or all of the 1-mer bases, the chance of it containing a particular 2-mer is low, the chance of it containing a particular 3-mer is even lower, the chance of it containing a particular 4-mer is remote, and the chance of it containing a 5-mer is zero.

Table D.1 Different ways of expressing the approximate number of bases that correspond to the sequences of a bacterial virus (phage), a bacterium such as *E. coli*, and man (*Homo sapiens*; haploid DNA content)

4^1	4		
4^2	16		
4^3	64		
4^4	256		
4^5	1024	10^3	
4^6	4096		Phage
4^7	16384		
4^8	65536		
4^9	262144		
4^{10}	1048576	10^6	
4^{11}	4194304		Bacterium
4^{12}	16777216		
4^{13}	67108864		
4^{14}	268435456		
4^{15}	1073741824	10^9	
4^{16}	4294967296		Man
4^{17}	17179869184		

Likewise, while we can bet that a genome sequence of 16 bases (the 4^2 row) has a good chance of containing each of the possible 2-mers, its capacity for 3-mers is low, and for 4-mers even lower. There is a very remote possibility that it might contain a particular 16-mer sequence, but zero possibility that it might contain a 17-mer.

Following this reasoning, we would not be surprised to find a particular 6-mer in a bacteriophage genome (corresponding to the 4^6 row), but higher order k-mers would become increasingly unlikely as k-mer size increased above 6. Similarly, we would not be surprised to find a particular 11-mer in a bacterial genome (e.g. *E. coli*, corresponding to the 4^{11} row), but higher order k-mers would become increasingly unlikely as k-mer size increased above 11. By the same logic, the human genome (approximating to row 4^{16}), would be increasingly unlikely to accommodate k-mers of orders higher than 16. While any order of k-mer can be biologically significant, multiple copies of a particular 17-mer should be especially so.

As discussed in Chapter 5 (see “PR2 revisited”) even more intriguing is how close the frequency of a higher order k-mer is to the frequency of its *complementary* k-mer. Comparisons of natural chromosomes with their shuffled equivalents (e.g. Figs. 4.2, 4.3) show that natural chromosomes are significantly less likely to display such parity [4, 5]. This partly reflects the fact that the k-mers of a particular set are not all born equal [6]. For example, the deficit of CpG in certain genomes (Chapter 18) implies a low abundance of k-mers containing the CpG dinucleotide.

References

1. Chargaff E (1977) A grammar of biology. In: *Voices in the Labyrinth. Nature, Man and Science*. Seabury Press, New York
2. Dyson F (2015) Interview by T Lin (March 26th) in *Quanta Magazine* (online at <https://www.quantamagazine.org/>)
3. Forsdyke DR, Bell SJ (2004) Purine-loading, stem-loops, and Chargaff's second parity rule: a discussion of the application of elementary principles to early chemical observations. *Applied Bioinformatics* 3:3–8
4. Afreixo V, Rodriques JMOS, Bastos CA (2015) Analysis of single-strand exceptional word symmetry in the human genome: new measures. *Biostatistics* 16:209–221
5. Zhang S-H (2015) Persistence and breakdown of strand symmetry in the human genome. *Journal of Theoretical Biology* 370:202–204
6. Chor B, Horn D, Levy Y, Goldman N, Masingham T (2009) Genomic DNA k-mer spectra: models and modalities. *Genome Biology* 10:R108

Appendix E: No Line?

The ancient wisdoms are modern nonsenses. Live in your own time, use what we know and as you grow up, perhaps the human race will finally grow up with you and put aside childish things.

Salman Rushdie. Step Across the Line (2003) [1]

In the struggle for literary existence this book may not survive to provoke the comments of future historians. In the event that it does, it is probable that, but for this appendix, they would have wryly noted that, while at the turn of the twentieth century religious extremism was the center of attention – the self-immolation of terrorists and the so-called “intelligent design” version of creationism – this book, like many other contemporary evolution texts, disregarded the topic. Indeed, evolutionists (or rather, evolutionists that get published in high-profile journals) tend to draw a politically-correct line between science and religion, arguing for separate, non-overlapping, domains, or “magisteria” [2, 3]. But, as indicated in Part VII, certain approaches used by scientists can be applied to issues deemed religious, while remaining within bounds deemed scientific.

Setting the Stage

Apart from the genetic information stored in our DNA, we also have non-genetic “mental” information (Chapter 1). The latter is stored in some undefined form (Chapter 20), and we may choose to assign it to external earth-bound databases, often in digital form. Human memory being limited and humans being ephemeral, each generation depends on those proceeding to have selected and assigned some of their non-genetic information to such databases, and to have preserved it there in forms that can be accessed. Databases *per se* have not, as far as we know, acquired the ability to manipulate information in the way humans do – a function which we

broadly name “consciousness.” Human consciousness keeps alive (i.e. holds in trust) what would otherwise be dead information. As Butler pointed out, it is like money in your pocket, dead until you decide to spend it. And human consciousness is itself kept alive by passage of DNA from generation to generation. The dependence on previous generations is of much consequence.

It is an inescapable fact that your grandparents appeared on this planet before your parents, and your parents before you. Thus, your ancestors had the opportunity, which they often availed themselves of, to ‘set the stage.’ On arriving you found that much around you, and the accompanying admonitions, made sense. They told you not to touch the candle flame. Ooops! Won’t do that again! And like many of your seven billion fellow planetary residents, you came to ask how we came to be here and, the not necessarily related question, how we should conduct ourselves during a stay that, with the passage of years, can appear increasingly brief.

Having been wisely guided on less substantial issues, many of the seven billion are inclined not to doubt the wisdom of ancestral answers to these two substantial ones. Salman Rushdie – whose rebellion against religious authority (reminiscent of Samuel Butler’s a century earlier) [4], led a head of state with parliamentary support to condemn him to death (issue a fatwa) – put it this way [1]:

You will be told that belief in ‘your’ stories [i.e. your people’s stories] and adherence to the rituals of worship that have grown around them, must become a vital part of your life in this crowded world. They will be called the heart of your culture, even of your individual identity. It is possible that they may at some point come to feel inescapable, not in the way that truth is inescapable, but in the way that a jail is. They may at some point cease to feel like the texts in which human beings have tried to solve a great mystery, and feel, instead, like the pretexts for other properly anointed human beings to order you around. And it’s true that human history is full of the public oppression wrought by the charioteers of the gods. In the opinion of religious people, however, the private comfort that religion brings more than compensates for the evil done in its name.

William Bateson put it no less emphatically in 1889 having proposed marriage to Beatrice within two weeks of their first meeting during a cultural trip to Dresden:

For me and for most other people in this year of grace I believe the practice of religion to be an outward and visible sign of inward and spiritual duplicity. Of course there are a very few men who feel things heavenly as vividly as things earthly, but they are very rare. For me to be married in a church would be acting a lie and though I love the old services as I do, the old buildings, as some of the fairest things left to us in an age of pollution and shoddy[ness], yet my feeling is that it is ours as a trust somehow from our forefathers, in which we have no part lot, and I should feel just as false if I went to Church and took credit for sanctity as I should if passed a false cheque.

Needless to say, the stage had been well set. They were married in a church [5]. And the stage may have been so well set that you also may have been persuaded that the two substantial questions are indeed related – namely, that the origin of planetary life came replete with instructions (“commandments”) as to how that life should be lived.

Rules of Thumb

Science deals with observations such as that the sun rises in the east and sets in the west, and that the surface of the earth appears flat. As has been shown in these pages, through hypothesis, experiment, and fresh observations, scientists arrive at schemes of relationships that correspond to reality (i.e. are ‘true’) to the extent that they increase our understanding of, and hence potentially our command over, our environment. Thus, for a given set of circumstances, it becomes more probable that we will correctly decide how to proceed. However, in addition to this general approach (i.e. *the scientific method*), there are three ‘rules of thumb’ that scientists – indeed, thinking people in general – have found to be valuable, *albeit not infallible*, adjuncts.

The first, attributed to William of Occam (Chapter 3), is that, given a number of alternative hypotheses to explain a set of observations (facts, data, phenomena), one should first go by (i.e. have most faith in) the simplest consistent with the observations, and only turn to the more complex when the simplest has been tested and found wanting [6].

The second is a variant of the first and is sometimes attributed to Albert Einstein. It is that one should indeed prefer the simple over the complex, but not the *too* simple. For example, the hypothesis that everything is a balance between the forces of good and evil can be made to explain anything. The sun rises in the east because the forces of good are supreme, and sets in the west when the forces of evil get the upper hand. You were wide awake at the beginning of a lecture because of the forces of good. You were fast asleep at the end because the forces of evil had prevailed.

Finally, there is what can be called “the feather principle.” Confronted with a number of competing hypotheses (of which one is “none of the above”), be most dubious of the hypothesis that is pressed by those who have much to gain materially should the hypothesis come to be generally accepted. In other words, suspect those who may be seeking, either consciously or unconsciously, to ‘feather their own nests’ by advocating a particular hypothesis. As Butler pointed out [7]: “It is not the rat catcher’s interest to catch all the rats; and, as Handel observed so sensibly, ‘Every professional gentleman must do his best for to live’.”

As discussed in Chapter 1 and the Epilogue, scientists themselves can have vested interests in particular viewpoints that may play the feathering role to the extent that their power and authority in a hierarchical system are sustained. This may sometimes be seen in its negative form – a failure to admit the existence of a particular rival hypothesis.

Even if we accept the assertion that the scientific method *per se* cannot be applied to religious issues [2, 3], nevertheless we can still apply the three rules of thumb. The hypothesis of one, quasi-anthropomorphic, intelligent creator, accords well with Occam’s razor principle. Quite simply, an entity exists that understands, and can ordain, everything. You have a problem? Be assured that things are that way because that entity decided it should be so. Next question please!

Of course, this argument contradicts the second rule of thumb. The argument is just too simple. It explains everything, yet it explains nothing. Furthermore, applying the feather principle, we can note that the material benefits accruing to its advocates have often been substantial. Indeed, for millennia thinking men and women who lack material resources have opted for the cloistered life where, in return for dispensations of what Karl Marx would have called “opium” to the populace, they have been left in peace to pursue whatever they might find agreeable.

It is likely that Mendel took this path. The necessity for a governing Prince to simulate religiosity if he wished to continue to enjoy the benefits of his office was stressed by Machiavelli [8], and modern politicians disregard this at their peril [1]. However, the creationists do quite well in terms of the negative feather principle – their opponents tend not to cite their texts, whereas creationists tend not to reciprocate by omitting citation.

Probabilities

On balance, creationism fails by the rule-of-thumb principle. But what can science offer instead? Can what science offers also resist rule-of-thumb scrutiny? First we should recognize that, as noted by Thomas Huxley [9], it is difficult for a scientist to be a 100% atheist. A scientist, to remain credible as a scientist, must assign a probability to each member of a set of competing hypotheses. A hypothesis deemed to be most unlikely, such as that of the existence of a supreme creator of the universe (if that is what “God” means), can be assigned a very low probability, but it cannot be excluded.

For a scientist, a declaration that one is say, a Darwinist, merely means that, on balance, one adopts (i.e. has most faith in) Darwin’s view of the power of natural selection. Similarly, one may declare oneself to be an atheist, meaning that, on balance, one prefers (i.e. one believes in) this hypothesis over agnostic and religious alternatives. This means that scientists go into combat with one hand tied behind their backs when arguing the merits of a case with a non-scientist who is not so encumbered. In Rushdie’s words [1]:

One of the beauties of learning is that it admits its provisionality, its imperfections. This scholarly scrupulousness, this willingness to admit that even the most well supported of theories is still a theory, is now being exploited by the unscrupulous. But that we do not know everything does not mean that we know nothing.

Given all this, can a scientist speak objectively about religion in general, and about creationism in particular? By now, scientists have a good track-record of providing examples where apparently objective human perceptions have been shown false. Early man perhaps considered the earth to be flat. Scientific observation of astronomical objects, dating back to the Greeks [9], led to the realization that we – now seven billion of us – are standing on the surface of an approximately spherical earth. The hypothesis was proven by circumnavigation of the planet by early explorers. So, unless there is something we have overlooked, the surface of the earth taken

as a whole (i.e. disregarding small segments) is not flat. More precisely, the probability that the planet's surface is flat may be considered very low. Perhaps, likewise, scientists can show that some religious postulates may be insecure.

No Beginning?

A fundamental observation that underlies much creationist thinking is that everything around us appears to have a beginning and an end. To create, after all, means making something begin. If there is no beginning there can be no creationism. I have sketched out here and in my previous books [5, 11] – with a degree of detail that I believe even the most recalcitrant of bickering evolutionists will be able to live with – a likely path from complex organic molecules to complex living organisms. Others have made the case for the derivation of complex organic molecules from inorganic precursor molecules. Let's examine this further.

The building up of inorganic molecules from even simpler units has been, and remains, an area of intense study based on the premise that there is some fundamental entity ("strings" or "superstrings" that can vibrate in many dimensions) from which everything is constituted [12]. The important point is that there is a unified chain of relationships down from highly complex organisms to a fundamental substratum which, quite simply, *exists*.

Thus, as a point of departure, most scientists accept the concept that there is something fundamental, of a relatively *low order of complexity*, which exists through all time. Their concern is to find what that fundamental is. They feel no need to invoke an abstract creator who might have brought that fundamental into existence. In short, scientists can accept something as *being*. Accordingly, there is no beginning or, necessarily, end, for that something. Thus, it is a law of thermodynamics that energy can be neither created nor destroyed. Energy, whatever it is, just *is*. Or, if it is destroyed, it converts into matter, whatever that is, according to Einstein's famous equation. Energy and matter interconvertably *exist*.

So scientists can accept the concept of being. They accept the concept of existing through time. But so can creationists. In their case the entity that is constant through time is the creator who, by "intelligent design," has put it all together. Whereas the constant entity for the scientists is relatively simple, the constant entity for the creationists would appear to be, given its key attribute as creator, *highly complex*. The two curves (A and B) shown in Figure E.1 indicate the roles of this complex entity, that some deem 'supernatural,' first to create the fundamental units out of which everything derives, and then, *perhaps*, to prescribe (influence) the path this derivation will take. The fact that the latter curve (B) is not obligatory for the supernatural concept is overlooked by some scientists who hold that our ability to explain the path without resorting to a supernatural power destroys the concept [13, 14].

One scientist has embraced both simple and complex constant entities. Disregarding the writings of mathematicians and physicists on what may be the true nature of time (as, through ignorance, do I), the priest and anthropologist Pierre

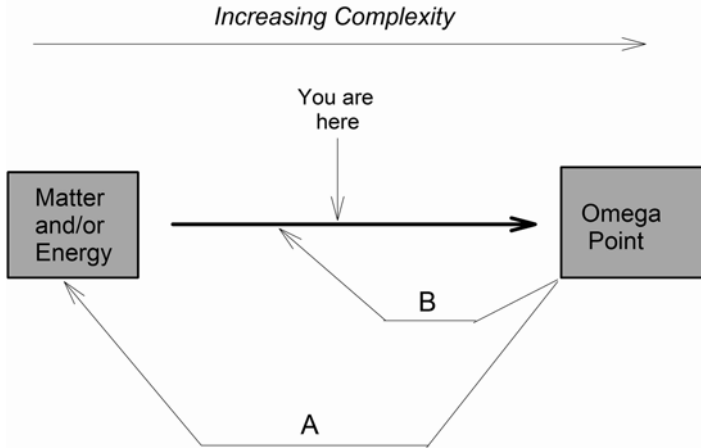


Fig. E.1 Two extremes of the complexity scale (boxed) that exist through time. Intermediate states appear ephemeral. The curved arrows indicate the possibilities of feedbacks either at the beginning (A), or during (B), the process of increasing complexity (the thick horizontal arrow upon which you are located). A minimal hypothesis (matter/energy just exists) would not have these feedbacks. Creating ability being assumed as associated with high complexity, then an ‘active creationist’ hypothesis would have both feedbacks. A ‘lazy creationist’ hypothesis would just have feedback A. The possibility of another thick arrow, separate from that drawn, which allows for an independent evolution of mind from matter/energy, was raised in Part VII

Teilhard de Chardin stood back from the phenomena before him, and saw just one process – matter becoming aware of itself [15]. This required time that would stretch endlessly backwards and forwards from the present, so that what we can conceive as happening, either has already happened (if it is possible to happen at all), or is part of some recursive cycle, or seems intrinsically without end (like the number of decimal places in the ratio of the circumference of a circle to its radius; π).

In Chardin’s view the process is ongoing and tends to produce forms and relationships of increasing complexity that can be extrapolated to an ultimate “omega point” where the degree of complexity would approach a maximum, tending to be constant in time. All this can be summarized as two boxes and an arrow (Fig. E.1). The difference between Chardin’s omega point and the creationists’ supernatural entity, is that the latter is often held capable of (and intent on) influencing the path from simple to complex (curved arrows), whereas the former does not. There is the possibility, doubtless explored in the science fiction literature, of a distant omega point having arisen independently of us, which might choose somehow to influence our evolutionary path, but Chardin does not countenance this.

Chardin had to bide his time before the church would allow his works to be made public. When they appeared they were attacked by the establishment of his day no

less vigorously than those of Butler had been attacked in his time. One lofty observer thought he saw through Chardin's chicanery [16]:

Just as compulsory primary education created a market catered for by cheap dailies and weeklies, so the spread of secondary and latterly of tertiary education has created a large population of people, often with well-developed literary and scholarly tastes, who have been educated far beyond their capacity to undertake analytical thought.

Can Intermediate States Exist?

Under the scheme shown in Figure E.1, we seven billion humans are located somewhere in space and time along the arrow leading from low to high complexity. The following question can then be asked: If the two extremes can, quite simply, *exist*, is it not possible that an intermediate state, such as that in which we currently find ourselves, could also, quite simply, *exist*, or, at least, have the potential to exist? To this, most of the seven billion would answer with a resounding "no." Everything that they see – including themselves, their friends, and their relatives – appears ephemeral, impermanent. Everything has a beginning and an end. However, many presumptively ephemeral, impermanent, things can be rendered less ephemeral, and perhaps permanent, *if we so will it*. This, in itself, might serve to dispel a creationist mythology that is partly driven by fear of death.

For example, for millennia people have agreed to live together on the banks of the River Thames in a geographic region named London. Just as the individual molecules of water in the river are forever changing, so are the people who make up London's population. But London remains. Those who will this far outnumber, and/or so far have defeated, those who might will otherwise. So London has the potential to persist through time to constitute part of that which will exist at Chardin's omega point.

Similarly, through each person – indeed, through each organism on the planet – there is a constant flux of molecules. By recycling, replacement and repair (Chapter 2), today's new molecules substitute for yesterday's. We, as individuals – a collective of molecules – persist. And, hypothetically, there is no limit to the duration of this persistence, so that we also have the potential to constitute part of that which will exist at Chardin's omega point. Thomas Huxley saw this in 1879 [17]:

The animal mechanism is continually renewed and repaired; and though it is true that individual components of the body are constantly dying, yet their places are taken by vigorous successors. A city remains notwithstanding the constant death-rate of its inhabitants; and ... an organism ... is only a corporate unity, made up of innumerable partially independent individualities.

He concluded:

That all living beings sooner or later perish needs no demonstration, but it would be difficult to find satisfactory grounds for the belief that they needs must do so. The analogy of a machine, that sooner or later must be brought to a standstill by the wear and tear of its parts, does not hold.

Cells taken from cats, dogs or humans thrive equally well when cultured away from the body, and can appear potentially immortal. Yet cats and dogs seldom live

beyond twenty years – the time it takes humans to reach their physical prime. Our understanding of the biology of this is increasing (Chapter 17). To some scientists it does not seem unrealistic to suppose that, in centuries hence (some hyperoptimists/hyperpessimists think sooner), humans, perhaps having opted finally to liberate themselves from religious mythology, will choose to ‘permanise’ themselves as entities that, quite simply, *exist*. Quoting from Benjamin Disraeli in *Endymion*, Butler considered it a matter of will [18]:

“There is nothing like will; everybody can do exactly what they like in this world, provided they really like it.” If this is as true as I believe it to be, “the longing after immortality,” though not indeed much of an argument in favour of our being immortal at the present moment, is perfectly sound as a reason for concluding that we shall one day develop immortality, if our desire is deep enough and lasting enough.

Biological knowledge is approaching a state such that all diseases could be either prevented or cured. Considering the process of biological aging as a disease [19], it also might be slowed, or even halted. We can envisage a scenario such that, as perceptual dust and cobwebs clear away, and births and deaths and the associated rituals become infrequent events, people will increasingly come to recognize the existential nature of their being. There will still be churches and priests. But the latter will bring comfort through truth, rather than through mythology. They will describe religions as relics of ancient ways of thinking, celebrate the present, and help people to look afresh at the two fundamental questions. Our evolution as biological entities will slow, giving way to the evolution of the information each generation holds in trust.

Contradiction Shall Reign!

So, where will the priests (‘officiants’) say we came from? It seems that we came from something fundamental, from which inorganic matter is composed, and which, quite simply, *exists*. How shall we conduct ourselves? Again, quite simply, in such ways as should further an agreeable existence. And we will bicker endlessly, without fear of recrimination, on all topics, including that of what constitutes an agreeable existence. In short, we will be free! In Rushdie’s words [1]: “Freedom is that space in which contradiction can reign, it is a never ending debate.” Or, as Samuel Butler put it a century earlier [20]: “Whenever we push truth hard she runs to earth in contradiction in terms, that is to say, in falsehood. An essential contradiction in terms meets us at the end of every enquiry.”

We also will debate possible omega points towards which we might collectively strive, and some may come to agree with Thomas Huxley [21] that:

The purpose of our being in existence, the highest object that human beings can set before themselves, is not the pursuit of some chimera as the annihilation of the unknown; but it is simply the unwearied endeavour to remove its boundaries a little further from our sphere of action.

Butler of course said it more simply, and with his usual twist of humor [20].

If I thought by learning more and more I should ever arrive at the knowledge of absolute truth, I should leave off studying. But I believe I am pretty safe.

In other words, we will strive to further the evolution of the sort of information that will help our understanding.

How does all this accord with the three rules of thumb? Quite well with the first two. In not invoking elusive supernatural powers, one arrives at a scenario of moderate simplicity. And those who advocate this view, including those trained in science, while having much to gain intellectually from its wide acceptance, usually do not gain materially to such an extent as the ardent advocates of religious viewpoints. Thus, we should not be overly suspicious of the motives of those advocating the view (third rule of thumb).

We should neither regard such individuals as shallow, nor impute an absence of sensitivity, or an innate inability to be moved at the wonder of it all. Atheists are no less sympathetic to the needs of others than the religiously inclined; nor are they moved less by great music, great art, the quiet splendor of a sunset, and the silver path to the moon at sea. The lives of Samuel Butler and William Bateson attest to this. The difference is that atheists consider the latter to be works of individual genius, or of natural forces, not of a God working through man or through Nature. When things go wrong they see that “the fault, dear Brutus, lies not in the stars, but in ourselves” [22]. But, as we saw in Part VII, perhaps there is no line between us and “the stars.”

References

1. Rushdie S (2003) *Step Across the Line*. Random House, Toronto
2. Gould SJ (1999) *Rocks of Ages. Science and Religion in the Fullness of Life*. Ballantine, New York
3. Ruse M (2001) *Can a Darwinian be a Christian? The Relationship between Science and Religion*. Cambridge University Press, Cambridge
4. Jones HF (1919) *Samuel Butler. A Memoir*. Macmillan, London
5. Cock AG, Forsdyke DR (2008) “Treasure Your Exceptions.” *The Science and Life of William Bateson*. Springer, New York
6. Birchler JA (2015) Mendel, mechanism, models, marketing and more. *Cell* 163:9–11 [“A simple explanation that does not explain the facts is to be discarded. Sydney Brenner introduced the concept of ‘Occam’s Broom,’ which is used to sweep inconvenient truths under the rug to salvage the ‘simplest’ explanation. Recognizing when to use the razor and avoid the broom is a useful reflection in evaluating scientific models as the subtle Mendel and McClintock examples attest.”]
7. Butler S (1880) *Unconscious Memory*. David Bogue, London.
8. Machiavelli N (1950) *The Prince and the Discourses*. Random House, New York
9. Huxley TH (1948) *Selections from the Essays*. Huxley. Castell A (ed) AHM Publishing, Northbrook
10. Lucretius (1951) *The Nature of the Universe*. Latham RE (trans) Penguin Books, Harmondsworth
11. Forsdyke DR (2001) *The Origin of Species, Revisited. A Victorian who Anticipated Modern Developments in Darwin’s Theory*. McGill-Queen’s University Press, Montreal

12. Wilczek F, Devine B (1987) *Longing for the Harmonies*. Norton, New York
13. Crick F (1988) *What Mad Pursuit*. Basic Books, New York
14. Dawkins R (1986) *The Blind Watchmaker*. Longman, Harlow, p ix
15. Chardin PT de (1959) *The Phenomenon of Man*. Collins, London
16. Medawar PB (1961) Review of 'The Phenomenon of Man.' *Mind* 70:99–106
17. Huxley TH (1879) *The Crayfish. An Introduction to the Study of Zoology*. Kegan, Paul & Trench, London, p 127
18. Butler S (1881) *Alps and Sanctuaries of Piedmont and the Canton Ticino*. David Bogue, London, p. 23 [Disraeli was also Lord Beaconsfield.]
19. Schäfer D (2002) 'That senescence itself is an illness': a transitional medical concept of age and ageing in the eighteenth century. *Medical History* 46:525–548
20. Butler S (1985) *The Notebooks of Samuel Butler*. Jones HR (ed) Hogarth Press, London, pp. 59, 299
21. Huxley TH (1896) *Darwiniana Essays*. Macmillan, London, pp 447–475
22. Shakespeare W (1599) Julius Caesar. In: Rowse AL (ed) *The Annotated Shakespeare*. Orbis, New York (1988)

Acknowledgements

I am indebted to Jim Gerlach and David Siderovski for much patient advice in the 1980s as I struggled to understand computers. My bioinformatic studies were enthusiastically assisted by a host of computer-literate students in the Department of Biochemistry at Queen's University in Canada. Special thanks to Isabelle Barrette, Sheldon Bell, Labonny Biswas, Yiu Cheung Chow, Anthony Cristillo, Kha Dang, Previn Dutt, Scott Heximer, Gregory Hill, Janet Ho, Robin Lambros, Perry Lao, Justin LeBlanc, Shang-Jung Lee, Feng-Hsu Lin, Christopher Madill, James Mortimer, Robert Rasile, Jonathan Rayment, Scott Smith, Theresa St. Denis, and Hui Yi Xue.

Insight into the differing perceptions of the 'info' people and the 'bio' people came in the 1990s when Janice Glasgow of the Department of Computing and Information Science at Queen's University invited me to join her in initiating a graduate course in Bioinformatics, and Paul Young of the Department of Biology invited me to contribute to his undergraduate Genomics course. His text, *Exploring Genomes*, was published in 2002. Andrew Kropinski of the Department of Microbiology and Immunology advised on bacteriophage genomes. Peter Sibbald advised on bits and bytes. David Murray of the Department of Psychology provided valuable biohistorical information and, among many, the works of biohistorian William Provine (1942-2015; Cornell University) were of great help in my getting 'up to speed' in that area.

Special thanks are due to Mary Catherine Bateson, and to Alan G. Cock (1926-2005) and his family, for facilitating access to historical materials relating to William Bateson. Charlotte Forsdyke copy-edited. Polly Forsdyke advised on German usages and translated the French and German versions of Delboeuf's 1877 paper into English. Sara Forsdyke advised on Latin and Greek usages. Ruth Forsdyke believed in me, sometimes more than I did myself, and suggested the term "reprototype." My wife Patricia is a constant source of inspiration and encouragement.

Sheldon Bell, Christopher Madill and James Mortimer wrote programs that greatly facilitated our bioinformatic analyses. Michael Zuker of the Rensselaer Polytechnic Institute kindly made available his nucleic acid folding programs. Dorothy Lang gave me information on mirror repeats prior to publication. The

Genetics Computer Group (GCG) suite of programs (“the Wisconsin package”) acquiesced to modification by my primitive UNIX scripts. In recent years there have been a productive collaboration with Chiyu Zhang (Jiangsu University) and Ji-Fu Wei (Nanjing Medical University), and helpful correspondence with Michel E. B. Yamagishi (Cidade Universitária Zeferino Vaz-Barão Geraldo) and Virgil Reese (Bartlesville, Oklahoma).

For the first edition, Joseph Burns, Deborah Doherty and Marcia Kidston (Springer, Norwall) smoothed the passage from manuscript to final copy. Melanie Wilichinsky, Meredith Clinton and Andrea Macaluso (Springer, New York) did likewise for the second, and Noreen Hensen and Lesley Poliner (Springer, New York) and Joseph Daniel and Sasi Reka (SPi Technologies India Private Ltd) did likewise for the third. Original data for figures were generously provided by Kenneth Wolfe and Paul Sharp (Fig. 7.3), by Honghui Wan and John Wootton of the National Center for Biotechnology Information (Fig. 14.1), and by Daiya Takai and Peter Jones of the University of Southern California (Fig. 18.2). Most other data were publicly available from sources such as GenBank and the ExInt Database (Fig. 13.2). The photograph of Erwin Chargaff was from the collection of the National Library of Medicine, Washington. The photograph of Samuel Butler is from a self-portrait held by St. John’s College, Cambridge, with permission of the Master and Fellows. The photograph of Friedrich Miescher was from the University of Basel. The photograph of Gregory Bateson was from the Imogen Cunningham Trust, California. The photograph of George Romanes was from the Royal Society collection. The photograph of Richard Goldschmidt is from the personal papers that Alan G. Cock entrusted to me in 2004 (now in the Queen’s University Archives). The photograph of William Bateson and Raphael Weldon is also from Alan’s personal copy of the William Bateson papers. Figure 2.1 was kindly adapted by Richard Sinden of Texas A & M University from his book *DNA Structure and Function*. Elsevier Science gave permission to use this, and also Figure 5.6 from an article by Michael Zuker and his colleagues in *Methods in Enzymology*, and Figure 16.3 from Max Lauffer’s *Entropy Driven Processes in Biology*. Figure 2.5 is adapted from an article by Austin and Marianne Hughes with permission of the authors and Nature Publishing Group. Figures 12.2 and 12.3 are reproduced from *PLOS One* with the permission of Donal Hickey (Concordia University, Montreal), and Elizabeth Clare (University of Guelph) provided supporting data. Figure 20.1 is reproduced under Creative Commons Licence from *Frontiers in Human Neurosciences* (Frontiers Media, Lausanne).

Permissions to reproduce or adapt materials from my own works were provided by Springer (Figs. 4.2 to 4.4, 5.5, 5.6, 7.2, 8.2, 10.1, 10.3, 11.4, 12.12, 13.4, 13.5), Cold Spring Harbor Laboratory Press (Fig. 6.2), McGill-Queen’s University Press (Figs. 8.1, 8.3, 8.4, 9.6 and 17.3), Elsevier Science (Tables 10.1, 13.1; Figs. 5.3, 5.4, 5.8, 6.1, 6.7, 13.6, 13.7, 14.2, 14.4 to 14.7, 14.9, 15.1 to 15.3, 15.6, 16.5), Adis International Limited (Table 4.3; Figs. 12.1 and 12.4 to 12.10), Oxford University Press (Fig. 13.3), World Scientific Publishing Company (Figs. 11.3, 16.1 and 16.2), and the Royal Society (Figs. 9.3 to 9.5). For space reasons many important studies are not cited. However, the studies that are cited often contain references to those studies.

My web pages are hosted by Queen's University (for internet locations see p. xxxiii in Prologue, References and Notes). These display full-texts of various scientific papers from the nineteenth century onwards, and much supplementary material, including links to videos of my Evolution Academy offerings (see also at YouTube: <https://www.youtube.com/user/forsdyke1/playlists>). Where possible the webpages will contain regular updates on new work that appeared after the book went to press. To supplement my books, I have written short biographies of W. Bateson, E. Chargaff, J. B. S. Haldane, G. J. Mendel, H. J. Muller and G. J. Romanes. Nature Publishing Group placed these in both on-line and paper editions of the *Nature Encyclopedia of Life Sciences* (2002) and/or the *Nature Encyclopedia of the Human Genome* (2003). In 2005 these became the *Encyclopedia of Life Sciences*, published by John Wiley & Sons, and are available on-line at <http://www.els.net/>. I have also provided an overview of introns in Chapters 4 and 8 of Lewin's *GENES XI* (2014; Jones & Bartlett, Sudbury).

Finally, there is a rather unusual acknowledgement of an anticipatory rather than retrospective nature. First, a quotation from *Unconscious Memory* concerning Samuel Butler's attempts to make the Victorian evolutionists recognize both the ideas of Hering and himself, and their own intellectual debts to Georges Louis Leclerc de Buffon and Charles Darwin's grandfather, Erasmus Darwin:

My own belief is that people paid no attention to what I said, as believing it simply incredible, and that when they come to know that it is true, they will think as I do concerning it. ... My indignation has been mainly roused ... by the wrongs ... inflicted on dead men, on whose behalf I now fight, as I trust that some one – whom I thank in anticipation – may one day fight on mine.

The reader will recognize that, a century after his death, I have here fought on Butler's behalf. I respectfully accept his thanks, while noting his caution *To Critics and Others* that the future might come to "see in me both more and less than I intended". But the battle is not yet won. So, in turn, I thank in anticipation that some one who may, one day, continue our work.

Index

A

Academic

- cloister, 428
- discipline, 58
- field of specialization, xxviii
- haven, 4
- ivory tower, 400
- mule, 59
- sterility, 58
- suicide, 404
- tie with industry, 403
- tomb, 405

Accent

- base composition as, 136, 177
- as class barrier, 47
- is deeper than content, 46
- of genes, xxx, 200
- of genomes, 136, 177
- linguistic divergence by, 136, 177
- as secondary information, 44, 128
- as species barrier, 47

Accidents, 32, 39, 390

Achilles heel

- of AIDS virus, 243
- of humans, 243

AC-pressure, 257, 271

Action

- consciousness of, 7, 317, 388, 425
- instinctual (innate), 8, 17, 335

Adaptation

- conserved by reproductive isolation, 177
- constrained by what exists, 57
- countermending, 49, 284
- of DNA to high temperature, 204
- facilitatory, 284
- for intra-strand pairing, 83

by low complexity segment, 263

menopause as an, 316

simultaneous, 54

treadmill of, 224

uncoupled from speciation, 137, 273

Adaptation and Natural Selection by
Williams, 201

Adenine deamination, 360

Agency

- base composition as, 201
- causing geographical isolation, 169
- as designer, 19
- internal or external, 19
- natural selection as, 51, 169, 201
- not natural selection, xxiii
- physiological selection as, 159, 223
- as preserver, 200
- secondary information as, 52

Aggregation. *See also* Heteroaggregation;

Homoaggregation

- chaperones defend against, 317
- entropy-driven, 307
- specificity of, 306
- temperature increases, 296, 308

Aging

- Butler on, 333, 432
- and cancer, 311
- as a failure to repair DNA, 333
- Huxley on, 431
- Weinberg on, 347

AG-pressure (*see* Purine-loading)

AIDS virus

- Achilles heel of, 243
- copackaging of genomes of, 185
- copackaging signal of, 243
- dimer linkage sequence of, 185, 243

- AIDS virus (*cont.*)
 evolutionary strategy of, 184
 flushing from reservoirs, 243
 genome of, 21, 242
 latency of, 21
 as a low (G+C)% species, 185
 overlapping genes of, 245
 pandemic of, xxv
 positive selection of, 242
 stem-loop potential of, 241
 superinfection of, 185
 therapy of, 184, 243, 403
- Alarm
 amplification of, 279
 by antigen-antibody reaction, 282
 by dsRNA, 292, 295, 312
 intracellular, xxx, 286, 360
 universal, 417
 by virus, 312
- Alexandria, library of, xxiii, 51
- Allele
 biallelic expression of, 314
 conflict, maternal-paternal, 358
 conversion of, 197
 definition of, 143
 one pair insufficient, 337
 and recombination repair, 196
- Allometry, 381
- Allopatry, 151, 158, 169, 398
- Altruism
 Darwin's problem with, 289
 by virus-infected cell, 285
- Amino acid, 269
 in AG-rich gene, 226, 228, 229
 in AT-rich gene, 124, 175, 213
 basic, 124
 as building block, 17, 45, 308
 composition variability, 79, 188
 in GC-rich gene, 124, 175, 213
 hydrophilic, 124
 hydrophobic, 124, 257
 importance of order of, 189, 231
 letter code for, 123
 names of, 45
 not optimum for protein, 95, 175
 one-step mutation subsets, 124
 as placeholder, xiii, 223, 230, 249, 258, 262, 263
 repeats (*see* Repeats)
 substitutions, 131
- Amoeba. *See also* Protozoa
 contrasted with a ciliate, 284
 living molecules in, 9
 organelles of, 19
 sophistication of, 19, 284
- An Examination of Weismannism* by Romanes, 14
- Antibody
 class switching of, 273
 as cross-linking agent, 307
 diversification of, 273
 holes in repertoire of, 282
 intracellular repertoire of, 286, 287
 repertoire generation, 355, 360
 repertoire is host-specific, 283, 315
 repertoire molded, 282
 repertoire purged, 282, 283
 six steps in production of, 282
 as steric complement, 298
 strength of antigen-binding by, 283
 variable and constant regions, 286
 variable genes for, 283
 Y-shaped, 286, 307
- Antigen
 cryptic, 298
 intracellular, 286
 as a moving target, 184, 313
 as a not-self marker, 282
 at pathogen surface, 127, 266, 282
 paternal in embryo, 281, 283
 as a self marker, 292
- Antigenic determinant
 introduction of the term, 308
 as a not-self marker, 282
- Anti-recombination
 activity not localized, 337
 as key to speciation, 336
 keeps DNA intact, 201
 and multiple alleles, 338
- Apoptosis
 in multicellular organisms, 287, 316
 in protozoa, 285
 role of mitochondria in, 116
 triggering of, 317
- Archaea, 106, 205, 235, 256
- Aristotle, 5, 43
- Arms race
 between predator and prey, 239
 and complexity, 317
 counter adaptations in, 280
 with forgers, 40
 intracellular, 287, 299, 353
 limits exon structure, 241
 not for speciation, 136
- Arrangement
 of body parts, 121
 of chromosome constituents, 153
 of elements, 5
 of facts, 21
 linear, of pangens, 14

of molecules, 14
 phenomenon of, 15, 371
 as source of meaning, 154
A Short History of Genetics by Dunn,
 xxxiii, xxxiv
The Astonishing Hypothesis by
 Crick, 370
 Audition
 by oocytes, 346
 by T-lymphocytes, 292, 315, 346
 Autoimmune disease
 of brain, 314
 in Down syndrome, 345
 due to AIRE deficiency, 292, 315
 female susceptibility to, 344, 345
 microbial equivalent of, 354
 when self read as not-self, 283, 287
 Avery, O., 154
 Avicenna, xxxi, 376

B
 Bacteria. *See also Escherichia coli*
 codon position plots for, 213, 225
 CRISPR defense system in, 299
 detect magnetic field, 373
 memory in, 367
 Bacteriophage
 defense strategy of, 353
 detected as not-self, 299, 353
 DNA length (bases), 422
 genome replication, 109
 microisochores in, 198
 paradox of, 85
 Baisnée, P.-F., 78
 Ball, A., 86
 Barrier, 149, 151, 281
 to demarcate information, xxix
 to development (*see* Hybrid inviability)
 to development falls, 150
 to gametogenesis (*see* Hybrid sterility)
 gene subsets determining, 148, 158
 of hybrid inviability (*see* Hybrid inviability)
 of hybrid sterility (*see* Hybrid sterility)
 linguistic, 46
 need to breach, 49, 58
 to pathogen, 285
 placental, 281
 prezygotic, 148
 to recombination, occurs first, 200
 reproductive, 46, 164
 retained in retrovirus, 184
 successive, 46, 151, 184, 193, 201
 temporal, 158
 three fundamental isolating, 147, 169

 to transmission falls, 149, 150
 to transmission trumps others, 148, 149, 158
 Base, 86, 103, 127. *See* Pairing
 cluster of (*see* Base cluster)
 combinations of two from four, 257
 complementary duplets, 67
 complementary triplets, 67
 composition (*see* Base composition)
 conserved (evolving slowly), 127
 deaminated, 331, 355, 360
 departure from equipfrequency, 257
 methylated, 84, 331, 351
 mispairing, 328
 non-conserved, 127
 non-Watson-Crick pairing of, 328
 order of (*see* Base order; Pairing)
 stacking of, 29, 31, 92, 96, 126, 296
 substitutions, 125, 240, 243, 328
 symbols for, 28
 trade-offs between, 227, 229, 230
 unusual in DNA, 328, 331, 351
 unusual in tRNA, 84
 Base cluster
 as cluster of clusters, 104
 of consecutive Gs or Ts, 227, 297
 level of selection for, 112
 and purine-loading, 105
 and transcription direction, 104
 violation of PR2 by, 104
 Base composition
 adoption by transposed gene, 110,
 195, 197
 affects amino acid composition, 124, 175,
 223, 263
 average for a species, 213
 as basis of reaction pattern, 341
 correlation with recombination, 138
 dependent component of CG
 frequency, 357
 dependent component of stem-loop
 potential, 88, 90, 183
 dependent message, 415
 determines k-mer frequency, 73
 determines speciation, 169
 differences affect stem-loop
 extrusion, 181
 and growth temperature, 231
 mathematical regularities in, 80
 in phylogenetic analysis, 176, 184
 predicts folding energy, 183
 pressure by two bases, 257
 profile for a sequence, 105
 same in both strands, 65–66
 unidentified properties of, 76
 variability of, 188

- Base order
 dependent component of **CG** frequency, 357
 dependent component of stem-loop potential, 88, 241, 267
 dependent message, 415, 417
 determines k-mer frequency, 73
 opposing base composition, 91, 97
- Bat, 37, 284, 368
- Bateson, G.
 influenced by father, 397
 on information conflicts, 209
 on information levels, 54
 photo of, 55
 on religion, 370
- Bateson, P., 399
- Bateson, W.
 agreement with Romanes, 157
 on Butler, 7, 14, 141
 character of the man, 402
 character units of, 15, 51, 163
 on complementary factors, 164
 detractors of, 397
 disagreement with Darwin, 7, 396
 on discreteness of species, 52
 on enzymes, 14
 on genetic factors, 15
 on mutant sectors in gonad, 347
 photo of, 160
 as pioneer, xxiv, 4, 152, 396, 404
 on religion, 426
 on residue concept, 163, 168, 176, 396
 and rogue phenomenon, 359
 on sex and speciation, 339
 sponsor of, xxiv, xxxii
 on variation, 121
 and wax model, 14
- Bdelloid rotifers, 341
- Beau Geste* by Wren, 404
- Bechamp, A., 305
- Behaviour
 courtship, 335, 373
 prediction of, 388
- Berkovich, S., 374, 387
- Bernabei, R., 375
- Bernardi, G.
 and classical isochores, 193, 199
 and genome phenotype, 130
- Bias, 213, 288
 of codon (*see* Codon bias)
 in mutation (*see* Mutation bias)
 in research support, 4, 157, 320, 403, 407
 of unbalanced dice, 18
- Bibb, M., 198
- Binary digit, 36, 254, 371, 415
- Binomial distribution, 80, 145
- The Birds* by du Maurier, 49
- Biohistorians
 misinterpretations by, xxviii, 396
- Bioinformatics. *See also* EB
 of brain function, 367, 375
 gene-centered, 3, 157
 history of, xxviii, 157
 the new, xv, xxiv, xxvii, xxviii, 3, 5, 254
 for phylogenetics, 57, 255
- Bird. *See also* Chicken
 adaptation for flight, 284, 368
 brain size of, 38, 49
 detects magnetic field, 373
 DNA content of, 37, 288
 migration of, 8, 37, 38
 mitochondria of, 217
 nest-making by, 8, 9
 pigeon hybrids, 162
 red blood cells of, 42, 321
- Blending
 of accents, 47
 of characters, 54
 of genes, 177, 179, 200
 of genomes, 185, 200
 inheritance, 142, 159, 179
 inheritance of multigenic traits, 145, 340
 not with sex characters, 143, 337, 340
- Bök, C., 65
- Brain. *See also* Memory
 ablation of parts of, 370, 382
 alteration in structure of, 17
 and consciousness, xxxii, 426
 and dreams, 18, 390, 391
 hippocampus, 357
 hypotheses of information storage in, 386
 as immunologically privileged, 271, 369
 laterality of, 376, 391
 plasticity of, 386
 rapid methylation in, 378
 scans, 383
 size in birds, 38
 size in man, 381
 size in women, 388
 size not correlated with memory, 373, 374, 376
 as stand-alone entity, 373
 synaptic plasticity hypothesis, 370
 ventricles of, 383
- Brain disease
 Alzheimer's, 271, 345
 autoimmune, 314
 blindsight, 391
 depression, 391

- Down syndrome, 345
- Huntington's, 269
- hydrocephaly, 383
- microcephaly, 385
- multiple personality syndrome, 387, 391
- and prions, 271
- and protein aggregation, 271
- prions, xxiv
- psychosomatic, 288
- savant syndrome, 385
- schizophrenia, xxiv, 54, 387, 403
- seasonal influences on, 391
- spinocerebellar ataxia, 271
- and trinucleotide expansion, 271
- Branch points. *See* Phylogenetic analysis
- Brave New World* by Huxley, 403
- Brenner, S., xxiv, xxv, xxxiii, 190
- Bridges, C., 342
- Brücke, E., 284
- Buffon, G.L.L., 402, 437
- Bultrini, E., 76
- Butler, S.
 - on aging, 333, 432
 - as challenger of authority, 7, 21, 427
 - on concealing meaning, 401
 - on egg and hen, 7
 - on evolution of machines, 18, 406
 - as friend of Theophrastus, 406
 - his frustration by Grant Allen, 407
 - his frustration by Romanes, 8, 400
 - on heredity and memory as one, 7–9, 368
 - on history, xxvii
 - on hybrid inviability, 58, 280
 - on hybrid sterility, 59
 - known of by Eliot, 399
 - Lamarckian viewpoint of, 19
 - and language, xxvi, 16
 - on life, 18, 432
 - on memory running a gemmule, 17
 - on money, 426
 - and natural penetration, 14
 - on natural selection, 141
 - his need to self-publish, 404
 - on non-blending of sex characters, 143
 - not understood, 20, 400
 - portrait of, 7
 - as regarded by Bateson, 7, 14
 - his revulsion at natural selection, 19
 - on sexual reproduction, 48
 - as shifter of perspective, 6, 392
 - on species limits, 49
 - on variation as origin of species, 18, 141
 - on words as mere means, 375
 - water-dipper analogy of, 360
- C**
- Cancelling
 - of epigenetic states, 359
 - of memories, 11
- Cancer. *See also* Paraneoplastic disease
 - cells display self, 314
 - cure of, xxiv, 314, 398, 403
 - individualized response to, 313
 - as lost population homeostasis, 316
 - oncogene, 243, 311
 - oncoprotein, 94, 311
 - in post-reproductive life, 311
 - suppressor gene, 355
- Cat, 13, 25, 64
- Cell, 312
 - aggregation of, 306
 - cycle of mitotic division of, 331
 - cytotoxic (*see* Lymphocyte)
 - death at home or in exile, 315
 - death by apoptosis, 287, 316
 - death by necrosis, 316
 - detects subatomic particles, 372
 - division as allotting characters, 15, 328
 - division more in testes, 346
 - DNA content of, 368
 - as an elementary organism, 284
 - as an immune system, 286
 - medullary epithelial, 315, 346
 - mitotic division of, 284, 331
 - number in brain, 372
 - phagocytic, 316
 - population size homeostasis, 316
 - red blood (erythrocyte), 306
 - regeneration from stem, 271, 312
- Cerebral
 - alterations, 17, 20
 - event, 290
 - wiring, 17
- Certainty
 - of prediction, 388, 415
 - and uncertainty, xxiii, 254, 381
- CG** island (*see* CpG island)
- Chambers, R., 387
- Channel
 - capacity, xxx, 288, 418
 - metabolic pathway as, 33
- Character, 21
 - antigenic, 280
 - assortment among children, 161
 - blending of, 49, 340
 - constitutive, 21
 - dilution of a new, 54
 - dominant and recessive, 142
 - inducible, 21

Character (*cont.*)

- latent (*see* Latent)
- not transferable by breeding, 163
- recognition before classification, 279
- sexual, 21
- transferable by breeding, 163, 168
- unit as defined by Bateson, 15, 163

Chardin, P.T., 430

Chargaff, E.

- difference (skew), 79, 105
- difference (skew, cumulative), 109
- difficulties of, 404, 406
- as discoverer, xxviii
- his cluster rule, xxix, 116
- his GC-rule, xxix
- his PR1, xxviii, 27
- his PR2, xxix, 65, 77, 183, 335
- on preservation of information, 138
- PR2 compliancy in introns, 76
- PR2, precision of, 79, 96
- PR2, selection level for, 419
- PR2, violation of, 104, 110, 115, 183, 253, 271
- rules as correction guides, 335
- rules summarized, 249

Chen, J-H., 183

Chicken. *See also* Bird

- color of feathers, 15
- compact genome of, 37
- as an egg's way, 7, 10
- inborn capacity of, 8
- intron size differences, 38

Chromosome, 187. *See also* X-chromosome;

Y-chromosome

- 04 of humans, 74
- 19 of humans, 72, 74
- 21 of humans, 345
- accounting between, 79
- accounting within, 79
- assortment at meiosis, 162, 182, 329
- centromere, 98, 273
- chromatin configuration of, 360
- conjugation, 166, 179, 281
- diploidy of, 35
- dosage compensation by, 343
- ends (telomeres), 98
- extra copies of, 345
- genes located to, 161, 163
- haploidy of, 35
- incompatibility at meiosis, 153, 162, 168
- late replication of, 332
- linear pattern of, 154
- mechanism of inactivation, 344
- meiotic division of, 167, 281

- mitotic division of, 160, 167
- number of, 169
- number of sterility factors on, 341
- orientation, 79
- as reaction system (*see* Reaction system)
- repatting of, 56, 153, 340
- telomere strand asymmetry in, 272
- thread-like, 331
- transposition of segments, 169, 244

Class

- of adverse event, 39
- barriers, 47
- as blending or non-blending, 145
- as dangerous or non-dangerous, 279
- as friend or foe, 279
- as harmful or non-harmful, 283
- of genotypic change, 55
- of immunoglobulin, 273
- of macromolecule, 16
- of transcript, 298, 359
- that is unclassified, 397

Clausen, R., 186

Clifford, W.C., 388

Cluster, 112

- of bases (*see* Base cluster)
- of birth dates, 391
- of points, 409, 413
- rule (*see* Chargaff, E.)

Code

- active and passive, 210
- as agreed convention, 16, 43, 122
- for base composition, 417
- for base order, 417
- degeneracy of, 46, 123
- dimensions of, 43
- general or individual-specific, 417
- the genetic, 123, 404
- linking phenotype to genotype, 15
- needed for meaning, 43, 417
- plasticity of, 130
- universality of, 122

Codon

- bias as a dialect, 211
- bias as non-random, 86, 210
- bias as secondary, 213
- bias at functionally minor sites, 195
- bias not translation-driven, 195
- bias, species specificity of, 211
- coincident, 134
- cooperation between positions of, 222
- as five base unit, 126
- in mRNA, 84, 126
- position plot for a gene in many species, 216
- position plot for genes within a species, 217

- position plot for species, 214, 216
- position plot switch-over point, 219
- position plots, context-independent, 218
- reading frames for, 126
- for starting protein synthesis, 123
- for stopping protein synthesis, 123, 212
- synonymous sets of, 46, 130, 212
- third position, independence of, 123, 213, 218, 256, 261, 263
- as three base unit, 45, 126
- tissue-specific usage of, 200
- tracks base composition, 188
- translation rate and accuracy, 107, 129, 187, 213
- wobble, 205
- Coevolution
 - of prey and predator, xxx, 239, 320
 - of proteins in common cytosol, 311
- Complementarity
 - of bases, 28, 168
 - between antigen and antibody, 298
 - of codon and anticodon, 84, 126
 - key and lock, 28, 43, 168
 - between self and not-self RNA, 294
 - between sterility factors, 165
 - of sword and scabbard, 168
 - of templates, 29
 - testing for, 180
- Complexity. *See also* Low complexity
 - and collective psychology (*zeitgeist*), 389
 - and compressibility, 254
 - as divergence index, 139
 - and entropy, 254
 - of explanation, 427
 - measurement of, 253
 - of psychomotor activity, 8, 374
 - related to uncertainty, 415
 - scale of, 431
 - and the social organism, 388
 - and the supernatural, 429
 - and uncertainty, 254
- Conflict. *See also* Information conflict
 - between ancestral memories, 59
 - between certainty and uncertainty, 254
 - when context unclear, 418
 - between genomes, xxx, 358
 - within genomes, xxx, 335
 - between good and evil, 427
 - between hypotheses, 396
 - between intrinsic and extrinsic, 64
 - between mRNA structure and coding, 94–96
 - between nucleus and cytoplasm, 95, 245
 - between objectivity and subjectivity, 388
 - between protein and DNA, 239
 - between sympatry and allopatry, 398
 - resolution of, 22, 85, 237, 401
 - between RNA and DNA, 245
- Consciousness
 - of action, 7, 316
 - higher than individual, 389
 - of observer, 369
 - and subjectivity, 389
- Conservation. *See also* Preservation
 - of exon more than intron, 37
 - of intron more than exon, 241
 - low in unclassified transcripts, 298
 - not always function indicator, 289, 298
 - of phenotype and genotype correlated, 134
 - as phylogenetic premise, 57
 - of stem-loop potential, 243
- Context
 - independence of codon plots, 218
 - provides cues, 55, 418
- Continuity, 16
 - in DNA strand synthesis, 111
 - of genetic factors, 15
 - of germ-plasm (*see* Germ-plasm)
 - of inheritance, 28
 - between man and animals, 368
 - non facit saltum*, 189
 - as unnecessary difficulty, 157
 - of variation, 159, 167, 177
- Control experiment
 - coding versus non-coding, 79
 - to eliminate stress, 291, 360
 - 1-mer level shuffling, 99
 - neutral as frame of reference, 129
 - reciprocal experiment, 302
- Conventional phenotype, xxix
 - versus genome phenotype, xxix, 47, 132, 213, 335
 - and natural selection, 56
 - tracks environment, 17
- Copying
 - of DNA information, 17, 28, 107
 - of pattern, 32
 - from a template, 29, 104
 - of vibrations, 10
- CpG island
 - associated with *Alu*, 290
 - to identify hidden transcriptome, 289
 - major and minor, 355
 - is near promoter, 354
 - and self-complementarity of CG, 331
- Crick, F.
 - on base composition variability, 188
 - and brain function, xxxi, 369, 392

- Crick, F. (*cont.*)
 and DNA structure, xxiii, 27, 399
 his connectionist view of memory, 370
 on hybrid sterility, 58
 on mRNA discovery, 401
 his questioning of natural selection,
 xxviii, 34
 his unpairing postulate, 34, 179
 on religion, 370, 387
 and repetitive sequences, 273
- Crookes, W., 13
- Crowded cytosol
 as conducive environment, xxxi, 114
 early appearance in evolution, 305
 restricts water mobility, 309
 and volume exclusion, 321
- Crowther, C.R., 166
- Cytosine deamination
 in immune repertoire generation, 355
 in mitochondria, 115

D

- Darwin, C.
 on altruism, 289
 Butler's commentary on, 21, 400
 caveat of, 137, 396
 difficulties of, 169, 289
 disparaged classification, 4
 disparaged evolution in jumps, 189
 on external conditions, 141
 on formative matter, 5
 his prose, xxviii
 his theory of natural selection, xxiii, 19,
 137, 141, 159
 his theory of pangenesis, 13, 48
 his understanding of information concept, 5
 his work, 3, 339
 on human brain, 367
 and Lamarckism, 20
 on language, 25
 and latency, 21
 on material basis of heredity, 12
 as model for Grampus, 406
 questioned by Butler, xxvii
 questioned by Huxley, 157
- Darwin, E., 437
- Darwin, F., 59
- Data
 abundance of, 405, 409
 Animal Genome Size Database, 42
 extrapolation of, 367, 389
 graphical, xxxi
 mining, xxvii

- National Center for Biotechnological
 Information (NCBI), 68
 points in clusters, 356
 storage, 425
- Dawkins, R.
 on concealing meaning, 402
 criticism of Bateson by, 396
 criticism of Gould by, 398
 on genes and bodies, 7
 and panpsychism, 390
 and selfish genes, xxiii, 202
- Delbrück, M., 85
- Descartes, R., 382
- Development
 Butler's view on, 7, 9, 10, 19, 58
 as expression of a program, 5, 358
 fetal nutrition during, 358
 genes affecting, 45, 148, 166
 hybrid inviable when it fails, 148, 280, 336
 of immune repertoire, 282, 283, 286, 320
 of language, 5, 375
 Mendel's view on, 5
 and orthogenesis, 19
 as recall of memory, 7
 role of formative matter in, 5
 role of imprinting in, 358
 slow in human child, 375, 391
 stress-induced phenocopies during, 313
 switched by short lived proteins, 312
 tissue formation during, 307
 unscheduled transcription in, 291
- de Vries, H.
 hierarchical elements distinguished by, 14
 named gemmules as pangens, 14
 and polyploids, 187
- Differentiation
 cryptic, 340
 depends on recombination barrier, 200
 of a duplicated gene, 195
 of genes as microisochores, 218, 222
 of an isochore, 201
 of sex chromosomes, xxxi, 340
 of sexes, 337, 339, 342
 of sexes and species compared, 340
 of species, 147, 224
 within a species, 46, 193
- Dinucleotide. *See also* Oligonucleotide
 expected versus observed
 frequency of, 357
 frequency and genome design, 76, 356
 given methyl accent, 351
 self-complementary, 67, 74, 352
 species-specific frequency of, 355
 table of, 67, 92

- Diploidy
 - as adaptation for error-detection, 39, 185, 329
 - and chromosomes, 35
 - and protein concentration, 318
 - in retroviruses, 184
- Discontinuity
 - definition of, 53
 - in DNA strand synthesis, 111
 - facit saltum*, 189
 - of genes, 54
 - Huxley on, 157
 - language a mark of, 368
 - non-genic, 54
 - of texts, 49, 51
 - in variation, 53, 178
- Disorder
 - in our lives, 39
 - related to entropy, 295, 308
 - in the universe, 33
- Distribution
 - of accents in a population, 177
 - bell-curve of, 177
 - of heights in a population, 50, 145, 154
 - normal, 169
- Divergence, 136
 - as branching evolution, 137, 154
 - between camel and pig, 141, 150
 - controlled by genome phenotype, 130, 134, 136, 189
 - decreased by selective sweep, 129
 - essential precondition for, 171, 341
 - from ancestral sequence, 128, 169, 186, 201
 - between genome and its proteins, 134, 188
 - between globin genes, 195
 - between HIV and HTLV, 185
 - as indicator of phylogeny
 - (*see* Phylogenetic analysis)
 - between man and primates, 368, 375, 391
 - more at third codon positions, 195
 - between mouse and rat, 130, 135
 - between prokaryotes and eukaryotes, 198, 235
 - related to complexity, 139
 - thwarted by gene conversion, 195, 197
 - between twins, 357
- DNA, 31, 34, 68, 108, 111, 244, 289
 - accent of, 44, 45, 179
 - adaptation to high temperature, 204
 - alphabet of, 26, 331, 351
 - antiparallel strands of, 68
 - archive, 4
 - as a binary string, 36
 - breakage by sheering, 198
 - bubbles from duplex, 180
 - cell content of, 368
 - as containing life history, 11
 - damage, 33, 184, 227, 328, 329, 333, 334
 - depurination by acid, 103
 - as digital information store, 368, 371, 415
 - dinucleotides in, 67, 93
 - discovery as nuclein, 327
 - divergence as time axis, 131, 135
 - double helix model of, 29
 - duplex hemimethylation, 352
 - duplication as rule, 34
 - editing of, 351
 - as enzyme substrate, 98, 210, 224
 - five prime end of, 67, 111
 - from fossils, 149
 - G-quadruplex structure, 227, 297
 - G-T** mismatch in, 334
 - homostability regions in, 193
 - inconstancy drives protein treadmill, 99, 224
 - inversion, 34, 110
 - junk, xxx, xxxii, 57, 238, 288, 289, 295, 418
 - as language, xxi, xxv, xxvi, 44, 55
 - as living, 11
 - long range interactions in, 335
 - mapping of, 353
 - megabase deletions in, 288
 - melting of duplex, 296
 - methylation and brain memory, 369
 - methylation inactivates X-chromosome, 344
 - natural or artificial sequences, 418
 - Poland-Scheraga model, 100
 - polymerase (*see* DNA polymerase)
 - as a polynucleotide, 31
 - post-synthetic modification of, 353
 - promiscuity of, 33, 46
 - and protein divergences compared, 131, 188
 - quantity control, 39
 - recombination (*see* Recombination)
 - redundancy in, xxviii, 35, 36, 327, 341
 - repair of, 30, 32, 33, 116, 122, 224, 239, 328, 333, 334, 355
 - (*see also* Recombination repair)
 - repeats (*see* Repeats)
 - repetitive elements (*see* Repetitive element)
 - replication (*see* Replication)
 - restriction fragments of, 353
 - reverse and forward complements in, 68
 - satellite, 271, 273
 - in sea water, 75
 - separation of fragments of, 198, 353
 - stem-loop potential in
 - (*see* Stem-loop potential)

DNA (*cont.*)

- strand (*see* Strand)
- supercoiling of, 111, 181, 204, 274
- symmetry principle, 69, 271, 352
- synthesis (*see* Replication)
- terminology of repair of, 333
- three prime end of, 67, 111
- topoisomerase, 111
- transposition, 34
- trinucleotides in, 69
- uncoiling of, 86, 111
- vulnerability to heat, 203

DNA polymerase

- collision with RNA polymerase, 111
- direction of movement of, 108
- identification of, 29
- origin recognition by, 107
- repair thwarted by, 328

DNA Structure and Function by Sinden, 436

Dobzhansky, T., 142, 152

Dog, 13, 46, 70, 193, 224, 381

Dominance

- degree of, 54
- versus recessive, 142, 318, 337

Dosage compensation

- females accommodate to males, 342
- keeps aggregation pressure constant, 343
- and Muller's paradox, 342
- mutation affects females, 343

dsRNA

- dependent kinase (PKR), 292
- formation of, 113, 292, 293
- in Huntington's disease, 271
- as interferon inducer, 293
- as intracellular alarm, xxx, 271, 293, 297

Duret, L., 111

EEB. *See* Evolutionary bioinformatics

EBV

- as episome, 295
- latency of, 258
- nuclear antigen 1 (EBNA1), 258
- purine-loading of EBNA1 gene in, 297

Editing

- of DNA, 351
- of RNA, 360

Education

- avoiding excessive detail, 60
- limit of, 49
- videos, xxxi, xxxiii

Eichmann, K., 320

Einstein, A., 57, 399–401, 427, 429

Eject

- as inferred entity, 388
- as social thought process, 389
- as universal thought process, 389

Elements, 289

- hierarchical arrangement of, 14
- of language, 16
- permutation of, 5
- repetitive (*see* Repetitive element)

Eliot, G, xxxii, xxxiii, 23, 395, 399, 402, 405

Embryo

- as antigen source, 281
- as passive information recipient, 6
- as remembering parents, 6, 48
- unscheduled transcription in, 291, 312

Energetics

- of folding, 86, 107
- of helix formation, 92
- of rouleau formation, 307
- of supercoiling, 111

English

- Cockney accent of, 44
- has characteristic k-mers, 75
- language, xxxi, 44
- Pidgin version of, 64
- telegram to China, 16
- Tottenham Court Road, xxviii

Entropy

- and complexity, 254
- as driver of chemical change, 305
- as a force, 295, 308
- temperature increases, 29, 296
- and water, 296
- work done by, 295

Entropy-Driven Processes in Biology by

Lauffer, 295

Environment

- adaptation to, 19
- as cause of early death, 333
- colocalization in, 54
- as ecological niche, 223
- extreme, 202, 235
- of host anticipated by virus, 313
- incipient species as, 171, 185
- not matching species, 52
- permissive for peptide display, 312
- phenocopies induced by, 319
- of predators, xxx
- reproductive, 130, 170
- selects for survival, 85
- signalling to and from, 373
- temperature of, 230
- total DNA from, 75
- tracking change in, 4, 17

- twins to assess role of, 357
- uniformity of intracellular, 312
- Enzyme**
 - active site of, 127, 310
 - as ferment, 14, 305
 - cytochrome c oxidase, 215
 - deoxycytidine deaminase, 273
 - DNA-dependent RNA polymerase, 104, 107, 236
 - for DNA repair, 328
 - as an executive element, 14, 45
 - kinase as phosphate donor, 292
 - ligase, 33, 108, 294
 - loss of activity by aggregation, 307, 318
 - as a machine, 16
 - methyl transferase, 351
 - nuclease, 93, 353
 - phospholipase A₂, 240
 - as producer of color, 14, 318
 - protease, 33
 - purification, 305
 - restriction, 353
 - reverse transcriptase, 298
 - ribonuclease, 294
 - RNA-dependent RNA polymerase, 293
 - solubilization of, 305
 - substrates and products of, 305
 - topoisomerase, 111
 - xylanase, 231
- Epigenetic inheritance, 333, 351
- Epstein-Barr virus. *See* EBV
- Erewhon* by Butler, xxxii, 7
- Error**
 - catastrophy, 184
 - correction of, 26, 58, 182, 238, 327
 - detection of, 26, 48
 - in type-setting, 25, 153
- Error-checking**
 - needs redundancy, 48, 328
 - by non-message sequence, 238
 - not in AIDS virus, 184
- Escherichia coli*
 - CHI sequences in, 300
 - codon position plot of, 218
 - CpG dinucleotides in, 356
 - DNA length (bases), 422
 - same code as man, 417
 - trinucleotides of, 74
- Eukaryote**
 - codon position plot of, 221
 - definition of, 235
 - shares ancestor with prokaryote, 198
- Evolution, 195**
 - branching, between species, xxix, 137, 146, 150
 - classic triad of processes in, 150
 - concerted, 194, 197
 - dichotomies in terminology of, 56
 - divergent (*see* Divergence)
 - essence of problem of, 55, 211
 - of eye, 372, 375
 - general theory of, 398
 - linear, xxix, 137, 146, 177
 - and mankind's future, 4, 399, 426, 432
 - of nucleus, 235
 - of organic from inorganic, 377
 - from organic molecules, 3, 429
 - parallel among viruses, 133
 - rate and expression level, 344
 - rate and stem-loop potential, 243
 - rate of, 34, 122, 127
- Evolutionary bioinformatics.**
 - See also* Bioinformatics, the new
 - agenda of, xxv
 - emergence of, xxiv
 - founder of modern, 211
 - goals of, xxvii, 255
 - task of, xxvi, 3
- Exceptions**
 - GC-pressure extremes, 246, 258
 - protein-encoding pressure, 239
 - R-loaded EBNA1 gene, 258
 - R-loaded HBZ gene, 297
 - treasure your, xxvii, xxxiii, 385, 397
 - as unexplained residue, 387
 - W-loaded malaria genes, 263
- Exon**
 - conservation, 37, 241
 - discovery of, 235
 - location in DNA, 77
 - non-conservation, 241
 - oligonucleotide parity in, 76
 - size range of, 237
- F**
- Facts**
 - correctness of, 320, 402
 - Darwin as arranger of, 21
 - new way of looking at, 157, 400
 - simplest explanation of, 57, 427
 - and speculation, xxviii, 396
- Feather**
 - evolution of, 284
 - pigment of, 15
 - principle, 427

- Ferment
 - as enzyme, 14
 - is not itself transmitted, 14
- Fine-tuning
 - of intracellular antibody repertoire, 287
 - of molecular concentration, 40
 - narrows frame of reference, 310
 - of proteins in a common cytosol, 311, 343
- First parity rule (PR1). *See* Chargaff, E.
- Fish, 52, 288, 368
- Fisher, R., 141, 402, 404
- Fitch, W., 130
- Folding
 - equivalent patterns of, 107
 - path followed, 107
 - of randomized sequence difference, 91
 - of randomized sequence mean, 91
- Food
 - anorexia nervosa, 358
 - cannibalism, 357
 - has its own memory, 11
- Frontali, C., 263
- Fruit fly
 - CpG dinucleotides in, 356
 - hybrid sterility in, 341
 - paranemic pairing in, 180
 - strand symmetry in, 74, 76
 - Y-chromosome degeneration in, 342
- Function
 - collective as well as specific, 309
 - conservation correlated with, 51
 - conservation not correlated with, 52, 289
 - house-keeping, 355
 - loss through aggregation, 310
 - not location, as defining, 224
 - psychomotor, 8, 374
 - relative to gene dose, 342
 - units of, xxiii, xxx
- G**
- G+C**, 72 (*see also* Homostability regions; Isochore)
 - as accent of DNA, 179
 - controls strand pairing, 182
 - differences decrease recombination, 187, 195
 - differences early in speciation, 176
 - gene specificity of, 217
 - as a genic environment, 195
 - as indicator of reproductive compatibility, 176
 - in intergenomic recombination, 185
 - in intragenomic recombination, 340
 - intermediate levels confer flexibility, 222
 - as preserving agency, 201
 - pressure (*see* GC-pressure)
 - range among genes, 222
 - species specificity of, 213
 - uniformity at second codon position, 215
 - uniformity at third codon position, 199, 218
 - uniformity in a gene, 198
 - uniformity in a genome, 176
 - viruses that cohabit greatly differ in, 185
- Galton, F., xxxii, 12, 13, 15, 21, 47, 50, 53, 209
- Gamete
 - formation failure, 161, 166, 281, 336
 - formation of (gametogenesis), xxx, 162, 166, 316, 329, 346
 - imprinting of, 358
 - loss through exile, 316
 - as mailbag, 47
 - male and female equivalence, 6, 12
 - quality control of, 332, 346
 - transfer barrier to, 148, 151
 - two types in one sex, 66
- GC-pressure**
 - downward and upward, 214, 222, 257
 - genome-wide nature of, 72, 255
 - high downward in malaria parasite, 246, 263
 - versus protein pressure, 213
 - versus purine-loading, 226
- Gelbart, W., 4
- Gemmules
 - each run by a memory, 17
 - as germs, 6
 - named by Darwin, 13
 - as organic substances, 48
 - as pangens, 14
 - transfer from soma to gonads, 20
- GenBank, 68, 80, 414
- Gene, 285, 354
 - boundary of, 198, 202
 - centered bioinformatics, 3, 157
 - conversion, 194, 196, 239, 274, 330, 333
 - coorientation, 115, 245
 - copy-loss, 194
 - definition of, xxx, 51, 163, 200, 202, 236
 - discontinuity of, 54
 - distinctive (G+C)% of, 198, 201, 221
 - dosage (copy number), 318
 - dosage compensation, 342
 - duplication of, 194
 - encodes transferable characters, 168
 - for courtship display, 335
 - G-zero switch 2 (G0S2)*, 290, 355
 - G-zero switch 30 (G0S30)*, 357
 - hitch-hiking by, 129

- house-keeping, 332, 355
- incompatibilities between products of, 336
- isolation before change in function, 197, 224
- length in birds, 37
- length in EBV, 262
- length in malaria parasite, 246
- linkage (*see* Linkage)
- as microisochore, 218, 222, 334
- multicopy, 194, 197
- mutator, 224
- nef* in AIDS virus, 191
- non-allelic, 145
- overlapping another gene, 242, 245
- paradigm of, xxiv, 3
- restriction-modification complex, 354
- selfish (*see* Selfish gene)
- specific mutation rate, xxix, 132
- tissue specific expression of, 312
- transposition of, 110, 195
- as unit of antirecombination, 200
- as unit of preservation (*see* preservation)
- as unit of recombination, 200, 337
- as unit of transcription, 235
- versus species, 217
- with no protein product, 344, 354
- Xist* on X-chromosome, 354
- Genetic, 318, 338
 - assimilation, 319
 - code, 122
 - disease, xxv, 54, 269
 - dominance (*see* Dominance)
 - drift, 122, 144, 238, 407
 - engineering, xxv, 300
 - factors carried by gametes, 15
 - factors interpreted as genes, 15
 - fitness, 330
 - identity (twins), 357
 - linkage (*see* Linkage)
 - program, 4
 - revolution, 137
 - texts, xxi, xxiii
- The Genetical Theory of Natural Selection* by Fisher, 141
- Genetic science
 - attacked, 404, 407
 - classical, 58
 - launched, xxix
 - progress delayed in, 157, 397, 407
- Genocentrism, xxvi, 3, 22
- Genome, 130, 176
 - accent of, xxx
 - circular of lambdaphage, 109
 - circular of SV40 virus, 109
 - compactness, 34, 184, 243, 258
 - as a constellation of genes, xxiv
 - convergence of two, 186
 - divergence (*see* Divergence)
 - homostability (*see* Homostability)
 - human, xxiv, xxv, xxxiii, 422
 - human accelerated regions in, 375
 - hypothesis of Grantham, 211
 - imprinting, 358
 - as an information channel, xxvi, xxx, 3
 - information content of, 37, 253, 287
 - isochore (*see* Isochore)
 - phenotype (*see* genome phenotype)
 - rebooting of, xxxi, 329
 - selfish, xxiii, xxx, 223
 - sequencing project, xxiv, 219, 353, 403
 - size and strategy, 238, 317, 421
 - space limitation, xxvi, xxx, 64, 209, 238
 - species specificity of k-mers in, 74
 - understanding of, 4, 157
 - uniformity of (G+C) in, 176
 - as a unit of preservation, xxiii, 201
 - as a unit of recombination, 200
 - virus fragments in, 298
- Genome phenotype
 - codon position plot as index of, 215
 - controls divergence, xxv, 134, 189
 - as 'intragenomic phenotype', 139
 - multiple pressures on, 248
 - over-rules conventional phenotype, 130, 335
 - and physiological selection, 56
 - as reprototype, 47, 170
 - targeted by non-classical factors, 130, 211
- Genotype
 - change due to pattern change, 153
 - change enhances somatic flexibility, 55
 - as a constellation of genes, 4
 - destroyed by meiosis, 201
 - as a hierarchical element, 14
 - plasticity of, 313
 - primary constituents of, 16
- Geographical isolation. *See* Allopatry
- Geological Evidences on the Antiquity of Man* by Lyell, 5
- Germ. *See also* Gemmules
 - definition of, 13
 - equated with formative matter, 6, 47
 - recollections of, 8
 - reconstructed from scratch, 10
 - within-germs model, 10
- Germ-line
 - as distinct from soma, 13, 285
 - eternal cycling of, 148
 - house-keeping genes active in, 355
 - mutations in, 329

- Germ-line (*cont.*)
 selection within gonads, 346
 virus sequences in, 298
- Germ-plasm
 constituents of, 16
 continuity of, 13, 16, 284
 stirp, 24
- Giraffe, 368, 381
- Goldschmidt, R.
 character of the man, 402
 and chromosomal pattern, 137, 153, 176
 on complementary factors, 166
 distinguished genes and reaction system,
 153, 340
 informational viewpoint of, 153
 and macroevolution, 56, 397
 on monstrosity, 189
 non-genic viewpoint of, 152
 on phenocopies, 313
 and reaction system, 153, 186
 on sex and speciation, 339
 photo of, 56
- Goldsmith, O., 371
- Gonad, xxx, 16, 20, 147, 158, 162, 179,
 281, 334
- Gould, S.J.
 on biohistorians, xxvii, xxviii
 versus Dawkins, xxxi, 399
 on spandrels, 360
 supported Goldschmidt, 397
- Granovetter, M., 284
- Grantham, R., 211
- Gregory, R.P., 162
- Greig, D., 169
- Grouping. *See also* Selection of group
 of elements, 12
 of organisms, 396
 of points, 409, 413
 of symbols, 43
- Guyer, M., 162
- H**
- Hacking, I., 387
- Haeckel, E., 20
- Haldane, J.B.S., xxxiii, 335, 340
- Haldane, J.S., 396
- Hameroff, S., 390
- Hamilton, W., 376
- Hamming, R., 238
- Hand
 on deck, 291, 360
 information for, 17
 with six fingers, 18, 54, 121
 on text, xxxii
 that weaves, 16
- Haplosufficiency, 318, 330, 358
- Hebb, D., 370
- Hebert, P., 215
- Hedges, B., 137
- Height
 binomial distribution of, 145
 of child relative to parent, 50
 determinants of, 143
 dwarfism, 54, 347
 as Mendelian character, 161
 multigenic determination of, 145
 normal distribution of, 159
 of peas, 142
 of race, 146
 unigenic determination of, 155
- Helmholtz, H.L., 401
- Heredity. *See also* Inheritance
 as fundamental principle, 6
 as language, 14
 material basis of, 12
 and memory as one, 6, 9, 19, 333, 400
 as transfer of stored information, 6
- Heredity Genius* by Galton, 53
- Hering, E., 6, 9, 15, 22
- Hermaphroditism, 143, 148
- Herpes simplex* virus
 base composition of, 72
 trinucleotides in, 72
- Heteroaggregation
 advantage of, 311
 as defense mechanism, 313
 between mutated and normal
 proteins, 310
 as like with unlike, 306
 with virus protein, 312
- Heterozygote
 alleles differ in sequence, 330
 crossed with recessive, 337
 derivation from Greek, 144
 enzyme concentration in, 318
 returns to homozygosity, 196, 330
- Hilbert, D., 391
- History
 of bioinformatics, xxviii, 157, 425
 books, 404
 misinterpretation of, xxviii
 of science, 395, 398, 404
 sources for, 5
 as understanding, xxiv, 4, 402
- Hitch-hiking. *See also* Selection
 of fluctuation in base composition, 195
 of gene, 129

of k-mers, 97
 of papers, 52
 Holliday, R., 196, 203, 210, 273
Homo bioinformaticus, xxvi–xxvii, xxxii, 80
 Homoaggregation
 as defense mechanism, 313
 can inhibit activity, 307
 can promote activity, 307
 as like with like, 306
 physiological, 307
 prior to degradation, 307
 Homology search, xxx, 179, 188, 194, 239, 280, 281
 Homostability
 principle, xxx
 regions in DNA, 176, 198
 Homozygote
 derivation from Greek, 143
 derivation from heterozygote, 196, 330
 enzyme concentration of, 318, 342
 identical sequences of alleles in, 329
 recessive, 337
 Hood, L., xxvi, xxxiii
 Hooke, R, 401
 Hooker, J., 142
 Host
 to foreign organisms, 282
 as moving target, 313
 preimmunization of, 283
 stiffened defenses of, 280, 320
 Hotspot. *See* Meiosis; Mutation, hotspot;
 Recombination hotspot;
 Recombination, hotspot
 Huang, S, 407
 Human T-cell leukemia virus
 its HBZ protein like EBNA1, 297
 as a high (G+C) species, 186
 strategy of, 184, 295, 297
 Huxley, A, 403, 405
 Huxley, J., 35
 Huxley, T.H.
 on aging, 431
 on agnosticism, 428
 criticism of Darwin by, xxix, 157
 criticism of Romanes by, 160, 398
 and evolution in jumps, 189
 on experts, 413
 on purpose of life, 432
 and selfish genes, 6
 Hybrid
 discipline, xxvi, 58
 between DNA molecules, 36
 between RNA molecules, 35, 114
 vigor, 58, 335

Hybrid inviability
 as a developmental barrier, 148, 346
 and Haldane's rule, 336
 and not-self discrimination, 280
 as a reproductive barrier, 151
 view of Butler on, 58, 280
 Hybrid sterility
 and cryptic speciation, 340
 as a chromosome pairing barrier, 148
 cured by tetraploidy, 168, 187, 336
 cycle interruption by, xxx, 164
 as Darwin's weak point, xxix
 as a distinction in kind, 166, 169
 as expression of extreme outbreeding, 161, 335
 as expression of parental incompatibility, 170
 as expression of parental phenotype, 158
 as expression of partner-specific phenotype, 164
 as facilitator of divergence, 178
 and Haldane's rule, 336
 mule as example of (*see* Mule)
 as natural, 281
 as an isolating agency, 176
 as a journey from old to new, 170
 as a reproductive barrier, 47, 164
 view of Bateson on, 161, 164
 view of Butler on, 59
 view of Crick on, 58
 view of Crowther on, 166
 view of Naveira on, 341
 view of Romanes on, 158
 view of Wallace on, 170, 172

I

Ideas
 appear before words, xxvii, 5, 157, 376
 conflicting (*see* conflict)
 that fly, 9
 marketing of, 58, 201, 403, 407
 as memes, 52
 of Mendel, 52, 396
 when novel need more time, 401
Ignorance - How It Drives Science
 by Firestein, 386
 Immune receptor
 protein, 285, 311
 RNA, 285, 292, 299, 360
 Immune system. *See also* Not-self; Self;
 Thymus
 alerting of, 268, 271, 356
 autoimmune regulator (AIRE) protein in, 292, 312, 315, 345

- Immune system. *See also* Not-self; Self;
 Thymus (*cont.*)
 in bacteria, 299
 EBV evasion of, 297
 HTLV1 evasion of, 297
 in jawless vertebrates, 284
 malaria, evasion of, 268
 prototypic, 284, 285
 in snails, 284
 in unicellular organisms, 284
- Immunization
 against embryo, 281
 prophylactic, 266, 283
 against self, 313
- Impressions of Theophrastus Such*, by Eliot, 405
- Incompatibility
 between accents, 177
 between gametes, 158
 between meiotic chromosomes, 153, 197
 non-genic, 341
 between parents, 170
 of pattern, 153
- Inflammation
 extracellular, 287
 induced by necrosis, 316
 interferons and, 293
 intracellular, 287
- Information. *See also* Phenomenon of
 arrangement
 active versus passive in DNA, 210
 analog, 369
 channel capacity for, 418
 compressibility, 371
 concept of, 15, 47, 327
 conflict between forms of, xxix, 63,
 209, 418
 context of, 54, 418
 digital, 369
 direction of flow of, 46
 externalized mental, 373, 374, 425
 as held in trust, 432
 holographic storage of, 374
 levels of, xxix, xxx, 44, 54, 64, 210
 limit of scope of, xxix
 in macromolecules, 17
 meaning of, 417
 measurement of, 37, 253, 415
 mental, 6–9, 17, 44, 367, 371, 382, 425
 penetrance of, 22
 preservation of, 18, 46, 51, 200, 425
 primary as text, 44, 63, 128
 primary, as the conventional phenotype, 47
 primary, destroyed by shuffling, 44
 primary, potential to carry, 240, 415
 redundant, 26, 136, 184, 209, 235
 regulation of expression of, 22
 secondary, 44
 secondary, as accent, 44, 128
 secondary, as an agency, 52
 secondary, as the genome phenotype, 47
 secondary, impaired by primary, 128
 secondary, not destroyed by shuffling, 44
 secondary, potential to carry, 240, 415
 statistical analysis of, 50, 409, 415
 superposition of, 390, 392, 418
 symbolic representation of, 5, 9, 376
 syntax of, 375
 theory, 36, 222, 415
 trade-off between forms of, xxx, 210, 418
 use of word by Butler, 9
 use of word by Romanes, 8
 use of word by Wheeler, 377
 when stored is memory, 5, 367, 371
- Inheritance, 20. *See also* Blending; Heredity
 of acquired characters (*see* Lamarck, J.)
 discrete, 142
 epigenetic, 353
 horizontal, 147
 Mendelian pattern of, 162
 non-Mendelian pattern of, 161
 preformation model of, 10
 transgenerational, 359, 426
- Insect. *See also* Fruit fly
 extracellular RNA in, 297
 firefly, 373
 flight of, 284, 368
- Instinct
 and cerebral wiring, 17
 Romanes' view of, 8
- Intelligent design, 20, 377, 398, 425, 429
- Intracellular Pangenesis* by de Vries, 14
- Intron
 and conflict resolution, 237
 conserved in venom gene, 241
 evolution early or late?, 238, 241
 function of, 238
 in *Xist* gene, 344
 interrupts any information, 237
 length proportionate to exon length, 237
 and low complexity segment, 246
 oligonucleotide parity in, 76
 removal of, 236
 tends to pyrimidine-load, 246
 third codon position as mini, 262, 268
- Isochore
 as genome sector, 176
 hitch-hiking by, 195
 as intragenomic isolator, 193

- as macroisochore, definition of, 200
- micro, as a genic **GC**-niche, 222
- micro, differentiation of, 218
- as microisochore, definition of, 199
- origin of, 195, 198, 200

J

- James, W., 397
- Jenkin, F.
 - his sphere of variation, 49, 146
 - questioned natural selection, 53
- Jerne, N. K., 404
- Johannsen, W.
 - and mathematics, xxvii
 - pure lines of, 50, 146
 - and the word gene, 51
 - and words phenotype and genotype, 15
- Jurassic Park* by Crichton, 149

K

- Kalmus, H., 132, 200
- Kandel, E., 371
- Kant, I, 396
- Kantsaywhere* by Galton, xxxii
- Kimura, M., 86, 219
- Kissing
 - in chromosome pairing, 180
 - entropically driven, 296
 - favoured by crowded cytosol, 296
 - inhibition by purine-loading, xxix, xxx, 295
 - between loops, 84, 113, 180, 185, 197, 244
 - as metaphor, xxx, 182
 - between tRNA and mRNA, 84, 125
 - between *Xist* and *Tsix*, 344
- Kleckner, N., 180, 188
- Koëlreuter, J.G., 12
- Kornberg, A., 29
- Krisch, H., 97

L

- Lamarck, J.
 - acceptance by Butler, 20
 - acceptance by Darwin, 140
 - acceptance by Galton, 24
 - acceptance by Haeckel, 20
 - acceptance by Spencer, 20
 - evidence supporting, 24
 - and inheritance of acquired characters, 20, 54, 298, 359
- Lankester, E.R., 398

Language

- acquisition by a child, 64, 375
- allied, 44
- of brain ('mentalese'), 376
- confusion by, 402
- convergence of, 44
- divergence of, 44, 136
- of DNA, xxi
- of genes, 44, 55
- k-mer frequencies in, 75
- like a biological species, 5, 136
- as metaphor, 16, 65, 128
- sign, 373
- spoken, 12, 46, 128, 373, 415
- syntax of, 375
- textual, 373
- as uniquely human, 367, 375

Larmor, J., 60

Lashley, K., 370

Latent

- AIDS virus, 21, 295
- character, 21
- part, contrasted with patent part, 21
- phenotype unmasked, 319
- prions, xxv
- sterility factors, 164

Lauffer, M., 295, 308

Le Bourgeois Gentilhomme by Moliere, 64

Letter

- alphabet of, 14, 371
- base as, 30
- chromosome as carrier of, 153
- fifth, 351
- frequency in DNA, xxix
- frequency in a language, 44
- in print block, 12, 153
- permuted, 154
- shuffled, 44
- unread in mailbox, 47

Lewes, G.H., 8, 405

Lewontin, R., 360

Life

- difference from non-living, 11, 429
- history in DNA, 11
- how to live it, 426
- as memory, 18
- origin of, 368, 372, 377, 429
- as phenomenon, 43, 305

Limit

- of attention, 390
- in brain plasticity, 387
- in brain redundancy, 387
- in channel capacity, xxx, 418
- in character transfer, 168

Limit (*cont.*)

- of codon flexibility, 86, 124
- of concentration of a macromolecule, 40, 306, 309
- in death style, 316
- of education, 49
- in enzyme reaction rate, 318
- of gene flow, 168
- in genome space, 64, 209, 238, 283, 298
- of memory, 425
- of natural selection, 137, 367
- in number of base permutations, 116
- in protein concentration, 318
- of a race, 49
- of range of discourse, 65
- of scope of information, xxix
- of the simple, 57, 427
- of somatic flexibility, 55
- of a species, 49, 168, 311
- of a theme, 59, 137
- of variation, 18, 53, 141, 335
- of what mind can be, 389
- of what the mind can conceive, 201, 340, 370, 372, 376, 387, 395, 402, 408

Line

- absence of, 433
- arrangement of germs in a, 233
- best matching points, 409
- between fact and speculation, xxviii, 396, 399
- horizontal or tilted, 412
- marking a discontinuity, 367
- politically correct, 425
- statistical description of, 409

Linkage

- of cooperating genes, 337, 354
- facilitates hitch-hiking, 129
- of isochore to gene, 195

Little, P., xxvi

Locus poenitentiae, 48

London

- persistence of, 431
- in Victorian times, xxxii, 6, 10

Lorber, J., 383

Low complexity segment. *See also* Repeats

- abundant at **G+C** extremes, 255
- correlates with base **A**, 263
- correlates with surface domain, 268, 287
- detected with Seg program, 259
- identification of, 255
- in EBV, 258
- in malaria parasite, xxx, 263, 266
- non-repetitive, 263, 269
- and purine-loading, 246, 266

- role at nucleic acid level, 255, 268
- as simple sequence, 246, 255

Luck or Cunning, by Butler, 9

Lyell, C., 5

Lymphocyte

- auditioning, 292
- B-cell, 282
- cytotoxic, 308, 312, 313
- holes in repertoire of, 320
- positive selection by near-self, 320
- repertoire molded by self, 313, 315
- T-cell, 185, 292, 313

Lysenko, T., 404, 407

M

Machiavelli, N, 428

Macroevolution, 56

Macromolecules

- assembly from subunits, 17
- Bateson's factors as, 16
- crowding of, 305
- as enzymes, 98
- fine-tuning of concentrations of, 40, 342
- functions of, 16, 285
- informational, 16, 17
- self and not-self, 39
- turnover of, 33

Macromutation

- limited power of, 121, 169
- types of in DNA, 339

Magnetism, 28, 169, 373

Malaria parasite

- circumsporozoite protein of, 266
- codon position plot of, 221
- haploidy of, 35
- low complexity elements in, xxx
- purines and protein length, 228
- recombination barrier in, 186
- synonymous mutation in, 130

Markov chain analysis, 78, 97

Marx, K., 7, 428

Maside, X., 341

Materialism

- of Butler, 20
- of Hering, 8
- of Spencer, 12

The Material Basis of Evolution by Goldschmidt, 398*Materials for the Study of Variation* by Bateson, 14, 52, 160

Mathematics and biology, xxvi, 80, 161, 335, 390, 391, 407, 414

Matsuo, S., 195

- Matthew, P., 23
- Mayr, E., 142, 398, 399
- McCarthy, J., 374
- Meaning
- as abstract entity, 44
 - general, of codon position plots, 218
 - of a group of symbols, 43
 - of information, 417
 - of a message, 55
 - ultimate, xxxii, 430
- Medium
- as paper, not message, 30, 64
 - as a phosphate-ribose chain, 30, 64
- Meiosis. *See also* Chromosome, conjugation
- cell division during, 167, 332
 - chromosome pairing in, 179, 186, 281
 - failure of, 148, 162, 164, 179, 182, 281, 330, 334, 336
 - recombination during, 274
 - specific to gonad, 160
 - template-critical outcomes of, 329
 - three steps in, 329
- Melanin, 21, 313
- Memory. *See also* Brain
- cancelling of, 11, 359
 - in computer, 373
 - conflict between two parental, 58, 281
 - connectionist view of, 370
 - by embryo of parents, 6
 - as 'engrams', 370
 - exceptional, 385
 - extracorporeal, 372, 376, 386
 - between generations, 7, 143, 346
 - and heredity as one, 6, 9, 19, 333, 400
 - immunological, 282, 354
 - in extracellular immunity, 298
 - in intracellular immunity, 298, 299
 - long-term, xxxi, 368, 372, 381
 - magnitude of, 371
 - many modes of, 18
 - of parents, 48
 - as phenomenon of arrangement, 15, 371
 - as a process, 18
 - remote, as in cloud computing, 373, 390
 - runs a gemmule, 17
 - in savants, 371, 391
 - scaling to head size, 385
 - short-term, 374, 376
 - specific localization of, 382
 - as stored information, xxviii, 5, 11, 18, 368
 - unconscious, 8
- Mendel, G.
- and the academic cloister, 428
 - Bateson's advocacy of, 4, 161
 - cited by Romanes, 155
 - as a discipline-crosser, 58
 - ideas of, 52, 396
 - laws of, xxiii, 53, 162, 396
 - and pea breeding, 142, 337
 - ratios of, 144, 161
 - view on development, 5
- Menopause, 316
- Mental Evolution in Animals* by Romanes, 8
- Message. *See also* Information
- address of, 375
 - classifying, 55
 - erased (cancelled), 57
 - as a letter sequence, 64
 - material basis of, 376
 - may exist without meaning, 417
 - meta, 55
 - as pattern, 30
 - personal privacy of, 374
 - primary, 44, 55
 - secondary, 44
 - simple, for marketing, xxiv
 - syntax of, 375
 - as written on a medium, 30, 373, 417
- Metagenomics, 75
- Metaphor. *See also* Language
- accent as, 177
 - can mislead, 57, 58
 - from computer technology, 371, 386
 - friends and acquaintances as, 284
 - from fuel usage, 135
 - from money, 39
 - from politics, 154
 - kissing as, xxx
 - sword and scabbard as, 167
- Method of Mechanical Theorems*
- by Archimedes, 57
- Micromutation
- power of, 121
 - role in branching evolution, 169
 - types in DNA, 339
- Middlemarch* by Eliot, xxxiii, 405
- Miescher, J.F.
- on heredity as language, 14
 - photo of, 48
 - on sex, 48, 238, 327
- Migration
- of birds, 8
 - from equator, 21
- Mind
- asleep, 390
 - ejective hypothesis of, 388
 - as inferred entity, 388
 - in other than brain mode, 389

- Mind (*cont.*)
 and panpsychism, 390
 theory of (ToM), 388
 unconscious, 390
- Mitochondria
 barcode in DNA, 116, 215
 fusion of, 116
 gene transposition within, 110
 good for genomes studies, 116
 low in (G+C), 116
 maternal inheritance of, 116
 negative purine-loading in, 113
 pyrimidine-loaded RNAs of, 115
 role in apoptosis, 116
- Molecular biology
 classical period of, 16, 22
 and evolutionary theory, 399
- Molecular chaperone
 affects dose-response curve, 318
 client proteins of, 319
 as a folding agent, 107
 and the heat-shock response, 317
 as mutational buffer, 319
 opposes aggregation pressure, 318
 as professional interactor, 319
- Molecules
 archaeology of, 57
 capable of memory, 9
 capable of sex, 35
 competition between, 6
 primordial, 251
 resident intracellular, 286
 as Tekke Turcomans, 9
 turnover of, 33, 431
 vibration of, 11, 28
- Mongolism. *See* Brain disease, Down syndrome
- Morgan, T.H., 402
- Mule, 58, 59, 157, 169, 179, 281, 330
- Muller, H.
 and chromosome pairing, 179, 399
 and dosage compensation, 342
 genic viewpoint of, 142, 152
 and molecular vibration, 28, 87
 as mentor of Watson, 32
 paradox of, 342
- Müller, M., 367, 373
- Murat, P., 297
- Murray, D., 376
- Music, 334, 391
- Mutation, 130, 312, 339, . *See also* Phenocopy
 accepted, 10, 122, 329
 affects dosage compensation, 343
 affects function and/or solubility, 319
 amino acid-changing, 124
 back, 133
 bias, 78, 213, 288, 355
 buffering by molecular chaperones, 319
 collective, 151, 341
 compensatory, 132, 134, 224
 conditional, 319
 correlation between different types of, 132
 cryptic, 121, 160, 164, 189, 340
 in duplicated gene, 194
 effect depends on position, 127
 forward excludes back, 133
 genocopy not phenocopy, 312
 gonad-specific, 158, 329
 by insertion of retrovirus, xxv
 hotspot, 355
 lethality in homozygotes, 329
 lethality of second codon position, 124
 localized and general, 210
 macro (*see* Macromutation)
 meltdown in retroviruses, 184
 methylcytosine to thymine, 355
 micro (*see* Micromutation)
 as a microchemical accident, 32
 mutator genes as cause of, 224
 neutral, 86, 127, 129, 211, 288
 non-synonymous, 124
 path of base changes in, 125
 phenocopy not genocopy, 292
 pleiotropic effect of, 210, 311
 pre-meiotic, 346
 proximity between two, 133
 by radiation (*see* Radiation)
 rate of (*see* Mutation rate)
 registering for later repair, 332
 repair of, 122
 after reproductive life, 329
 spontaneous, 32, 121, 137
 spread in a population, 122, 169
 as a step-wise process, 125, 152, 153, 280
 synonymous (non-amino-acid-changing),
 124, 134
 synonymous site saturation, 131, 133
 synonymous sites rarer, 131
 systemic, 153, 176
 transition, 125, 328, 330
 transversion, 125, 330
 violation of neutral assumption for, 130
 vulnerability of CpG, 355
- Mutation rate
 accepted rate an underestimate, 329
 analogy with public speaking, 128
 gene-specific, 132
 high in DNA recognition proteins, 224
 high in testes, 346

- high in viruses, 311
 - low in highly expressed proteins, 344
 - ratio of rates of different types, 129, 134, 135
 - synonymous sometimes zero, 130
 - temporal calibration of, 128
- Muto, A., 213
- N**
- Nägeli, C.
- on material basis of heredity, 12
 - on mechanism of heredity, 5
 - and orthogenesis, 19
- Nakata, A., 299
- Natural selection
- as agency, 51, 146, 152, 201, 202
 - and codon bias, 211
 - contrasted with artificial selection, 137
 - contrasted with physiological selection, 56
 - elimination of long palindromes by, 244
 - empowered by reproductive isolation, 171, 179, 189, 281
 - function optimized by, xxiii, 85, 204
 - as fundamental principle, 6, 127, 201
 - greater in early life, 333
 - and hybrid sterility, 170
 - indirect action on species, 223
 - level of operation of, 223, 290
 - limited power of, 53, 137
 - preservation not guaranteed by, xxiv
 - questioned by Bateson, 53
 - questioned by Crick, xxviii, 34
 - questioned by Darwin, 140
 - questioned by Fisher, 141
 - questioned by Haeckel, 20
 - questioned by Hedges, 137
 - questioned by Jenkin, 53
 - questioned by Ohno, 76
 - questioned by Pagel, 136
 - questioned by Romanes, 406
 - questioning of sufficiency of, xxiii, 169
 - as reproductive barrier, 147
 - role in gene duplication, 194
 - and sexual selection, 137
- Nature
- acts of, xxv
 - confers sterility barrier, 281
 - experiments of, 382
 - level at which she wrote, 420
 - and nurture, xxxii, 357
 - her prime function, 330
 - not prescient, 238, 359
 - scope of her attention, 334
 - as stem-loop installer, xxix
 - under no obligation, 387
- Naudin, C., 142
- Naveira, H., 341
- Near-self
- contrast with 'altered self', 320
 - discrimination from self, 310
 - immunoreceptors, xxx
 - primes positive selection, 320
 - as a subset of not-self, 280
- Negative selection
- in exons, 38
 - in generation of immune repertoire, 282, 315
 - impairs reproductive success, 127
- Nematode worm
- chromosome accounting in, 80
 - recombination in, 180
 - strand symmetry in, 76, 97
 - telomeres of, 272
- Neumann, J. von, 372
- Newton, I, 399, 401
- Nirenberg, M, 404
- Not-self
- detection of, 310
 - includes the near-self subset, 280
 - recognition of nucleic acid as, 286
- Novella, I., 133
- Nucleotide. *See also* Oligonucleotide
- as building block, 30, 107
 - channeling, 33
 - structure, 30
- Nucleus
- ablation of parts of, 369
 - archival type in ciliates, 285
 - in bird red blood cells, 42
 - chromatin threads in, 14
 - as defining eukaryote, 198, 235
 - evolution of, 235
 - recombinational isolation in, 186
 - selection pressure differs from cytoplasmic, 95
 - S-phase volume doubling of, 332
 - working type in ciliates, 285
- Number
- of allelic pairs, 146
 - of brain cells, 372
 - of branch points, 136
 - of chromosomes, 169
 - of genes affecting development, 148
 - and meristic variation, 121, 145
 - and parity, 65
 - of sexes, 66, 327, 342

Number (*cont.*)

- of sterility factors, 341
- scaling of information to brain size, 385
- simple enumeration fails, 389
- of unit characters (genes), 163

Nussinov, R., 79, 97

Nutshell, this book in a, 418

O

Occam, W., 17, 57, 427, 433

Of Molecules and Men by Crick, 370

Ohno, S., 65, 76

Oligonucleotide. *See also* Dinucleotide;

Trinucleotide

- derivation from Greek, 72
- frequencies in phylogenetic analysis, 75, 421
- frequency as driving force, 188
- hierarchy in a species, 75
- overlap, 97
- parity is primary, 77
- relationship between orders of, 75, 77, 356
- reverse complements in introns, 77

Organelles. *See* Amoeba; Mitochondria

Organism

- computation of, xxvi
- elementary, 284
- as a microcosm, 6
- reconstruction from DNA, 149
- social, 388
- variant, 121, 137, 170

Orgel, L., 273

Origin

- of human mental power, 368
- of life, 368, 377, 429
- multiple for replication, 108
- of replication, 108
- of variation as origin of species, 141

Origin of species

- equated with variation, 18
- essence of problem of, 147, 163, 168, 169
- as interruption of reproductive cycle, 336
- like origin of sex, 339
- needs barrier breached, 49
- as a physiological question, 13

The Origin of Species by Darwin, 137, 141, 368

Osawa, S., 213

Ovule, 5, 158

Ovum

- implantation in uterus, 316
- and random drift, 144
- receptors on, 280
- as successful auditioner, 346

P

Pagel, M., 136

Pairing

- between **A** and **U** in RNA, 84
- antiparallel, 84
- chromosome barrier to, 166, 197
- between chromosomes in tetraploids, 186, 336
- energy values for dinucleotides, 92
- failure due to pattern change, 153
- between **G** and **T** in DNA, 87, 92, 355
- between **G** and **U** in RNA, 84, 94
- between **I** and **C** in RNA, 360
- interstrand, 83, 84
- intrastrand, 83, 96
- like-with-like, 168
- long-range, 79, 96
- between loops, 79
- of meiotic chromosomes, 153, 162, 179, 181, 281, 330
- paranemic, 180, 196
- between purines and pyrimidines, 28
- reversal to expose bases, 34, 179
- sense-antisense, 27, 70, 113, 344
- sense-sense, 27
- stability numbers of Tinoco, 88
- strong between **G** and **C**, 88
- Watson-Crick, 28, 31, 87, 168, 182, 204, 274
- weak between **A** and **T**, 88

Palimpsest as a bad metaphor, 57

Palindrome

- with **AT**-rich repeats, 244
- CG** as self-complementary, 352
- Greek derivation of, 65
- and inverted repeats, 97, 244
- in humans, 243
- as obeying parity rule, 65, 77
- properties of, 31
- quasipalindrome as precursor of, 97
- with trinucleotide repeats, 243
- in Y chromosome, 341

PAM matrices, 250

Pangensis

- extended by de Vries, 14
- extended by Miescher, 48
- implies representation of parts in gamete, 14
- theory of, 13

Paper

- of Gould, 398
- as medium, not message, 10, 30
- of Mendel, xxix, 52

Paradigm

- classical Darwinian, xxxii
- genic, xxiv, 4, 157, 342

- for hybrid sterility, 334
- shift, xxiv, 137
- Paradox
 - addressing of, 400, 401
 - of autoimmune disease, 314
 - bacteriophage, 85
 - GT** and recombination, 273, 300
 - Muller's, 342
 - of stability and variability combined, 10
 - of understanding history and subject, 5
 - viewed as absurd, 397
- Paraneoplastic disease, 313, 314
- Parity, 25
 - measurement of, 254
 - at oligonucleotide level, 420
 - within one strand, 66, 182, 420
 - primary at oligonucleotide level, 77
 - rules (PR1, PR2) (*see* Chargaff, E)
 - between two lines of text, 25
 - between two sexes, 66
 - between two strands, 65
- Pathogen. *See also* Virus
 - death if unmutated, 127
 - identification from base composition, 136
 - implication of its genome size, 317
 - as not-self, 285
 - preadaptation of, 313
 - recruitment of host self-antigens by, 313
 - resistance to heat, 203
 - RNA transcript as, 271
 - strategy of, 320
 - surface antigen of, 127
- Pattern
 - duplication of, 32
 - of genotype, 56, 153
 - incompatible chromosomal, 153
 - Mendelian, 161
 - of methylation, 351
 - non-Mendelian, 359
 - of stem-loops, 182
- Pea, 5, 142, 143, 146, 163, 359
- Pearson, K., 404
- Pellew, C., 359
- Penfield, W., 382
- Penrose, R., 390
- Perec, G., 65
- Periodical, 9
 - rhythms, 9–11
 - sequence elements (*see* Sequence, periodicity in)
- Phaestos Disk, 4
- Phenocopy
 - distinct from genocopy, 313
 - no genetic change, 319
- Phenomenon
 - of arrangement, 15, 371
 - associated with memory, 7
 - of imprinting, 358
 - of in-series redundancy, 35
 - of life, 43
 - of rogues in peas, 359
 - paranormal, 376
- Phenotype, 56
 - conventional (*see* Conventional phenotype)
 - determined by proteins, 45
 - genome (*see* Genome phenotype)
 - as a hierarchical element, 14
 - intragenomic, 139
 - introduction of the term, 15
 - not affected by sterility factors, 164, 340
 - organismal, 139
 - plasticity of, 56, 312, 313
 - as a set of characters, 4
 - sterility-of-offspring as a parental, 164, 170
- Phillips, G., 78
- Philosophy
 - agenda of, xxv
 - debate on certainty, 382
 - of mind, 388
 - and the soul, 387
- Phylogenetic analysis, 57, 75, 136, 137, 176, 184, 255
- Physics
 - biology not supplant, 377
 - and brain, 374, 390
 - as endless veil lifting, xxvii
 - laws of, 376
 - paradox in, 85
 - quantum mechanics, 369, 391, 401
 - subatomic particles, 372
- Physiological, 160
 - aggregation, 307
 - complements, 172
 - process of information retrieval, 18
 - reaction system, 153
 - selection (*see* Physiological selection)
 - self-destruction, 316
- Physiological selection
 - critics of, 160
 - follows collective variation, 159
 - initially cryptic, 159
 - named by Romanes, 158
- Pizzi, E., 263
- Planck, M., 401
- Plant seed
 - development of, 7
 - early germinating, 158
 - gametic contribution to, 5

- Plant seed (*cont.*)
 genetically modified, xxv
 memory lost when eaten, 11
 pure lines of, 50
- Poetry
 as an error-detecting device, xxviii, xxix, 64
 that rhymes, 391
 as unnatural, 63
- Polanyi, J., 405
- Polarity
 of biological reading, 107
 of biosynthesis, 106, 111
 of DNA strands, 31
 of proteins, 107
- Pollen, 5, 12, 158, 161
- Polymorphism
 due to microsatellites, 273
 as a hazard, 315
 as an individualizing factor, 313
 of malaria protein, 266, 268
 is natural, 333
 not neutral, 313
 prevents viral anticipation, 287, 298
 terminology relating to, 287
- Polyploidy. *See* Tetraploidy
- Population
 allele establishment in, 152
 apocalypses affecting, 333
 bottle-neck, 127, 129, 130
 diversity lost, 130
 expansion of, 239, 305
 homeostatic regulation of size of, 316
 purging of bad genes from, 330
 size among bird species, 217
 variation within, 177
 variation within subset, 159
- Positive selection
 for evolution of dosage compensation, 342
 for immune receptor function, 311
 for improved venom, 241
 of immune repertoire, 298, 319
 of a malaria gene, 266
 metaphor for, 239
 of species as a group, 170
 promotes reproductive success, 127
 to react with near-self, 282, 320
 role of codon position in, 197
- Power
 of assuming a character, 14
 to cause is transmitted, 15
 to destroy a pigment, 15
 of hereditary transmission equal in sexes,
 6, 12
 inborn, 8
 to invoke character, is transmitted, 14
 of recalling memory, 21
 to vary beyond Jenkin's sphere, 53
- Prabhu, V., 69
- Predator
 in arms race, 239
 of deer, 141
 external and internal, xxx
 of rodents, 240
- Preservation. *See also* Conservation
 of ancient texts, 52, 57
 gene as a unit of, xxx, 46, 201
 of a gene duplicate, 195, 230
 genome as a unit of, 201
 of hybrid sterility, 170
 imperfections of, 22
 of information, 18, 51, 200, 425
 of an isochore, 195
 as more fundamental than function, xxiii,
 52, 201
 needs a barrier, 58
 needs reproductive isolation, 176
 as precondition for selection, 52, 197, 396
 by secondary information, 46
 of structural element in protein, 244
 of words, 157
- Pressure
 for aggregation, 308
 definition of, 257
 general and local, 248
 summary of, 248
- Principle component analysis, 73, 76, 413
- Problems in Genetics* by Bateson, 15
- Process
 cyclic, 147
 of discovery, xxviii, 5, 396, 407
 evolutionary, 53, 430
 of heredity, 6
 mental, 18, 389, 391
 new, from triad to tetrad, 150
 its separateness from result, 121
- Program
 for cell death, 116
 of embryonic development, 5
 for an enzyme, 17
 genetic, 4
- Prokaryote
 archaea as, 205
 bacteria as, 198
 definition of, 235
 shares ancestor with eukaryote, 235
- Promoter
 differential methylation of, 358
 RNA polymerase binding to, 107, 236,
 290, 354
- Protein, 239

- active site of, 95–96, 127
- as an aggregating agent, 308
- aggregation in brain disease, 271
- as an amino acid chain, 45
- antigenic determinant of, 308
- as an immune receptor, xxx, xxxi, 285, 311
- circumsporozoite, 266
- coaggregation with unrelated proteins, 270, 310
- collective functions of, 307, 308, 343
- complexity of, 256
- concentration fine-tuning of, 309–311, 317, 342, 345
- concentration is independent of cell sex, 343
- defeats DNA, 244
- degradation to peptides, 308
- denaturation, 308
- as digital information store, 416
- divergence versus genome divergence, 134
- DNA recognized by, 210
- domain swapping, 238
- domain, functional, 246
- domain, non-functional, 268
- domain, structural, 230
- element, structural, 245
- as an enzyme, 45
- evolution after RNA, 239
- evolution and expression level, 344
- as an executive unit, 22
- factors determining concentration of, 310, 343
- fellow travelers of, 310
- flexibility to change, 95, 130, 215
- for peptide-display, 308, 312
- functions other than its specific, 285, 311
- heat-shock, 187, 291
- histone, 360
- insulin, 382
- interdomain space is for simple sequence, 258
- interdomain space is less in thermophiles, 230
- length affected by base composition, 229, 246, 266
- major histocompatibility complex (MHC), 308, 320
- as a mini-molecular chaperone, 319
- ORF encoding a, 126
- peptide chain in, 84
- plasma, 306, 307
- polarity of, 107
- pressure (*see* Protein-pressure)
- primary function of, 288, 311, 344
- rhodopsin, 375
- secondary functions of, 288, 311, 345
- sequence alignment gaps, 249
- sequence alignment scoring, 249
- solubility of, 269, 305, 309, 317
- species invariance of, 215
- specific activity of, 309
- structural (collagen), 33
- surface hydrophilic, 268, 287, 344
- synthesis favored by purine-loading, 296
- synthesis in brain, 374
- synthesis inhibited by dsRNA, 292
- synthesis of, 45, 85, 123, 129, 297
- tubulin as memory store, 369
- turnover, 33, 308
- variable parts of, 243
- Protein-pressure
 - on codons, 187, 213, 214, 222
 - versus DNA, 79, 239
 - dominance of, 245
 - extreme in overlapping regions, 246
 - for function, 222, 268
 - versus **GC**-pressure, 213, 218
 - versus purine-loading pressure, 225
 - versus **RNY**-pressure, 127
 - as selective force, 79
- Proteosomes, 312
- Protozoa, 176
 - amoeba (*see* Amoeba)
 - conjugation in, 285
 - sophistication of, 285
 - Tetrahymena, 176, 284
- Punnett, R.C., 347
- Purine-loading
 - in bacteria, 225
 - and base clusters, 105
 - as correction guide, 334
 - as dominant pressure on protein length, 266, 269
 - in EBNA1 gene, 258, 262, 297
 - versus fold pressure, 266
 - versus **GC**-pressure, 226
 - as informational constraint, xxix
 - as inhibitor of kissing, xxix, xxx, 113
 - inversely related to (G+C)%, 116
 - location in loops, 105, 112
 - in malaria parasite, 265
 - negative in mitochondria, 113
 - and optimum growth temperature, 204
 - placeholder amino acids
 - support, 228
 - versus protein-pressure, 225, 226
 - versus **RNY**-pressure, 127, 226
 - to self-discriminate, 293
 - in thermophiles, 105, 204
 - ways to achieve, 257
 - wide distribution among species, 113

Pyrimidine-loading

- by EBV, 258, 295
- by HTLV1, 295
- in mitochondria, 115
- rare in bacteria, 225
- in some viruses, 116

Q

Question

- chicken-and-egg type, 110, 255
- DNA variability versus amino acids, 188
- historical, xxvii
- horse-and-cart type, 77
- of genic and non-genic causes, 169
- of mind and brain, 370
- ultimate, xxxiii, 367, 430

R

Race

- cross between different, 340
- habitual actions of, 8
- as a long-lived individual, 49, 333
- and phenotypic differentiation, 340
- as a variety, breed or line, 340

Radiation, 19, 21, 32, 203, 312, 341, 372

Random

- assortment of chromosomes, 144, 329
- element at higher oligonucleotide orders, 78
- events, 25, 39
- fluctuation in base composition, 195
- forces produce parity, 66
- generation of antibody repertoire, 282
- genetic drift, 144, 238, 407
- inactivation of X-chromosome, 342
- interactions between DNA bases, 183
- mixing of base letters, 73, 88, 89, 91, 421
- mutation as, 124, 332
- position of a dice, 18
- selection for sequencing, 414
- turning of stones, xxviii
- usage of codons, 211

Ray, J., 166

Reaction system

- as a unit of recombination, 156
- Clausen's concept of, 186
- is different from genes, 153, 340
- sex chromosomes as example, 339

Reaney, D., 238

Recombination. *See also* Anti-recombination;

- Gene conversion; Homology search
- adaptive value of, 185, 239
- benefits of lost, 338

- between X and Y chromosomes, 337
- correlation with base composition, 138
- cut first model for, 180
- as destroyer, 185, 194, 201, 354
- of DNA segments, 34
- gene as a unit of, 200
- genome as a unit of, 200
- at **GT**-rich CHI sequences, 273, 300
- at **GT**-rich class-switch sites, 273
- homostability regions support, 193
- hotspot, 152, 202, 273
- illegitimate, 244
- intergenomic, 193
- intragenic, 267
- intragenomic, 193, 194
- less when selection is positive, 128
- pair first model for, 180
- recombinator sequence for, 210, 273
- repair by means of, 239, 329, 330, 333
- repair, as last court of appeal, 334
- significance for evolution, 34, 195
- somatic, 273
- trade-off with length, 258, 262
- transience of PRDM9 sites, 274

Reductionism, xxxii, 28, 58, 370

Redundancy

- in brain, 386
- in codons, 123, 130, 211
- elimination, 136
- four-fold, 35
- levels of, xxviii
- in mitochondria, 116
- in-parallel, 26, 329, 331
- in-series, 35, 329, 341
- needed for error-detection, 48, 328
- two-fold, 35, 328

Reese, V., 332

Regeneration

- of archival nucleus, 285
- of lost cells, 271, 312
- of phenotype from genotype, 6
- of plant from cuttings, 6
- by recycling, 33

Regression to the mean, 50, 146

Repair. *See* DNA, repair ofRepeats, 243. *See also* Low complexity segment

- copy number of, 96
- direct, 244
- distinct location in exons, 244
- inverted (*see* Palindrome)
- length of, 96
- mirror, 244, 435
- opposing stem-loop potential, 244

- satellite, 271, 273
- short tandem, 269
- and speciation, 273
- supporting stem-loop potential, 244
- Repetitive element
 - Alu*, 289, 355, 360, 369
 - discovery of, 36, 289
 - editing of, 369
 - fewer in birds, 38
 - near *GOS2* gene, 290
 - as variation-generating device, 298
- Replication
 - bidirectional, 108
 - continuous or discontinuous, 111
 - differential timing of, 332
 - fork multiplicity, 300
 - origin of, 107
 - termination of, 109
 - in a Y-shaped fork, 31, 108, 111
- Reproduction, 335
 - asexual, 8, 335
 - as a cyclic process, 147, 336
 - of habits, 8
 - as revealing memory, 7
 - sexual (*see* Sexual reproduction)
- Reproductive isolation
 - allopatric, 151
 - between camel and pig, 150
 - by chromosomal incompatibility, 182
 - as conserver of adaptations, 177
 - defined, 47, 146
 - by developmental barrier, 150
 - empowers natural selection, 179, 189
 - first manifest in offspring, 164
 - by hybrid sterility, 151
 - immediate, 187
 - as origin of species, xxix, 150, 164
 - partner-specific, 158, 179
 - as potential species multiplier, 223
 - as precondition for divergence, 176
 - role of repetitive sequences in, 273
 - role of secondary information, 46
 - secondary mechanisms of, 341
 - between strains of *Tetrahymena*, 176
 - sympatric, 151
 - by transmission barrier, 150
- Reprotype, 47, 56, 170, 171, 181, 182, 335, 435
- Research
 - decoys in, 85, 285, 289, 295, 320, 401
 - marketing of, xxiv, 399, 403, 407
 - priorities in, 3, 4, 157, 371, 403
 - refocusing, 367, 371, 400
 - residue of unexplained phenomena, 387, 397
- Retrovirus. *See also* AIDS virus; Human T-cell leukemia virus
 - ancient remnant in genome, 298
 - use in genetic engineering, xxv
 - inactivation by methylation, 354
 - mutational meltdown in, 184
 - phylogenetic analysis of, 184
- Revolution
 - genetic, 137
 - in ideas, 4
 - new bioinformatics as, xxxii
- Rhythm
 - communication of, 10
 - periodical, 9–11
 - recurrence of, 11
- Ribosomes, 85, 114, 187, 204
- RNA, 112
 - antibody, also expressing primary function, 291
 - antibody, innate repertoire of, 286
 - versus DNA, 245
 - double-stranded (*see* dsRNA)
 - editing of, 95, 297, 360
 - hidden transcriptome, xxx, 289, 297, 298, 312, 355, 360
 - immune receptors, xxx, 285, 294, 297
 - induced silencing complex (RISC), 293
 - interference (RNAi), 294
 - low copy number types, 289
 - molecules driving on same side, 115
 - mRNA, 85, 86, 235
 - mRNA heterogeneity, 285
 - mRNA, prediction of existence, 401
 - mRNA structure, 86, 94
 - mRNA-tRNA interaction, 84, 114, 187
 - non-protein-encoding, 106, 237, 245, 344, 354
 - piRNAs, 359
 - purine-loading of (*see* Purine-loading)
 - rRNA as most abundant RNA type, 85, 112, 401
 - rRNA has high (G+C) in thermophiles, 204
 - rRNA precursor with spacer, 235
 - short interfering (siRNA), 293
 - test-tube synthesis of, 35
 - tRNA, 84, 187
 - unscheduled expression of, 292
 - virus, evolution of, 133
 - world, 239
- RNA polymerase
 - copies template strand, 236
 - cryptic promoters for, 298
 - dedicated, 289
 - direction of movement of, 104
 - promoter for, 107, 290, 354

RNY rule

- as basis for periodicity, 213
- chemical basis of, 126
- as correction guide, 335
- to detect protein coding, 126
- determines reading-frame, 245
- and overlapping genes, 246
- pattern in both DNA strands, 126
- and **RNY**-pressure, 127, 226, 257
- violated by **AG**-pressure, 127, 225, 265

Romanes, G.

- alliance with Sanderson, 396
- cited Mendel, 143
- conflict with Butler, 8, 17, 57, 400, 406
- on female brain size, 388
- and a higher subjectivity, 390
- his collective variation postulate, 159, 169
- his fruitless pangenesis studies, 20
- and holy grail, 176
- on hybrid sterility, 158
- on instinct, 8
- and intrinsic peculiarity, 176
- known by Eliot, 399
- on pangenesis, 14
- and panpsychism, 390
- parallels with Gould of, 398
- photo of, 158
- on physiological selection, 56, 158
- as pioneer, xxiv, 160, 396
- and the term "ultra-Darwinian", 406
- and theory of mind, 388
- on transition of germinal substance, 16
- on Weismannism, 14, 142
- and world eject, 389

Rosetta Stone, 4**Rouleau**

- formed in polymerized albumin, 306
- like a pile of coins, 29, 92, 306
- like stacked DNA bases, 29, 92
- specificity of formation of, 306

Rushdie, S., 425, 428, 432**Russian doll model, 10, 14****S**

- Salser, W., 86
- Sanderson, J.B., xxiii, xxxiii, 396
- Sanger, F., xxiv
- Saunders, E., 146, 161
- Schaap, T., 210
- Schrödinger, E., 16, 22
- Science, 376
 - Arabic, 376
 - and the arts, two cultures, 370

- authority in, 201, 385, 427
- citation of, 428
- confusion in, xxxi
- diffidence in, 320, 428
- fiction, 376, 430
- hazards of technology, 3
- history of, 395, 404
- literature, 395, 398, 404
- logic of, 383, 387
- nature of observation in, 421, 427, 428
- nest-making in, 427
- politics of, xxxi, 4, 385, 395, 407
- and religion, 376, 425
- research in (*see* Research)
- rules of thumb of, 427
- scope of, 425, 427
- specialization in, xxviii

Second parity rule (PR2). *See* Chargaff, E.**Selection, 127, 160, 290**

- artificial, 137
- frequency-dependent, 66
- genic for survival within species, 137
- of group, 151, 171, 290
- hierarchical theory of, 398
- of kin (*see* Altruism)
- level of operation, 77, 78, 86, 223, 245, 255, 290
- negative (*see* Negative selection)
- by non-classical factors, 130
- non-genic, 137
- physiological (*see* Physiological selection)
- positive (*see* Positive selection)
- and preservation, xxiv, xxx
- as a "sweep", 129
- with wisdom, 405

Self. *See also* Near-self; Not-self

- aggregation with self, 306, 308
- defined afresh each generation, 283
- discrimination from near-self, 39, 310
- discrimination from not-self, 39, 279, 299, 310, 354
- discrimination, intracellular, 282, 284, 285, 293, 307, 343
- marker in DNA, 182, 353

Selfish gene

- as autonomous entity, 132
- concept of, 200
- conflict with selfish genome by, xxx, 223
- and contraception, 317
- explains altruism, 285, 290
- explains junk, 289
- explains repeats, 273
- and gametic imprinting, 359
- Huxley on, 6, 431

- and the menopause, 316
- as unit of self-preservation, xxiii, xxx
- Semon, R., 16, 22
- Sequence, 169, 246, 258, 341
 - alignment analysis of, 255
 - ancient, 57
 - complexity of (*see* Complexity)
 - crossover hotspot instigator (CHI) in, 273
 - decomposition into dinucleotides, 92
 - divergence (*see* Divergence)
 - microsatellite, 272, 273
 - periodicity in, 213, 269, 274
 - redundancy (*see* Redundancy)
 - simple (*see* Repeats)
 - simple defined, 246
 - spacer, 235
 - spurious alignments of, 255
 - statistics of, 409
 - tagging as suspect, 332, 359
- Sex, 337
 - chromosomes (*see* X- and Y-chromosomes)
 - differentiation like species
 - differentiation, 339
 - early origin of, 239
 - heterogametic, 336, 337, 341
 - homogametic, 337
 - molecular, 35
 - in protozoa, 285
 - sterility affects the heterogametic first, 336
 - two in a species, 66, 327, 342
- Sexual reproduction
 - ancient origin of, 239
 - biparental, 8
 - as a fragmenter of genomes, 201
 - leads to homology search, 280
 - permits sequential correction, 332
 - reason for, 47, 327, 335
- Shakespeare, W., 49, 51, 137, 148
- Sharp, P., 131
- Shaw, G.B., 47, 401, 407
- Sheep, 44, 141, 142, 279, 314
- Shuffling
 - to destroy base order, 88
 - to destroy letter order, 44
 - sequences, 73, 86, 95, 128
- Simon, J., 13
- Sinclair, R., 80
- Sinden, R., 29
- Skew analysis. *See also* Chargaff difference analysis
 - at origins of replication, 109
 - at recombination hotspots, 274
- Skin
 - cell death, 316
 - cells and radiation, 21
 - as perimeter, 285
 - pigmentation, 21, 287, 313
- Skull
 - capacity to store, 371
 - fusing of bones of, 383
 - open to check contents of, 382
- Smith, A., 405
- Smithies, O., 104
- Smith, J.M., 398
- Soma
 - as disposable, 148, 334
 - as distinct from germ-line, 13, 160, 285
 - limited flexibility of, 55
- Space
 - competition for, 64, 258
 - conflict with redundancy, 26, 34
 - dimensions of, 390
 - extracellular within skin, 285
 - within genome, xxvi, xxx, 64, 209, 238, 283
 - between protein domains, 230
- Spandrel. *See* Water-dipper
- Speciation
 - driven by accent change, 47
 - driven by synonymous mutations, 135
 - gene, 152, 224
 - holy grail of, 176
 - need for prior isolation, 136, 169
 - non-genic, xxix, 151
 - and repetitive sequences, 273
 - uncoupled from adaptation, 137
- Species
 - allied, 47, 52, 162, 164, 185, 281
 - arrival, 137, 142, 147, 170
 - barcode for, 215
 - definition of, 138, 171
 - discontinuity of, 52
 - discreteness of, 52
 - foreign, 281
 - versus genes, 217, 219, 222
 - incipient, 150, 171, 179
 - selection as group selection, 171, 218, 221, 223
 - sibling, 45
 - specific codon usage, 211
 - specificity of k-mer values in, 74, 136, 422
 - as a unit of antirecombination, 200
- Spencer, H., 8, 12, 19, 20
- Spermatozoon, 12, 144
 - compactness of, 34
 - seen as self, 280
 - two types in human males, 336
- Stacking of bases, 31, 92, 96

- Stem-loop extrusion
 - affected by base composition, 181
 - affected by base order, 181
 - asymmetrical, 98, 272
 - from DNA, 87, 180
 - favored by negative supercoiling, 86, 111, 180
 - favored entropically, 296
 - generates mismatches, 334
 - symmetrical, 97, 180
- Stem-loop potential
 - base composition-dependent, 182
 - base order-dependent, 240
 - calculation of, 88–90
 - conserved in AIDS virus, 243
 - conserved in intron, 241
 - of EBNA1 simple sequence, 262
 - factors affecting, 182
 - as flap in single strand, 68
 - widely distributed, 87
- Step Across the Line* by Rushdie, 425
- Steps to an Ecology of Mind* by Bateson, 54
- Sterility. *See also* Hybrid sterility
 - cryptic factors for, 164
 - factors fail to complement, 166
 - genic, 151, 162
 - non-genic, 151, 161
 - sporadic, 162
- Stone
 - positions of stability of, 53
 - in species wall, 171
 - which to turn?, xxviii
 - as a word, 9
- Storage
 - extra-corporeal, 372
 - holographic, 374
 - of information, 18, 371
- Strand
 - antiparallel in DNA, 31
 - asymmetry in folding energy, 271
 - breakage at genic boundary, 202
 - complementarity of, 168
 - displacement as loop, 197
 - guidance for error detection, 331, 334
 - identity in density, 271
 - inward-looking, 179
 - lagging in replication, 108, 111
 - leading in replication, 108, 111
 - migration of cross-over point, 196
 - mRNA synonymous, 104
 - mRNA template, 104
 - non-identity in density, 271
 - paranemic pairing by, 196
 - polarity of, 67
 - recorded in data bases, 67
 - RNY** pattern in both, 126
 - separation and annealing, 36
 - symmetry in folding energy, 97, 271
 - symmetry principle, 70
 - unpairing, 179
 - winding, 86
- Strategy
 - for accurate information transfer, 26
 - for detecting forgery, 39
 - for detecting not-self, 39, 356
 - evolutionary, 49
 - of host versus pathogen, 283, 317, 320
 - of predator, xxx
 - for purine-loading, 257
 - of recycling, 33
 - of repair, 33
 - of species-specific coding, 211
 - of virus, to evade defenses, 184, 293, 295, 299, 313, 320, 353
- The Strength of Weak Ties* by Granovetter, 284
- Stress
 - at a critical developmental stage, 319
 - and the heat-shock response, 317
 - increased transcription in, 291, 314
 - need to control for, 290
- Structure
 - calculation of secondary, 88
 - and conflict, 94, 230, 245
 - domains of, 230
 - of information, 375
 - mediated homology recognition, 179
 - needs of may dominate, 86, 95, 244
 - of neural synapses, 370
 - of oncoprotein mRNA, 94
 - precision in a protein, 215
 - primary and secondary, 64, 83
 - validation of, 93
- The Structure of Evolutionary Thought* by Gould, 398
- Subak-Sharpe, J.H., 76
- Sueoka, N., 78, 175, 176, 188
- Supek, F., 188, 231
- Survival
 - of the fittest, 19, 85, 127, 289
 - of incipient species, 171
- Suyama, A., 193, 198, 222
- Sympatry, 46, 52, 398
- Syndrome
 - Down, 345
 - multiple personality, 387, 391
 - savant, 382, 385

- Szybalski, W.
 and base clustering, 104
 on replication origin, 109
 transcription direction rule of, 112, 268, 290
- T**
- Task
 of biology, xxv
 of EB, xxvi, 3, 22, 209
- Tekke Turcomans, 9, 11
- Teleology, 19, 20
- Tellem, J., 297
- Temperature
 as aggregation promotor, 308
 change at ovulation, 316
 correlation with purine-loading, 204
 for DNA degradation, 203
 of duplex melting, 29, 114, 296
 favors entropic reactions, 296
 increase (pyrexia), 296
 of optimum growth, 204, 231
- Template
 concept for copying, 29
 dependent reactions, 17
- Tennyson, A., 141
- Tetraploidy
 in cell cycle G2-phase, 331
 as cure for hybrid sterility, 168, 187, 336
 in tobacco, 186
- Text
 ancient, 51, 57
 that charms, xxiii
 of a DNA strand, 30
 need for complete, xxv
 and pretext, 426
- Theophrastus, as first botanist, 406
- Thermophiles, 96, 105, 203, 230
- Thiselton-Dyer, W, 398
- Thymus. *See also* Autoimmune disease
 AIRE protein located in, 292, 345
 T-cell education in, 315
- Tiger, 170, 171
- Tinoco, I., 88
- Tomizawa, J.-I., 113
- Trade-off
 between **A** and **C** preferred, 227, 230
 between codon positions, 228
 between forms of information, xxvi
 between length and recombination vulnerability, 258
 between melanin and Vitamin D, 21
 between menopause and procreation, 317
 not of diploidy, 38
- Transcription
 directional polarity of, 104
 factor acting at promoter, 354, 357
 factor at hotspots (PRDM9), 152
 failure to terminate, 290
 promiscuous, 291
 rate of, 310, 344
 regulation of, 112
 and RNA processing, 236
 same direction as replication, 109
 unidirectional when genes coorientated, 115, 245
 unscheduled, 291
- Translation. *See also* Codon
 efficiency of, 213
 of germinal into somatic, 16
 polarity of, 107
 as rate-limiting, 187
- Transposition
 of chromosome segments, 110, 169, 244
 of gene, 110, 197
- Treffert, D.A., 371
- Trinucleotide
 expansion mechanism, 269
 frequencies of, 73
 homology recognition unit, 188
- U**
- Unconscious Memory* by Butler, 9, 400, 437
- Universe
 as a computer, xxvi
 dark matter in, 374
 holographic, 374, 377
 unity of, 392
 unseen, 367
- V**
- Vaccinia virus
 base composition of, 66
 dinucleotides in, 70
 trinucleotides in, 70
- Variation
 as a character difference, 121
 collective, 157, 159, 169
 continuity of, 159, 177, 189
 cryptic, 159, 160, 176, 189, 340
 decreased by positive selection, 129
 discontinuity of, 53, 160, 177, 189
 discrete, 18

Variation (*cont.*)

- due to nurture, 50
- fine gradations in, 18
- as a fundamental principle, 6
- homeotic, 121
- intrinsic versus extrinsic origin of, 159, 162
- Jenkin's sphere of, 49
- limits of its scope, 18, 50, 171
- meristic, 121
- spontaneous, 18, 137, 159
- sporadic, 159, 169
- statistical analysis of, 50, 409
- substantive, 121
- summation of, 166
- unbiased, 19
- when storage imperfect, 18
- wider for base composition, 188

Vegetative propagation, 6

Venditti, C., 136

Vibration. *See also* Periodical rhythms

- change as marker, 310
- between conformations, 87, 114, 183
- copying of, 10
- of gene, 200, 334
- as memory, 11, 369
- molecular, 296, 306
- as molecular resonance, 43

Victorians, xxvii, xxviii, xxxii, 5, 6, 18, 47, 58, 400, 404, 437

Virus. *See also* AIDS virus; EBV

- anticipates host environment, 286, 298, 313
- cattle plague, 13
- chemokine response to, 293
- coat protein of, 85, 286, 307
- coinfection by, 185
- cowpox, 66
- evolution of, 133, 185
- as first named, 13
- hepatitis C, 243
- infection as stress, 291
- infection during pregnancy, 291, 312
- interferon response to, 293
- latent, 295
- origin of, 285
- rubella, 291
- smallpox, 13, 66
- strategy of, 286, 293, 299, 311, 313, 320, 353
- superinfection, 185
- tobacco mosaic (TMV), 307

Vitamin D, 21

Vitiligo, 313

Voltaire, 405

W

Wada, A., xxx, 193, 198, 218, 222, 334

Wahls, W., 273

Wallace, A.R., 20, 142, 170, 172, 340, 367, 373, 375, 387, 398, 406

Wan, H., 255

Water

- dipper analogy, 360
- ordering of molecules of, 92, 305, 308

Watson, J., xxiii, 27, 399

Wax model, 14

Weaver, I., 357

Weinberg, W., 347

Weiner, B.M., 180

Weismann, A.

- Butler's disagreement with, 19
- as Darwinist, 142, 406
- and the germ-line, 13
- and theory of germ-plasm, 14, 284
- and translation of germ, 16

Weldon, R., 160

Wheeler, J., 377

White, H., 86

White, M., 397

Whitehead, A.N., 397

Williams, G.C.

- and definition of gene, xxiii, xxx, 34, 200, 202
- on recombination and sex, 239
- and selfish genes, 132

Window

- counting bases in, 44, 182
- moving, 105
- as segment in a sequence, 44, 199
- size optimum, 105
- of time, 432

Winge, Ö., 168, 187

Wittgenstein, L., 382, 388

Wolfe, K., 131

Wolfe, V., xxxii

Wootton, J., 255

Word

- coining of, 5, 157, 288, 308, 396
- difference in active and passive code, 210
- as a handle, 6, 288, 375
- after idea, xxvii, 5, 15
- k-mer frequencies of, 75
- meaning of, 5, 295, 351, 353
- mind, associations of, xxxi
- misprint of, 154
- religious implication of, 394
- rhyming, 391
- stone as a, 9, 375
- structure of, 27

Worm. *See also* Nematode worm
marine, 238
Wyatt, G., 27, 399

X

X-chromosome
extra product dose in females, 342
heteromorphy with Y, 339
homomorphy with Y, 337
inactivation by methylation, 354
meiotic pairing of, 281
mutations affecting dosage
compensation, 343
in ovum, 66
pseudoautosomal region of, 340
in spermatozoon, 66, 336

in successive male and female cells, 343
transposition with Y changes (G+C), 197

Y

Y-chromosome
confers maleness, 336
degeneration of, 342
meiotic pairing of, 281
pseudoautosomal region of, 197, 340
in spermatozoon, 66
in a succession of male cells, 342
Yomo, T., 76

Z

Zuker, M., 88