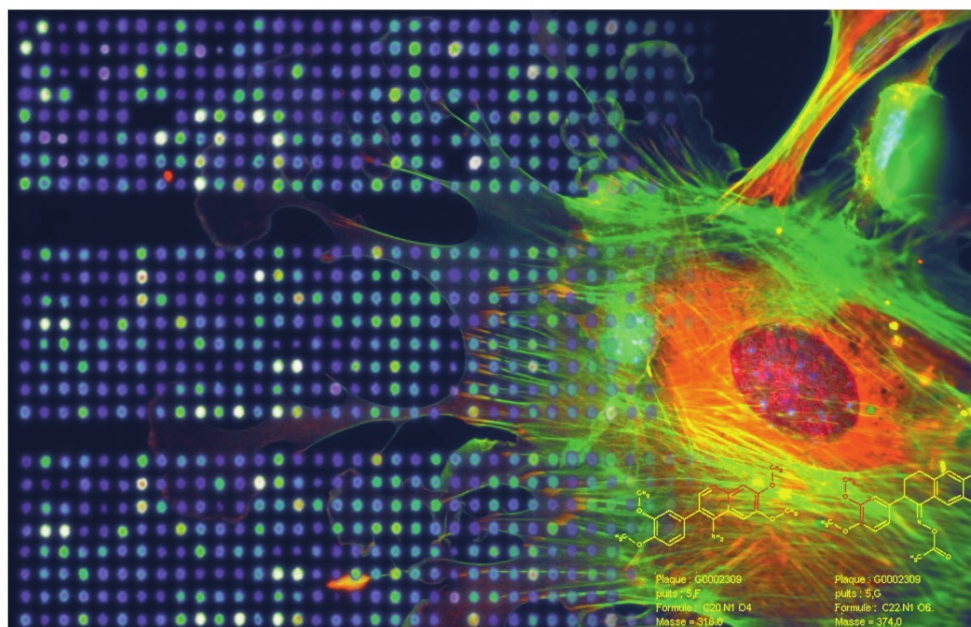




— SELECTED BY GRENOBLE SCIENCES —

Chemogenomics and Chemical Genetics



Chemogenomics and Chemical Genetics

Grenoble Sciences

The aims of Grenoble Sciences are double:

- ▶ to produce works corresponding to a clearly defined project, without the constraints of trends or programme,
- ▶ to ensure the utmost scientific and pedagogic quality of the selected works: each project is selected by Grenoble Sciences with the help of anonymous referees. Next, the authors work for a year (on average) with the members of an interactive reading committee, whose names figure in the front pages of the work, which is then co-published with the most suitable publishing partner.

Contact: Tel.: (33) 4 76 51 46 95 - E-mail: Grenoble.Sciences@ujf-grenoble.fr
website: <http://grenoble-sciences.ujf-grenoble.fr>

Scientific Director of Grenoble Sciences: **Jean BORNAREL**,
Emeritus Professor at Joseph Fourier University, Grenoble, France

Grenoble Sciences is a department of Joseph Fourier University, supported by
the **French National Ministry for Higher Education and Research**
and the **Rhône-Alpes Region**.

Chemogenomics and Chemical Genetics is an improved version of the original
book ***Chemogénomique - Des petites molécules pour explorer le vivant***
sous la direction de Eric MARÉCHAL, Sylvaine ROY et Laurence LAFANECHÈRE,
EDP Sciences - Collection Grenoble Sciences, 2007, ISBN 978 2 7598 0005 6.

The Reading Committee of the French version included the following members:

- ▶ **Jean DUCHAINE**, Principal Advisor of the Screening Platform, Institute for Research in Immunology and Cancer, University of Montreal, Canada
- ▶ **Yann GAUDUEL**, Director of Research at INSERM, Laboratory of Applied Optics (CNRS), *Ecole Polytechnique*, Palaiseau, France
- ▶ **Nicole MOREAU**, Professor at the *Ecole Nationale Supérieure de Chimie*, Pierre and Marie Curie University, Paris, France
- ▶ **Christophe RIBUOT**, Professor of Pharmacology at the Faculty of Pharmacy, Joseph Fourier University, Grenoble, France

*Translation performed by **Philip SIMISTER***

*Typesetted by **Centre technique Grenoble Sciences***

*Cover illustration: **Alice GIRAUD***

(with extracts from a DNA microarray image - *Biochip Laboratory/Life Sciences Division/CEA* - and a photograph of actin filaments array and adhesion plates in a mouse embryonic cell - *Yasmina SAOUDI, INSERM U836 Grenoble, France*)

Eric Maréchal • Sylvaine Roy • Laurence Lafanechère
Editors

Chemogenomics and Chemical Genetics

A User's Introduction for Biologists,
Chemists and Informaticians



Springer

Editors

Dr. Eric Maréchal
Laboratory of Plant Cell Physiology
UMR 5168, CNRS-CEA-INRA-
Joseph Fourier University
Rue des Martyrs 17
38054 Grenoble Cedex 9
France
eric.marechal@cea.fr

Sylvaine Roy
Laboratory of Plant Cell Physiology
UMR 5168, CNRS-CEA-INRA-
Joseph Fourier University
Rue des Martyrs 17
38054 Grenoble Cedex 9
France
sylvaine.roy@cea.fr

Laurence Lafanechère
Albert Bonniot Institute
Department of Cellular Differentiation and Transformation
Rond-point de la Chantourne
38706 La Tronche Cedex
France
laurence.lafanechere@ujf-grenoble.fr

Translator:

Philip Simister
Weatherall Institute of Molecular Medicine
University of Oxford
Oxford OX3 9DS, UK

Originally published in French: *Chemogénomique - Des petites molécules pour explorer le vivant* sous la direction de Eric MARÉCHAL, Sylvaine ROY et Laurence LAFANECHÈRE, EDP Sciences - Collection Grenoble Sciences, 2007, ISBN 978 27598 0005 6.

ISBN 978-3-642-19614-0 e-ISBN 978-3-642-19615-7
DOI 10.1007/978-3-642-19615-7
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930786

© Springer-Verlag Berlin Heidelberg 2011

This work is subject to copyright. All rights are reserved, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilm or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the German Copyright Law of September 9, 1965, in its current version, and permission for use must always be obtained from Springer. Violations are liable to prosecution under the German Copyright Law.

The use of general descriptive names, registered names, trademarks, 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.

Cover illustration: Alice Giraud

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

CONTENTS

Preface.....	1
Introduction.....	3

FIRST PART **AUTOMATED PHARMACOLOGICAL SCREENING**

Chapter 1 - The pharmacological screening process: the small molecule, the biological screen, the robot, the signal and the information	7
<i>Eric MARÉCHAL - Sylvaine ROY - Laurence LAFANECHÈRE</i>	
1.1. Introduction	7
1.2. The screening process: technological outline.....	8
1.2.1. Multi-well plates, robots and detectors	8
1.2.2. Consumables, copies of chemical libraries and storage	10
1.2.3. Test design, primary screening, hit-picking, secondary screening	10
1.3. The small molecule: overview of the different types of chemical library	12
1.3.1. The small molecule	12
1.3.2. DMSO, the solvent for chemical libraries.....	12
1.3.3. Collections of natural substances	12
1.3.4. Commercial and academic chemical libraries.....	14
1.4. The target, an ontology to be constructed	14
1.4.1. The definition of a target depends on that of a bioactivity.....	14
1.4.2. Duality of the target: molecular entity and biological function	15
1.4.3. An ontology to be constructed	16
1.5. Controls	17
1.6. A new discipline at the interface of biology, chemistry and informatics: chemogenomics	17
1.7. Conclusion.....	18
1.8. References	19

Chapter 2 - Collections of molecules for screening: example of the french national chemical library	23
<i>Marcel HIBERT</i>	
2.1. Introduction	23
2.2. Where are the molecules to be found?	25
2.3. State of progress with the European Chemical Library	27
2.4. Perspectives	28
2.5. References	28

Chapter 3 - The miniaturised biological assay: constraints and limitations 29*Martine KNIBIEHLER*

3.1. Introduction	29
3.2. General procedure for the design and validation of an assay	30
3.2.1. Choice of assay	31
3.2.2. Setting up the assay	33
3.2.3. Validation of the assay and automation	35
3.3. The classic detection methods	36
3.4. The results	36
3.4.1. The signal measured: increase or decrease?	36
3.4.2. The information from screening is managed on three levels	37
3.4.3. Pharmacological validation	40
3.5. Discussion and conclusion	40
3.6. References	41

Chapter 4 - The signal: statistical aspects, normalisation, elementary analysis 43*Samuel WIECZOREK*

4.1. Introduction	43
4.2. Normalisation of the signals based on controls	44
4.2.1. Normalisation by the percentage inhibition	44
4.2.2. Normalisation resolution	44
4.2.3. Aberrant values	46
4.3. Detection and correction of measurement errors	48
4.4. Automatic identification of potential artefacts	49
4.4.1. Singularities	49
4.4.2. Automatic detection of potential artefacts	50
4.5. Conclusion	52
4.6. References	52

Chapter 5 - Measuring bioactivity: K_i , IC_{50} and EC_{50} 55*Eric MARÉCHAL*

5.1. Introduction	55
5.2. Prerequisite for assaying the possible bioactivity of a molecule: the target must be a limiting factor	55
5.3. Assaying the action of an inhibitor on an enzyme under Michaelian conditions: K_i	56
5.3.1. An enzyme is a biological catalyst	57
5.3.2. Enzymatic catalysis is reversible	57
5.3.3. The initial rate, a means to characterise a reaction	59
5.3.4. Michaelian conditions	59
5.3.5. The significance of K_m and V_{max} in qualifying the function of an enzyme	60
5.3.6. The inhibited enzyme: K_i	60
5.4. Assaying the action of a competitive inhibitor upon a receptor: IC_{50}	62
5.5. Relationship between K_i and IC_{50} : the CHENG-PRUSOFF equation	63
5.6. EC_{50} : a generalisation for all molecules generating a biological effect (bioactivity)	64
5.7. Conclusion	64
5.8. References	65

**Chapter 6 - Modelling the pharmacological screening:
controlling the processes and the chemical, biological and experimental information..... 67***Sylvaine ROY*

6.1. Introduction	67
6.2. Needs analysis by modelling.....	68
6.3. Capture of the needs	69
6.4. Definition of the needs and necessity of a vocabulary common to biologists, chemists and informaticians	69
6.5. Specification of the needs	69
6.5.1. Use cases and their diagrams	70
6.5.2. Activity diagrams	72
6.5.3. Class diagrams and the domain model.....	73
6.6. Conclusion.....	78
6.7. References	78

Chapter 7 - Quality procedures in automated screening..... 79*Caroline BARETTE*

7.1. Introduction	79
7.2. The challenges of quality procedures.....	79
7.3. A reference guide: the ISO 9001 Standard.....	80
7.4. Quality procedures in five steps	82
7.4.1. Assessment	82
7.4.2. Action plan - planning.....	83
7.4.3. Preparation	83
7.4.4. Implementation.....	83
7.4.5. Monitoring.....	83
7.5. Conclusion.....	84
7.6. References	84

SECOND PART**HIGH-CONTENT SCREENING AND STRATEGIES IN CHEMICAL GENETICS****Chapter 8 - Phenotypic screening with cells and forward chemical genetics strategies..... 87***Laurence LAFANECHÈRE*

8.1. Introduction	87
8.2. The traditional genetics approach: from phenotype to gene and from gene to phenotype.....	88
8.2.1. Phenotype	88
8.2.2. Forward and reverse genetics	89
8.3. Chemical genetics	89
8.4. Chemical libraries for chemical genetics	90
8.4.1. Chemical library size.....	91
8.4.2. Concentration of molecules.....	91
8.4.3. Chemical structure diversity.....	91
8.4.4. Complexity of molecules	93

8.4.5. Accessibility of molecules to cellular compartments	93
8.4.6. The abundance of molecules	94
8.4.7. The possibility of functionalizing the molecules	94
8.5. Phenotypic tests with cells	94
8.6. Methods to identify the target	96
8.7. Conclusions	99
8.8. References	99

Chapter 9 - High-content screening in forward (phenotypic screening with organisms) and reverse (structural screening by NMR) chemical genetics 103

Benoît DÉPREZ

9.1. Introduction	103
9.2. Benefits of high-content screening.....	104
9.2.1. Summarised comparison of high-throughput screening and high-content screening	104
9.2.2. Advantages of high-content screening for the discovery of novel therapeutic targets	104
9.2.3. The nematode <i>Caenorhabditis elegans</i> : a model organism for high-content screening.....	105
9.2.4. Advantages of high-content screening for reverse chemical genetics and the discovery of novel bioactive molecules	108
9.3. Constraints linked to throughput and to the large numbers	110
9.3.1. Know-how	110
9.3.2. Miniaturisation, rate and robustness of the Assays	110
9.3.3. Number, concentration and physicochemical properties of small molecules ..	111
9.4. Types of measurement for high-content screening	111
9.4.1. The critical information needed for screening	111
9.4.2. Raw, numerical results	111
9.4.3. Results arising from expert analyses	112
9.5. Conclusion.....	112
9.6. References	112

Chapter 10 - Some principles of Diversity-Oriented Synthesis..... 113

Yung-Sing WONG

10.1. Introduction	113
10.2. Portrait of the small molecule in DOS	114
10.3. Definition of the degree of diversity (DD).....	116
10.3.1. Degree of diversity of the building block	116
10.3.2. Degree of stereochemical diversity	118
10.3.3. Degree of regiochemical diversity	119
10.3.4. Degree of skeletal diversity	121
10.4. Divergent multi-step DOS by combining elements of diversity	124
10.5. Convergent DOS: condensation between distinct small molecules	127
10.6. Conclusion.....	130
10.7. References	130

THIRD PART

TOWARDS AN IN SILICO EXPLORATION OF CHEMICAL AND BIOLOGICAL SPACES

Chapter 11 - Molecular descriptors and similarity indices..... 135*Samia ACI*

11.1. Introduction.....	135
11.2. Chemical formulae and computational representation.....	136
11.2.1. The chemical formula: a representation in several dimensions	137
11.2.2. Molecular information content.....	137
11.2.3. Molecular graph and connectivity matrix	139
11.3. Molecular descriptors.....	140
11.3.1. 1D descriptors	141
11.3.2. 2D descriptors	141
11.3.3. 3D descriptors	144
11.3.4. 3D versus 2D descriptors?	146
11.4. Molecular similarity	147
11.4.1. A brief history	147
11.4.2. Properties of similarity coefficients and distance indices	147
11.4.3. A few similarity coefficients.....	148
11.5. Conclusion.....	148
11.6. References	150

Chapter 12 - Molecular lipophilicity: a predominant descriptor for QSAR..... 153*Gérard GRASSY - Alain CHAVANIEU*

12.1. Introduction.....	153
12.2. History.....	153
12.3. Theoretical foundations and principles of the relationship between the structure of a small molecule and its bioactivity	154
12.3.1. QSAR, QPAR and QSPR.....	154
12.3.2. Basic equation of a QSAR study.....	155
12.4. Generalities about lipophilicity descriptors	155
12.4.1. Solubility in water and in lipid phases: conditions for bioavailability.....	155
12.4.2. Partition coefficients	156
12.4.3. The partition coefficient is linked to the chemical potential.....	156
12.4.4. Thermodynamic aspects of lipophilicity	157
12.5. Measurement and estimation of the octanol/water partition coefficient.....	158
12.5.1. Measurement methods	158
12.5.2. Prediction methods.....	160
12.5.3. Relationship between lipophilicity and solvation energy: LSER	163
12.5.4. Indirect estimation of partition coefficients from values correlated with molecular lipophilicity.....	163
12.5.5. Three-dimensional approach to lipophilicity	165
12.6. Solvent systems other than octanol/water.....	166
12.7. Electronic parameters.....	167
12.7.1. The HAMMETT parameter, σ	167
12.7.2. SWAIN and LUPTON parameters	168

12.8. Steric descriptors	169
12.9. Conclusion.....	169
12.10. References	169

Chapter 13 - Annotation and classification of chemical space in chemogenomics 171

Dragos HORVATH

13.1. Introduction.....	171
13.2. From the medicinal chemist's intuition to a formal treatment of structural information	171
13.3. Mapping structural space: predictive models.....	174
13.3.1. Mapping structural space	174
13.3.2. Neighbourhood (similarity) models	175
13.3.3. Linear and non-linear empirical models	178
13.4. Empirical filtering of drug candidates.....	180
13.5. Conclusion.....	181
13.6. References	181

Chapter 14 - Annotation and classification of biological space in chemogenomics..... 185

Jordi MESTRES

14.1. Introduction.....	185
14.2. Receptors.....	186
14.2.1. Definitions.....	186
14.2.2. Establishing the 'RC' nomenclature	187
14.2.3. Ion-channel receptors.....	188
14.2.4. G protein-coupled receptors.....	189
14.2.5. Enzyme receptors	190
14.2.6. Nuclear receptors	190
14.3. Enzymes	191
14.3.1. Definitions.....	191
14.3.2. The 'EC' nomenclature	192
14.3.3. Specialised nomenclature.....	193
14.4. Conclusion.....	193
14.5. References	194

Chapter 15 - Machine learning and screening data..... 197

Gilles BISSON

15.1. Introduction.....	197
15.2. Machine learning and screening.....	199
15.3. Steps in the machine-learning process	202
15.3.1. Representation languages.....	203
15.3.2. Developing a training set	205
15.3.3. Model building.....	206
15.3.4. Validation and revision	207
15.4. Conclusion.....	209
15.5. References and internet sites	209

Chapter 16 - Virtual screening by molecular docking.....	213
<i>Didier ROGNAN</i>	
16.1. Introduction	213
16.2. The 3 steps in virtual screening.....	213
16.2.1. Preparation of a chemical library	213
16.2.2. Screening by high-throughput docking.....	216
16.2.3. Post-processing of the data.....	218
16.3. Some successes with virtual screening by docking.....	220
16.4. Conclusion.....	221
16.5. References	222

APPENDIX

BRIDGING PAST AND FUTURE?

Chapter 17 - Biodiversity as a source of small molecules for pharmacological screening: libraries of plant extracts.....	227
<i>Françoise GUERITTE, Thierry SEVENET, Marc LITAUDON, Vincent DUMONTET</i>	
17.1. Introduction	227
17.2. Plant biodiversity and North-South co-development	229
17.3. Plant collection: guidelines	230
17.4. Development of a natural-extract library	231
17.4.1. From the plant to the plate	231
17.4.2. Management of the extract library	231
17.5. Strategy for fractionation, evaluation and dereplication	233
17.5.1. Fractionation and dereplication process.....	233
17.5.2. Screening for bioactivities.....	235
17.5.3. Some results obtained with specific targets	236
17.5.4. Potential and limitations.....	239
17.6. Conclusion.....	240
17.7. References	240
Glossary	241
The authors	253

PREFACE

Jean CROS

Having completed the reading of this work, one can only feel satisfied for having encouraged Laurence LAFANECHÈRE, Sylvaine ROY and Eric MARÉCHAL, who attempted and succeeded in achieving the impossible: writing, along with colleagues from the public sector, a book that will endure, concerning a technology the mastery of which had remained until this point the domain of the pharmaceutical industry. Indeed, this work has arisen from the competence and practical knowledge of fifteen or so academic scientists who, often against the tide of strategies defined by their host organisations, have established automated pharmacological screening for fundamental research ends. It is important to recall that the first book *High throughput screening*, edited in 1997 by John P. DEVLIN, which enabled all scientists to discover the importance of robotics in the discovery of new medicines, was written by about a hundred contributors, all of whom were industrial scientists involved in drug discovery.

Over the last ten years, we have seen appear in the scientific literature much more about ‘small’ molecules coming from robotic screens that have been used with success in revealing new biological mechanisms. From drug candidate, the molecule has thus become a research tool.

The successful experience at Harvard is a fertile example which should serve as a model for some of our research centres: basic research in chemical genetics, discovery of new drug candidates and training of young researchers. May this book, which has developed out of training workshops organised by the CNRS, CEA and INSERM, be the stimulus for future careers in a field which is eminently multidisciplinary and which brings together biologists, chemists, informaticians and robotics specialists. The great merit of this book is to have simply, from everyday experiences, united researchers and competencies that until now had not associated with one another.

Beyond the new terms that we discover or rediscover throughout the chapters: chemical genetics, cheminformatics, chemogenomics etc., there are the techniques, certainly, but also and above all there are the scientific questions to which these technologies will henceforth help to find answers. In addition, there are the economic issues that from now on become the duty of every researcher to take into account.

Congratulations to all of the authors and editors.

INTRODUCTION

André TARTAR

Over the last two decades, biological research has experienced an unprecedented transformation, which often resulted in the adoption of highly parallel techniques, be it the sequencing of whole genomes, the use of DNA chips or combinatorial chemistry. These approaches, which have in common the repeated use of trial and error in order to extract a few significant events, have only been made possible thanks to the progress in miniaturisation and robotics informatics.

One of the first sectors to put into practice this approach was within pharmaceutical research with the systematic usage of high-throughput screening for the discovery of new therapeutic targets and new drug candidates. Academic research has for a long time remained distanced from this process, as much for financial as for cultural reasons. For several years, however, the trivialisation of these techniques has led to a considerable reduction in the cost of accessing them and has thus permitted academic groups to employ such methods in projects having generally more cognitive objectives.

Nevertheless, it is no less vital, as with all involved methods, to take into account the cost factor as a fundamental parameter in the development of an experimental protocol relative to the expected benefit. The value of a chemical library is in effect an evolving notion resulting from the sum of two values that evolve in opposite directions:

- » On the one hand, the set of physical samples whose value will fatally decrease due both to its consumption in tests, but above all to the degradation of the components. The experience of the last few years also shows that it will be subjected to the effects of fashion, which will contribute rapidly to its obsolescence: no-one today would assemble a chemical library as would have been done only five years ago. Since the great numbers that dominated the first combinatorial chemical libraries, a more realistic series of criteria has progressively been introduced, bearing witness to the difficulties encountered. 'Drugability' has thus become a keyword, with LIPINSKI's rule of 5 and the 'frequent hitters' becoming the *bête noire* of screeners having given them too often cause for hope, albeit unfounded.
- » On the other hand, the mass of information accumulated over the different screening tests is ever increasing and will progressively replace the physical chemical library. With a more or less distant expiry date, the physical chemical

library will have disappeared and the information that it has allowed to accumulate will be all that remains. This information can then be used either directly, constituting the ‘specification sheet’ of a given compound, or as a reference source in virtual screening exercises or *in silico* prediction of the properties of new compounds.

A very simple strategic analysis shows that with the limited means available to academic teams, it is easier to be competitive with respect to the second point (quantity and quality of information) than to the first (number of compounds and high throughput). This also shows that the value of an isolated body of information is much less than that of an array organised in a logical manner based on two main dimensions: the diversity of compounds and the consistency of the biological tests.

It is in this vein that high-content screening should become established, permitting the collection and storage of the maximum amount of data for each experiment. This high-content screening will be the guarantee for the optimal evaluation of physical collections. It is interesting to note that the problem of information loss during a measurement was at the centre of spectroscopists’ preoccupations a few decades ago. In the place of dispersive systems (e.g. prisms, networks) that sequentially selected each observation wavelength but let all others escape, they have substituted non-dispersive analysis techniques entrusting deconvolution algorithms and multi-channel analysers with the task of processing the global information. Biology is undergoing a complete transformation in this respect. Whereas about a decade ago one was satisfied by following the expression of a gene under the effect of a particular stimulus, today, thanks to pan-genomic chips, the expression profile of the whole genome has become accessible. It is imperative that screening follows the same path of evolution: no longer losing any information will become the rule. In the longer term, it will be necessary for this information to be formatted and stored in a lasting and reusable manner.

With this perspective, this book appears at just the right moment since it constitutes a reference tool enabling different specialists to speak the same language, which is essential to ensure the durability of the information accrued.

FIRST PART

AUTOMATED PHARMACOLOGICAL SCREENING

Chapter 1

THE PHARMACOLOGICAL SCREENING PROCESS: THE SMALL MOLECULE, THE BIOLOGICAL SCREEN, THE ROBOT, THE SIGNAL AND THE INFORMATION

Eric MARÉCHAL - Sylvaine ROY - Laurence LAFANECHÈRE

1.1. INTRODUCTION

Pharmacological screening implements various technical and technological means to select from a collection of molecules those which are active towards a biological entity. The ancient or medieval pharmacopeia, in which the therapeutic effects of mineral substances and plant extracts are described, arose from pharmacological screening but whose operative methods are either unknown or very imprecise (see chapter 17). Due to a lack of documentation, one cannot know if this ancient medicinal knowledge resulted from systematic studies carried out with proper methods or from the accumulation of a collective body of knowledge having greatly benefitted from individual experiences. Over the centuries, along with the classification and archiving of traditional know-how, the research into new active compounds has been oriented towards rational exploratory strategies, or screens, in particular using plants and their extracts. The approaches based on systematic sorting have proved their worth, for example, through the research into antibiotics.

The recent progress in chemistry, biology, robotics and informatics have, since the 1990s, enabled an increase in the rate of testing, giving rise to the term **high-throughput screening**, as well as the measurement of multiparametric signals, known as **high-content screening**. Beyond the medical applications, which have motivated the growth of screening technologies in pharmaceutical firms, the **small molecule** has become a formidable and accessible tool in fundamental research. The know-how and original concepts stemming from robotic screening have given rise to a new discipline, **chemogenomics** (BREDEL and JACOBY, 2004), a practical component of which is **chemical genetics**, which we shall more specifically address in the second part of this book.

Pharmacological screening involves very diverse professions, which have their own culture and jargon, making it difficult not only for biologists, chemists and informaticians to understand each other, but so, too, for those within a given discipline. What is an ‘activity’ to a chemist or to a biologist, a ‘test’ to a biologist or to an informatician, or even a ‘control’? **Common terminology** must remove these ambiguities. This introductory chapter briefly describes the steps of an automated screening **process**, gives a preview of the different types of collections of molecules, or **chemical libraries**, and finally tackles the difficult question of what are the definitions of a **screen** and of **bioactivity**.

1.2. THE SCREENING PROCESS: TECHNOLOGICAL OUTLINE

1.2.1. MULTI-WELL PLATES, ROBOTS AND DETECTORS

Automated pharmacological screening permits the parallel testing of a huge number of molecules against a biological target (extracts, cells, organisms). For each molecule in the collection, a test enabling measurement of an effect on its biological target is implemented and the corresponding signal is measured. Based on this signal a choice is made as to which of the molecules are interesting to keep (fig. 1.1).

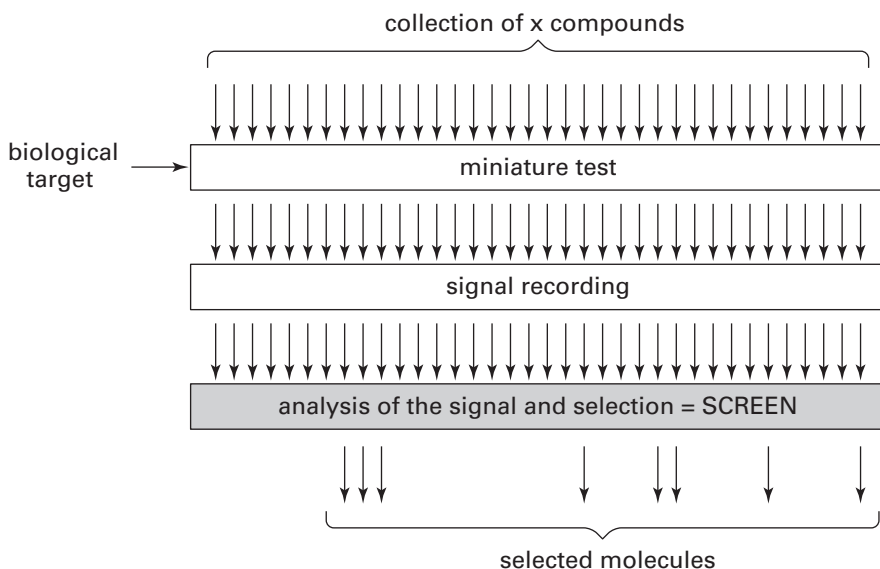


Fig. 1.1 - Scheme of a pharmacological screening process

The mixture of molecules and target as well as the necessary processes for the test are carried out in plates composed of multiple wells (termed **multi-well plates**, or **microplates**, fig. 1.2). These plates have standardised dimensions with 12, 24, 48, 96, 192, 384 or 1536 wells.

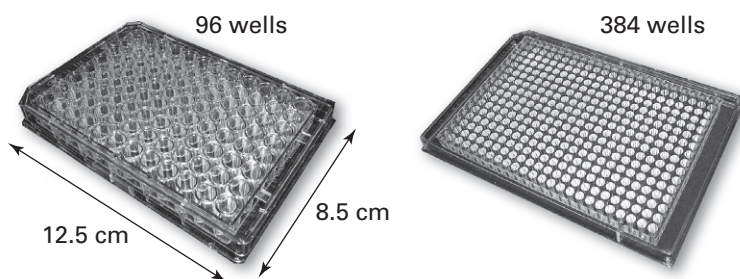


Fig. 1.2 - Multi-well plates with 96 and 384 wells (85 × 125 mm)

When more than 10,000 components are screened per day, the term ‘**micro HTS**’ (**μHTS** or **uHTS**) is employed. The financial savings made by miniaturisation are not negligible. The recent technological developments are also directed towards miniaturised screening on **chips** (microarrays). However, screening is a technological and experimental compromise aimed at the simplicity of setting up, the robustness of the tests and the reliability of the results; robotic screening is not heading relentlessly on a track towards miniaturisation. Rather than an unrestrained increase in the testing rate and over-miniaturisation, the current developments in screening are turning towards methods that allow the maximum amount of information to be gained about the effects of the molecules tested, thanks to the measurement of multiple parameters or even to image capture using microscopes (high-content screening; see the second part). The current state of automated screening technology still relies, and probably will for a long time, on the use of plates with 96 or 384 wells.

Parallel experiments are undertaken with the help of robots (fig. 1.3). These machines are capable of carrying out independent sequential tasks such as dilution, pipetting and redistribution of components in each well of a multi-well plate, stirring, incubation and reading of the results. They are driven by software specifically adapted to the type of experiment to be performed. The tests are done in standard microplates identified by a barcode and manipulated by the robotic ‘hand’: for example, it may take the empty plate, add the necessary reagents and the compounds to be tested, control the time of the reaction and then pass the plate to a reader to generate the results. For visualising the reactions arising from the molecule’s contact with the target in the wells, different methods are used based on measurements of absorbance, radioactivity, luminescence or fluorescence, or even imaging methods. The process of screening and data collection is controlled by a computer. Certain steps can be carried out independently of the robot, such as the detection of radioactivity or the analyses done manually using microscopes etc.

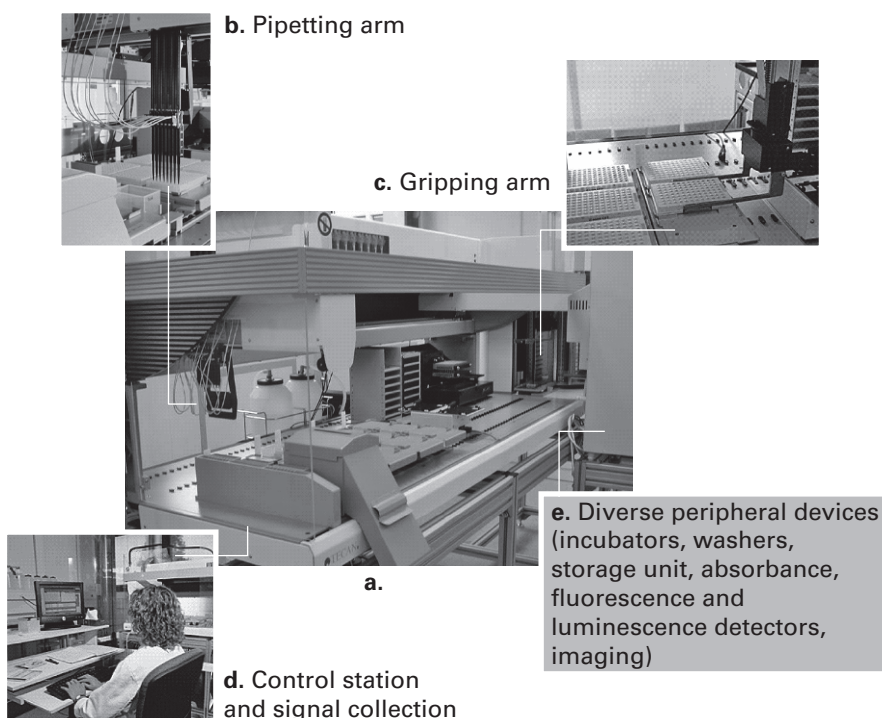


Fig. 1.3 - An example of a screening robot

(Centre for Bioactive Molecules Screening, Grenoble, France)

The robot ensures a processing sequence to which the microplates are subjected (pipetting, mixing, incubation, washing etc.) and the measurement of different signals according to the tests undertaken (e.g. absorbance, fluorescence, imaging). Each microplate therefore has a 'life' in the robot. A group of control programs is required to optimise the processing of several microplates in parallel (plate scheduling).

1.2.2. CONSUMABLES, COPIES OF CHEMICAL LIBRARIES AND STORAGE

With each screen, a complete stock of reagents is consumed. More specifically, a series of microplates corresponding to a copy of the collection of targetted molecules is used (fig. 1.4).

1.2.3. TEST DESIGN, PRIMARY SCREENING, HIT-PICKING, SECONDARY SCREENING

The screening is carried out in several phases. Before anything else, a **target** is defined for a scientific project motivated by a fundamental or applied research objective. We shall see below that the definition of a target is a difficult issue. A **test** is optimised in order to identify interesting changes (inhibition or activation) caused by the small molecule(s) and collectively referred to as **bioactivity**.



Fig. 1.4 - Preparation of a chemical library for screening

A chemical library is replicated in batches for single use. One copy (batch), stored in the cold, is used for screening.

Often, different types of test can be envisaged to screen for molecules active towards the same target. At this stage, a deeper consideration is indispensable. This consideration notably must take into account the characteristics of the chemical libraries used and attempt **to predict** as well as possible in which circumstances the implementation of the test might end with erroneous results ('false positives' and 'false negatives', for example). The biological relevance of the test is critical for the relevance of the molecules selected from the screen, with respect to subsequent interest in them as *drug candidates* and/or as *research tools* (chapter 3).

- » A **primary** screen of the entire chemical library is undertaken in order to select, based on a threshold value, a first series of candidate compounds.
- » A **secondary** screen of the candidate molecules enables their validation or invalidation before pursuing the study. In order to perform this secondary screen, the selected molecules are regrouped into new microplates. The sampling of these **hit molecules** is done with a robot. This step is called **hit-picking**.

1.3. THE SMALL MOLECULE: OVERVIEW OF THE DIFFERENT TYPES OF CHEMICAL LIBRARY

1.3.1. THE SMALL MOLECULE

Small molecule is a term often employed to describe the compounds in a chemical library. It sums up one of the required properties, *i.e.* a molecular mass (which is obviously correlated with its size) less than 500 daltons. A small, active molecule is sought in collections of pure or mixed compounds, arising from natural substances or chemical syntheses.

1.3.2. DMSO, THE SOLVENT FOR CHEMICAL LIBRARIES

Dimethylsulphoxide (DMSO, [fig. 1.5](#)) is the solvent frequently used for dissolving compounds in a chemical library that were created by chemical synthesis. DMSO improves the solubility of hydrophobic compounds; it is miscible with water.

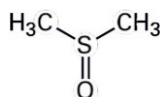


Fig. 1.5 - Dimethylsulphoxide (DMSO),
the solvent of choice for chemical libraries

One of the properties of DMSO is also to destabilise the biological membranes and to render them porous, allowing access to deep regions of the cell and may provoke, depending on the dose, toxic effects. Although DMSO is accepted to be inert towards the majority of biological targets, it is important to determine its effect with appropriate controls before any screening. In case the DMSO is found harmful for the target, it is critical to establish at what concentration of DMSO there is no effect on the target and consequently to adapt the dilution of the molecules in the library. Sometimes, it may be necessary to seek a solvent better suited to the experiment.

1.3.3. COLLECTIONS OF NATURAL SUBSTANCES

Natural substances are known for their diversity (HENKEL *et al.*, 1999) and for their structural complexity (TAN *et al.*, 1999; HARVEY, 2000; CLARDY and WALSH, 2004). Thus, 40% of the structural archetypes described in the data banks of natural products are absent from synthetic chemistry. From a historical point of view, the success of natural substances as a source of medicines and bioactive molecules is evident (NEWMAN *et al.*, 2000).

Current methods for isolating a natural bioactive product, called bio-guided purifications, are iterative processes consisting of the extraction of the samples using solvents and then testing their biological activity (see chapter 17). The cycle of purification and testing is repeated until a pure, active compound is obtained. While allowing the identification of original compounds arising from biodiversity, this type of approach does present several limitations (LOCKEY, 2003). First of all,

the extracts containing the most powerful and/or most abundant bioactive molecules tend to be selected, whereas interesting but less abundant compounds, or those with a moderate activity, would not be retained. Cytotoxic compounds can mask more subtle effects resulting from the action of other components present in the crude extract. Synergistic or cooperative effects between different compounds from the same mix may also produce a bioactivity that disappears later upon fractionation. Pre-fractionation of the crude extracts may, in part, resolve these problems (ELDRIDGE *et al.*, 2002). Mindful of these pitfalls, some pharmaceutical firms choose to develop their collections of pure, natural substances from crude extracts. This strategy, despite requiring significant means, can prove to be beneficial in the long term (BINDSEIL *et al.*, 2001). Lastly, with chemical genetics approaches (second part), the strategies adopted for identifying the protein target may necessitate the synthesis of chemical derivatives of the active compounds, which can present a major obstacle for those natural substances coming from a source in short supply and/or that have a complex structure (ex. 1.1).

Depending on the synthesis strategy used (see chapter 10), **two types of collection can be generated: target-oriented collections**, synthesised from a given molecular scaffold, and **diversity-oriented collections** (SCHREIBER, 2000; VALLER and GREEN, 2000). Each of these types of collection has its advantages and disadvantages. Compounds coming from a **target-oriented** collection have more chance of being active than those selected at random, however, they may only display activity towards a particular class of proteins. A diversity-oriented collection (chapter 10), on the other hand, offers the possibility of targetting entirely new classes of protein. Each individual molecule has, however, a lower probability of being active.

Example 1.1 - An anti-cancer compound from a sponge

Obtaining 60 g of discodermolide, an anti-cancer compound found in *Discodermia dissoluta* (GUNASEKERA *et al.*, 1990), a rare species of Caribbean sponge, would require 3,000 kg of dry sponges, *i.e.* more sponges than in global existence. Therefore, chemists have attempted to synthesise the discodermolide molecule. Only in 2004 did a pharmaceutical group announce that, after two years of work, they managed to produce 60 g of synthetic discodermolide, by a process consisting of over thirty steps (MICKEL, 2004). Discodermolide is now under evaluation in clinical studies for its therapeutic effect on pancreatic cancer.

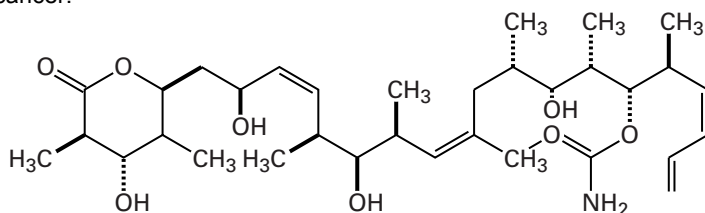


Fig. 1.6 - Discodermolide

Several groups have attempted to reproduce, with the help of combinatorial organic synthetic methods, the diversity and complexity of natural substances. The current developments in combinatorial synthesis are moving towards the simultaneous and

parallel synthesis, on a solid support, of a great number of organic compounds using automated systems. A strategy often employed is ‘split/mix’ synthesis or ‘one bead/one molecule’ synthesis ([Harvard] web site, see below in *References*). Based on various structural archetypes – often heterocyclic or arising from natural products – bound to a solid phase, several groups have developed collections of molecules possessing a complex skeleton (GANESAN, 2002; LEE *et al.*, 2000; NICOLAOU *et al.*, 2000). These strategies in combinatorial chemistry face two difficulties: first, obtaining the molecules in sufficient quantity for screening and second, synthesising *de novo* the molecule of interest. These constraints involve following the history of the steps in the synthesis of each molecule, aided, for example, by coding systems (XIAO *et al.*, 2000).

1.3.4. COMMERCIAL AND ACADEMIC CHEMICAL LIBRARIES

It is possible to purchase compound collections showing a great diversity, or even target-oriented collections, from specialised companies. In general, these collections are of high purity and are provided in the form of microplates adapted for high-throughput screening. We speak of **chemical library formatting**.

For a decade, several initiatives have been underway to make available the collections of molecules developed in academic laboratories. The National Cancer Institute (NCI) in the USA, for example, has different collections of synthetic compounds and natural substances (<http://dtp.nci.nih.gov/repositories.html>). The French public laboratories of chemistry are organised in such a way as to classify and format in microplates the different molecules that they collectively produce, in a move to promote them in the public interest (<http://chimiotheque-nationale.enscm.fr/>; chapter 2). The public collections contain several tens of thousands of compounds.

1.4. THE TARGET, AN ONTOLOGY TO BE CONSTRUCTED

1.4.1. THE DEFINITION OF A TARGET DEPENDS ON THAT OF A BIOACTIVITY

When a small molecule interacts with an enzyme, a receptor, a nucleotide sequence, a lipid, an ion in solution or a complex structure, it can induce an interesting functional perturbation on a biological level. Thus, we say that the molecule is active towards a biological process, that it is **bioactive**. In addition, small molecules are also studied for their specific binding to particular biological entities, thereby constituting probes, markers and tracers, for visualising these species in cell biology, for example, without notable effects upon the biology of the cell. Therefore, to different degrees, **bioactivity covers two properties of a molecule**:

- › **binding to a biological object** (binding to a receptor, a protein, a nucleotide, or engulfment of an ion), and
- › **interference with a function** (for instance, the inhibition or activation of an enzyme, or of a dynamic process or cellular phenotype).

The term **bioactivity** began appearing in the scientific literature at the end of the 1960s (SINKULA *et al.*, 1969). This term, which **removes the ambiguity from the word ‘activity’** (with its varying meanings in biology, biochemistry, chemistry etc.), has progressively established itself (fig. 1.7).

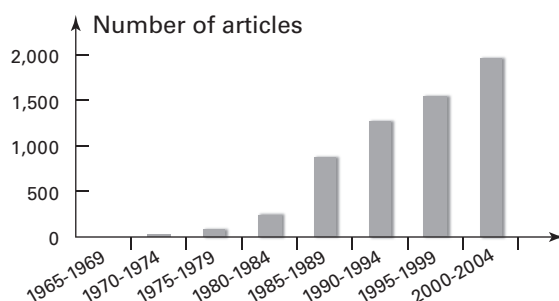


Fig. 1.7 - Number of publications citing the term bioactivity in a pharmacological context, in the biological literature since the 1960s (histogram constructed from data in PubMed; National Center for Biotechnology Information)

1.4.2. DUALITY OF THE TARGET:

MOLECULAR ENTITY AND BIOLOGICAL FUNCTION

Which biological object should be targetted? Which biological function should be disrupted? What we mean by ‘target’ is the answer to both of these questions. By target, we may mean the physical biological object, such as a receptor, a nucleotide, an organelle, an ion etc., but it may also refer to a biological function, from an enzymatic or metabolic activity to a phenotype on the cellular level or on that of the whole organism (fig. 1.8).

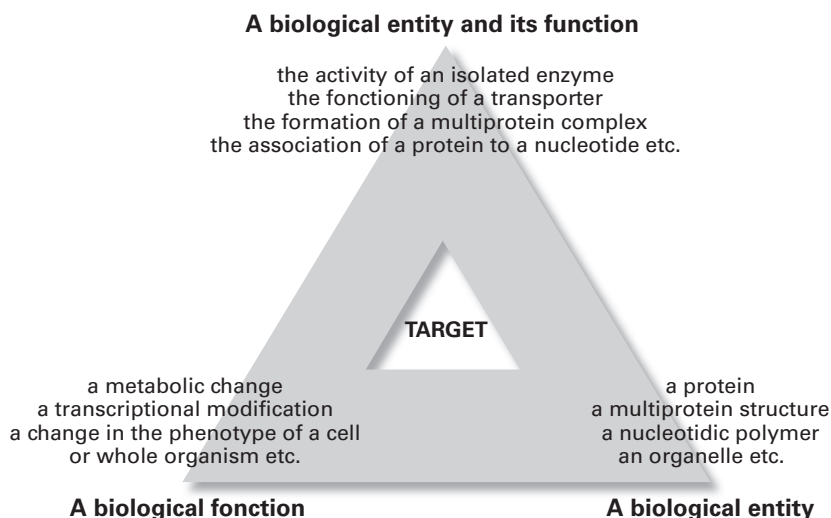


Fig. 1.8 - Defining the biological target

It may be characterised structurally (biological object) and/or functionally (biological function).

In a way consistent with the definition of the target, a **test** is developed (chapter 3). The test is itself composite in nature, conceived to be carried out in parallel (implemented in series) in order to screen a collection of molecules. **Enzymatic screens** and **binding screens** utilise tests that quantify relatively simple processes *in vitro*, in which the target has been previously characterised in terms of its structure. Binding screens can also be based on structural knowledge of the target and the molecule. This type of screening done *in silico* is also called **virtual screening** (chapter 16). Lastly, for **phenotypic screens**, the physical nature of the target is unknown and must be characterised *a posteriori* (chapters 8 and 9).

1.4.3. AN ONTOLOGY TO BE CONSTRUCTED

An ontology refers to the unambiguous representation of the ensemble of knowledge about a given scientific object. In its most simplified (or simplistic) form, the ontology represents a hierarchical vocabulary. Rather than linearly defining a complex notion, we attempt to embrace its complexity by a diagram representing the different meanings by which this notion is clarified. The well-known example is that of the gene. Is the gene a physical entity on a chromosome? If this is the case, is it just the coding framework, with or without the introns that interrupt it, with or without the regions of DNA that regulate it? Is the function of a gene to code for a protein, or does the gene's function overlap with that of the protein? What is then the function of this protein? Is it the activity measured *in vitro*, or rather the group of physiological processes that depend on it? Are genes related in evolution equivalent in different living species? A consortium has been set up to tackle the complexity of **gene ontology** (ASCHBURNER *et al.*, 2000; <http://www.geneontology.org/>) in its most simple form, *i.e.* a hierarchical vocabulary. This short paragraph shows clearly that the question of the target is similar to the question of the gene. A reflection on the **ontology of the target** must be undertaken in the future (chapter 14). For the particular case of **phenotypic screening of whole organisms** (chapter 9), the description of the target can readily include the **taxon** in which the species is found according to an ontology arising from the long history of evolutionary science (see the taxonomic browser at the National Center for Biotechnology Information, <http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html>). For cells **phenotypic screens**, an ontology of the cell is even more in its infancy (BARD *et al.*, 2005).

Based on the model of the very popular database, PubMed (<http://www.ncbi.nlm.nih.gov/pubmed>), which has been available for a number of years to the community of biologists, a public database integrating comprehensive biological and chemical data, PubChem (<http://pubchem.ncbi.nlm.nih.gov/>), has been recently introduced for chemogenomics.

In this book, the concept of a target will, as far as possible, be described in terms of its molecular, functional and phenotypic aspects.

1.5. CONTROLS

It is crucial to define the **controls** that permit the analysis and exploitation of the screening results. This point is so critical that it deserves a paragraph, albeit a brief one. There is nothing less imprecise than the term ‘control’. What do we mean by a positive control, a negative control and a blank?

Example 1.2 - What is a positive control?

Is the positive control the biological object without additives, functioning normally?

⇒ In this case, what is ‘positive’ would be the integrity of the target. The positive control would be then measured in the absence of any bioactive molecule.

Is the positive control a molecule known to act upon the target?

⇒ In this case, what is ‘positive’ would be the bioactivity of the molecule, *i.e.* exactly the opposite of the preceding definition. ▲

Example 1.3 - What is a blank?

- ▶ The mixture without the target?
- ▶ The mixture without the bioactive molecule?
- ▶ The mixture without the target's substrate? ▲

In order to remove this ambiguity, an explicit terminology is necessary. The notion of bioactivity allows, in addition, a comparison of screening results.

- » Thus, by **control for bioactivity**, we mean a mixture *with* a molecule known to be active towards the target (**bioactive**). The concentration of this bioactive molecule used as a control can be different from the concentration of the molecules tested (it is possible to screen molecules at 1 μ M while using a control at 10 mM, if this is the concentration necessary to measure the effect on the target).
- » On the contrary, the **control for bio-inactivity** is a mixture *without* a bioactive molecule. The mixture can be chosen without the addition of other molecules and yet containing DMSO, in which the molecules are routinely dissolved.

1.6. A NEW DISCIPLINE AT THE INTERFACE OF BIOLOGY, CHEMISTRY AND INFORMATICS: CHEMOGENOMICS

The costly technological developments (namely: robotics, miniaturisation, standardisation, parallelisation, detection and so on) which have led to the creation of screening platforms, were initially motivated by the discovery of new medicines. As with all innovation, there have been the enthusiasts and the sceptics. To undertake an **assessment of the contribution of automated screening to the discovery of new medicines** is difficult, in particular due to the length of the cycles of discovery (on average 7 years) and development (on average 8 years) for novel molecules (MACARRON, 2006). We counted 62 candidate molecules arising from HTS in 2002, 74 in 2003 and 104 in 2005, numbers that are very broadly underestimated

since our enquiries did not cover all laboratories and were restricted for reasons of confidentiality. Furthermore, the success rate is variable depending on the targets, with regular success with certain targets and systematic failure with others, which may be taken to be ‘**not druggable**’. MACARRON (2006) notes on this topic that failure can also be due to the quality of the chemical library. Thus, after the firms GlaxoWellcome and SmithKline Beecham merged, certain screens that had initially failed with the library from one of the original firms, actually succeeded with the library of the other. The question of the ‘right chemical library’ is thus a critical one and an open field of study.

To close this chapter, we ask the question of the place of automated screening in the scientific disciplines. Is this technology just scientific progress enabling the miniaturisation of tests and an increase in the rate of manual pipetting? According to the arguments of Stuart SCHREIBER and Tim MITCHISON from Harvard Medical School, Massasuchetts, a new discipline at the interface of genomic and post-genomic biology and chemistry was born thanks to the technological avances brought about by the development of automated screening in the academic community: **chemogenomics**. This emerging discipline combines the latest developments in genomics and chemistry and applies them to the discovery as much of bioactive molecules as of the targets (BREDEL and JACOBY, 2004). More widely, the object of chemogenomics is to study the relationships between the space of biological targets (termed **biological space**) and the space of small molecules (**chemical space**) (MARECHAL, 2008). This ambitious objective necessitates that data from both biological and chemical spaces be structured optimally in **knowledge-bases** in order for them to be efficiently explored using **data-mining** techniques. Fig. 1.9 shows, with an example of the strategy applied to reverse chemical genetics, the place of chemogenomics at the fulcrum of the three disciplines: biology, chemistry and informatics.

This book does not treat chemogenomics as a mature discipline, but as a nascent discipline. Above all, we shall shed light on what biologists, chemists and informaticians can today bring, and find, when their curiosity leads them to probe the encounter between the living world and chemical diversity.

1.7. CONCLUSION

Motivated initially by the research into novel medicines, pharmacological screening today offers the opportunity to select small molecules that are active towards biological targets for fundamental research ends. New tools (chemical libraries, screening platforms) and new concepts (the small molecule, the target and bioactivity) have founded an emerging discipline that necessitates very strong expertise in chemistry, biology and informatics (chemogenomics), with different research strategies (forward and reverse chemical genetics). Multidisciplinarity requires a shared language. There is no ideal solution. Nevertheless, the concept of a

bioactive molecule seems sufficiently central to help remove the ambiguities over the terms target, screen, test, signal and control. In case of doubt, the reader is encouraged to consult the general glossary at the end of this book.

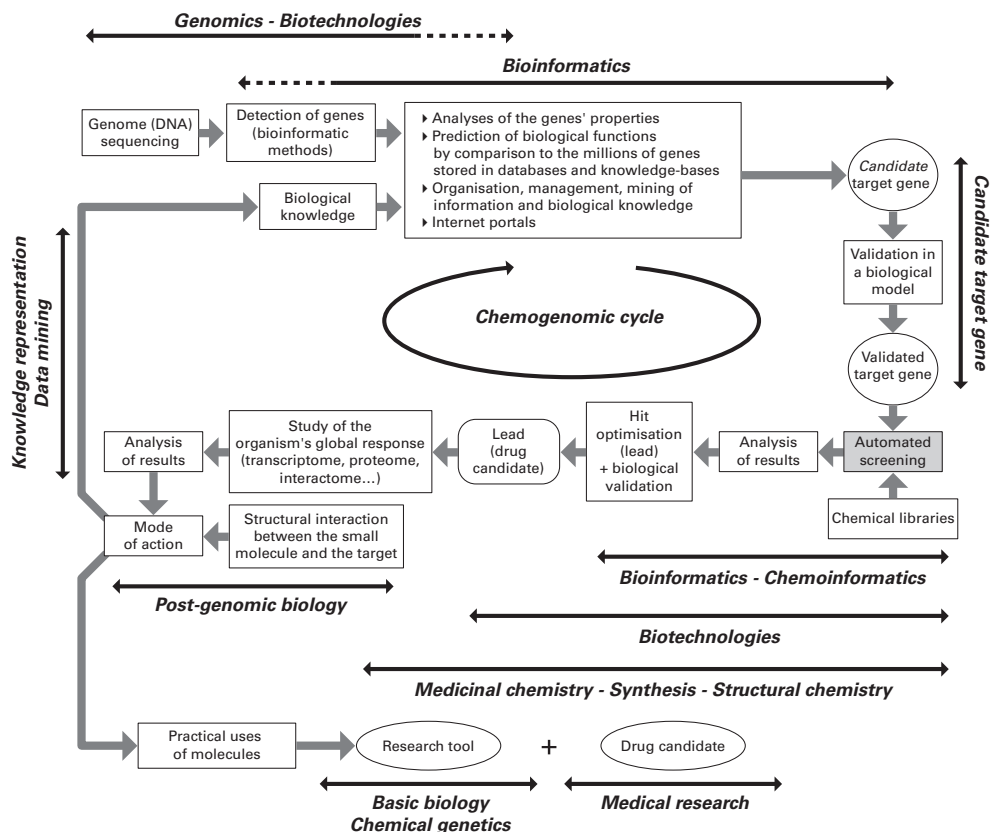


Fig. 1.9 - Chemogenomics, at the interface of genomics and post-genomic biology, chemistry and informatics

Chemogenomics aims to understand the relationship between the biological space of targets and the chemical space of bioactive molecules. This discipline has been made possible by the assembly of collections of molecules, the access to automated screening technologies and significant research in bioinformatics and chemoinformatics.

1.8. REFERENCES

[Harvard]

http://www.broad.harvard.edu/chembio/lab_schreiber/anim/animations/smdbSplitPool.php

ASHBURNER M., BALL C.A., BLAKE J.A., BOTSTEIN D., BUTLER H., CHERRY J.M., DAVIS A.P., DOLINSKI K., DWIGHT S.S., EPPIG J.T., HARRIS M.A., HILL D.P., ISSEL-TARVER L., KASARSKIS A., LEWIS S., MATESE J.C., RICHARDSON J.E., RINGWALD M., RUBIN G.M., SHERLOCK G. (2000) Gene ontology: tool for the unification of biology.

The Gene Ontology Consortium. *Nat. Genet.* **25**: 25-29

- BARD J., RHEE S.Y., ASHBURNER M. (2005) An ontology for cell types. *Genome Biol.* **6**: R21
- BINDSEIL K.U., JAKUPOVIC J., WOLF D., LAVAYRE J., LEBOUL J., VAN DER PYL D. (2001) Pure compound libraries; a new perspective for natural product based drug discovery. *Drug Discov. Today* **6**: 840-847
- BREDEL M., JACOBY E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **5**: 262-275
- CLARDY J., WALSH C. (2004) Lessons from natural molecules. *Nature* **432**: 829-837
- ELDRIDGE G.R., VERVOORT H.C., LEE C.M., CREMIN P.A., WILLIAMS C.T., HART S.M., GOERING M.G., O'NEIL-JOHNSON M., ZENG L. (2002) High-throughput method for the production and analysis of large natural product libraries for drug discovery. *Anal. Chem.* **74**: 3963-3971
- GANESAN A. (2002) Recent developments in combinatorial organic synthesis. *Drug Discov. Today* **7**: 47-55
- GUNASEKERA S.P., GUNASEKERA M., LONGLEY R.E., SCHULTE G.K. (1990) Discodermolide: a new bioactive polyhydroxylated lactone from the marine sponge *Discodermia dissoluta*. *J. Org. Chem.* **55**: 4912-4915
- HARVEY A. (2000) Strategies for discovering drugs from previously unexplored natural products. *Drug Discov. Today* **5**: 294-300
- HENKEL T., BRUNNE R.M., MULLER H., REICHEL F. (1999) Statistical investigation into the structural complementarity of natural products and synthetic compounds. *Angew. Chem.* **38**: 643-647
- LEE D., SELLO J.K., SCHREIBER S.L. (2000) Pairwise use of complexity-generating reactions in diversity-oriented organic synthesis. *Org. Lett.* **2**: 709-712
- LOKEY R.S. (2003) Forward chemical genetics: progress and obstacles on the path to a new pharmacopoeia. *Curr. Opin. Chem. Biol.* **7**: 91-96
- MACARRON R. (2006) Critical review of the role of HTS in drug discovery. *Drug Discov. Today* **11**: 277-279
- MARECHAL E. (2008) Chemogenomics: a discipline at the crossroad of high throughput technologies, biomarker research, combinatorial chemistry, genomics, cheminformatics, bioinformatics and artificial intelligence. *Comb. Chem. High Throughput Screen.* **11**: 583-586
- MICKEL S.J. (2004) Toward a commercial synthesis of (+)-discodermolide. *Curr. Opin. Drug Discov. Devel.* **7**: 869-881
- NEWMAN D.J., CRAGG G.M., SNADER K.M. (2000) The influence of natural products upon drug discovery. *Nat. Prod. Rep.* **17**: 215-234
- NICOLAOU K.C., PFERFFERKOM J.A., ROECKER A.J., CAO G.Q., BARLUENGA S., MITCHELL H.J. (2000) Natural product-like combinatorial libraries based on privileged structures. 1. General principles and solid-phase synthesis of benzopyrans. *J. Am. Chem. Soc.* **122**: 9939-9953

SCHREIBER S.L. (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**: 1964-1969

SINKULA A.A., MOROZOWICH W., LEWIS C., MACKELLAR F.A. (1969)
Synthesis and bioactivity of lincomycin-7-monoesters. *J. Pharm. Sci.* **58**: 1389-1392

TAN D.S., FOLEY M.A., STOCKWELL B.R., SHAIR M.D., SCHREIBER S.L. (1999)
Synthesis and preliminary evaluation of a library of polycyclic small molecules for use in chemical genetic assays. *J. Am. Chem. Soc.* **121**: 9073-9087

VALLER M.J., GREEN D. (2000) Diversity screening versus focused screening in drug discovery. *Drug Discov. Today* **5**: 286-293

XIAO X.Y., LI R., ZHUANG H., EWING B., KARUNARATNE K., LILLIG J., BROWN R., NICOLAOU K.C. (2000) Solid-phase combinatorial synthesis using MicroKan reactors, Rf tagging, and directed sorting. *Biotechnol. Bioeng.* **71**: 44-50

COLLECTIONS OF MOLECULES FOR SCREENING: EXAMPLE OF THE FRENCH NATIONAL CHEMICAL LIBRARY

Marcel HIBERT

2.1. INTRODUCTION

The technological progress in molecular biology and the **genomic** revolution marked the 1990s by the race to sequence the whole genomes of viruses, bacteria, plants, yeasts, animals and pathogenic organisms. As for the human genome, we now have available thousands of novel genes whose biological functions and therapeutical interest remain to be elucidated. The challenge of the **post-genomic** era is now to explore this macromolecular space, which is characterised by an unprecedented amount of information. The relationship:

gene (DNA polymer) $\longleftrightarrow$ **protein** (polymer of amino acids)

can be addressed thanks to high-throughput technologies (**transcriptomics** for the transcription of DNA to RNA; **proteomics** for the characterisation of proteins). The question of the relationship between the gene and what its presence implies for the organism (the structures and functions governed by the gene) is much more difficult. We speak of a **phenotype** to designate the set of structural and functional characteristics of an organism governed by the action of genes, in a given biological and environmental context.

gene / protein $\longleftrightarrow$ **phenotype ?**

A lengthy phase of dissection and integration of the molecular and physiological mechanisms relating genes and phenotypes is underway. The recent years have seen the emergence or the strengthening of such disciplines as bioinformatics, genomics, proteomics and genetics. Each approach is complementary and must be employed in a similar manner in order to elucidate the possible function(s) of genes and the proteins encoded by them (see chapter 1 and [fig. 2.1](#)). Together, however, these approaches turn out to be incomplete. The inactivation of a gene by mutation theoretically permits the study of the phenotype obtained and hence elucidation of the function of the gene concerned.

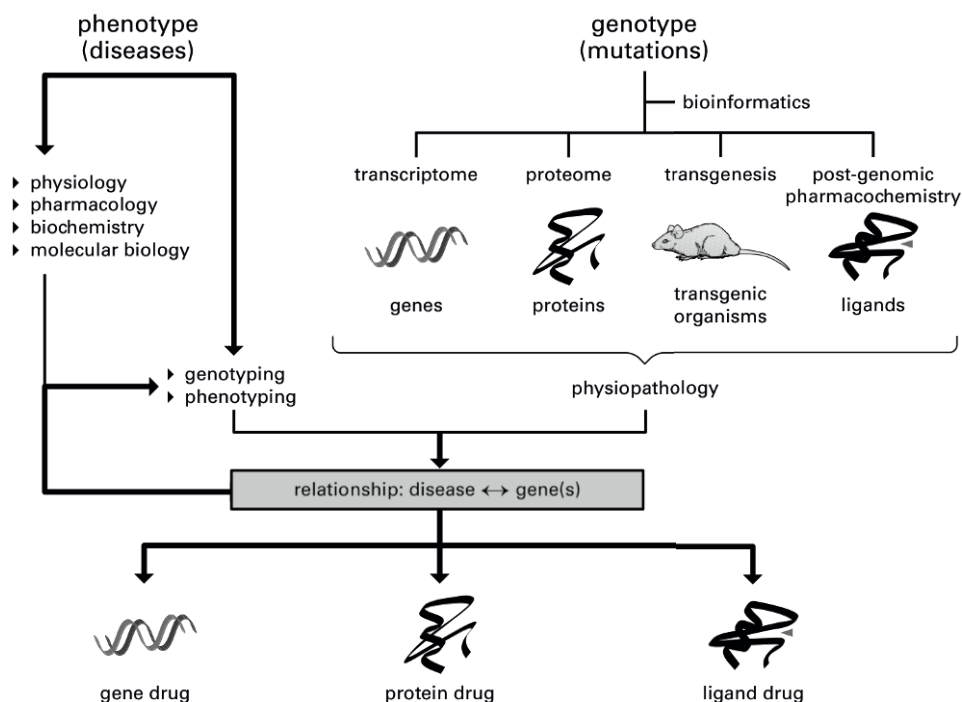


Fig. 2.1 - Strategies for post-genomics

In the above scheme, the different strategies for post-genomic research are illustrated for the biomedical context. The 'disease' is characterised by a phenotype that diverges from the state observed for a 'healthy' patient. One aim is to relate this phenotype to the macromolecular genomic or proteomic data and to open suitable therapies.

However, **conventional genetics cannot achieve everything:**

- › certain genes are duplicated in the genome forming multi-gene families, which compensate for the effect of a mutation in one of its members,
- › some mutations have no effect under the particular experimental study conditions,
- › some mutations are lethal and no clear information can be derived after introducing the mutation,
- › certain organisms quite simply cannot be mutated at will (plants, for example).

The search for potent, specific and efficacious ligands is a very promising complementary strategy that can overcome these difficulties (see the second part and [fig. 2.2](#)). Indeed, small molecules that are effective towards biological targets constitute flexible research tools for exploring molecular, cellular and physiological functions in very diverse conditions (e.g. dose, medium, duration of activation), without involving the genome *a priori*.

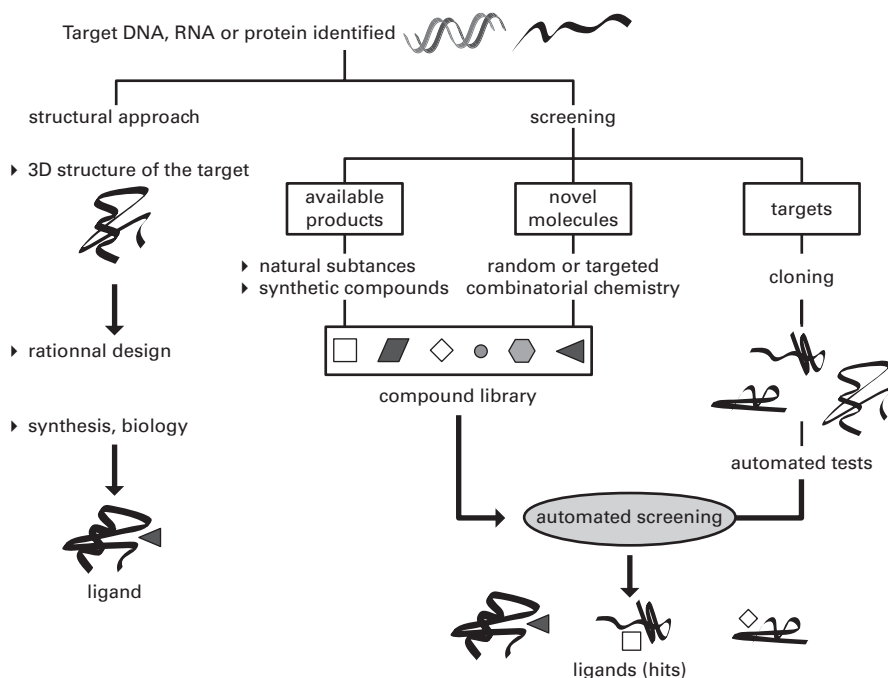


Fig. 2.2 - Strategies in chemogenomics

This scheme summarises the different strategies by which ligands of biological macromolecules can be identified. It does not display the strategies for phenotypic screening, which is the subject of chapters 8 and 9.

The second part of this book will explore more particularly what is meant today by **chemical genetics**. Two heavy investments must be made to enable the development of **chemogenomics** in an academic environment: the acquisition of **screening robots**, and the assembly of collections of molecules or natural substances destined for screening, *i.e.* **chemical libraries**.

2.2. WHERE ARE THE MOLECULES TO BE FOUND?

Where are the molecules and natural substances necessary for screening to be found? A large number of chemical libraries are commercially available, which can be globally classified into three categories:

- » The collections of molecules retrieved from diverse medicinal chemistry laboratories in several countries: such collections offer a huge diversity of molecular structures, and the possibility to initiate scientific partnerships between biologists and chemists in order to optimise the hit into a useful pharmacological probe or drug candidate.
- » Synthetic chemical libraries arising from combinatorial chemistry: these chemical libraries are huge in size, but usually consist of poor structural diversity. The hit rate they deliver is often disappointing.

- » Targetted chemical libraries based upon pharmacophores: these chemical libraries are small in size and are generally more limited in structural diversity, but are well suited to afford a hit rate above average in screening campaigns on their targets (see chapter 10).

In which category are public chemical libraries? There exists principally a large public collection of molecules aimed at cancer screening, available from the **National Cancer Institute** (NCI, <http://dtp.nci.nih.gov/discovery.html>), USA, as well as some smaller-scale initiatives such as a specialised chemical library developed for AIDS in Belgium (DE CLERCQ, 2004). Access to these libraries is currently restricted and their sizes are modest. The development of a collection of small molecules and natural substances more freely exploitable by (and for) public research has motivated the constitution of a wider public chemical library in France, whose molecules and substances come from a pooling of those available in public research laboratories or be synthesised or collected *de novo*. This initiative has led to the creation of the **French National Chemical Library (in French, Chimiothèque Nationale)**, while awaiting the creation of a **European Chemical Library (HIBERT, 2009)**. A major objective has been for the components of the French National Chemical Library to be inventoried in a centralised public database, freely accessible to the scientific community, and for each to be stored in a standard format compatible with robotic screening. Initiated and validated by a few research groups, the chemical library and the network of laboratories to this day links together 24 universities and public institutes. **Copies** of this collection (**replicas**) are, if needed, to be negotiated with academic laboratories or industrialists to be screened in partnerships..

In practical terms, the establishment of the *Chimiothèque Nationale* involved:

- the identification, collection (weighing-in) and organisation of synthetic molecules, natural substances or their derivatives existing in academic laboratories,
- their recording in a database that is computationally managed,
- the standardisation of bottling and labelling stocks,
- the production of several copies of the entire range of products in 96-well plates, known as mother plates,
- the production from the mother plates, according to need, of daughter plates at 10^{-4} - 10^{-6} M destined to be made available for screening,
- the management of collaborations by contracting.

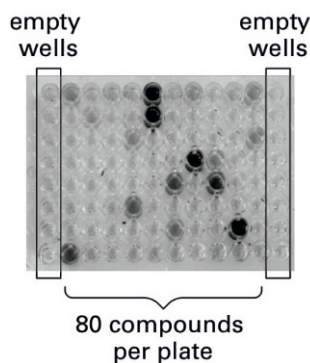
A molecule from the chemical library thus follows quite a course from its creation (the scientific context that motivated its synthesis, the chemist who designed and produced it), its collection, its weighing-in, its formatting, to its potential identification in the course of screening (the scientific context that motivated the screening of a given target, the researchers who carry out the biological project). The actions that we have just listed thus highlight three important constraints for building a chemical library: the significant effort of organisation and standardisation, the need

to be able to trace the course of each molecule and finally the necessary contracting to provide an operating framework.

In answer to these constraints, the *Chimiothèque Nationale* relies on some simple general principles:

- » The *Chimiothèque Nationale* is a federation of chemical libraries from different laboratories. The laboratories remain in charge of their own chemical library (management, valorization) but participate in concerted, collective action,
- » The members of the *Chimiothèque Nationale* adopt agreed communal conventions:
 - › recording of the molecules and natural substances in a centralised communal database in which, as a minimum requirement, feature the 2D structures of the molecules and their accessible structural descriptors (mass, $c \log P$ etc.; see chapters 11, 12 and 13). In the case of natural substances for which the structures of the molecules present are unknown, the identifiers and characteristics of the plants/extracts/fractions are to be indicated. For all substances, the names and contact details of the product managers are given as well as information for stock monitoring (available/out-of-stock, in plates or loose),
 - › an identical format for plate preparation: 96-well plates, containing 80 compounds (molecules, extracts, fractions) per plate at a concentration of 10^{-2} M, in DMSO. The first and last columns remain empty so as to accommodate the internal reference solutions during screening (fig. 2.3),
 - › a similar material transfer agreement.

Fig. 2.3 - A mother plate from the *Chimiothèque Nationale*
In this example of a plate, certain compounds, which are chromophores, display a characteristic colour.



2.3. STATE OF PROGRESS WITH THE EUROPEAN CHEMICAL LIBRARY

In terms of organisation, in 2003 the *Chimiothèque Nationale* became a service-oriented division of the French National Centre for Scientific Research, CNRS (see the website <http://chimiotheque-nationale.enscm.fr>). To date, the national database has indexed more than 40,000 molecules and more than 13,000 plant extracts

available in plates from partner laboratories. The *Chimiothèque Nationale* will be expanded to the European level. In terms of scientific evaluation, the existing chemical libraries have already been tested on hundreds of targets in France and other countries, leading to the emergence of several research programmes at the interface of chemistry and biology. Several innovative research tools as well as some lead compounds with therapeutic applications have been discovered and are currently being studied further. The most advanced drug candidate derived from the *Chimiothèque Nationale* screening is Minozac currently in clinics in Phase II for the treatment of Alzheimer's disease.

2.4. PERSPECTIVES

In parallel to the development of this chemical library, a **network of robotic screening platforms** is being realised based on existing academic facilities and those newly emerging. The smooth integration of the *Chimiothèque Nationale*, screening platforms and the scientific projects designed around the targets, has led and will continue to lead more quickly to the discovery of original research tools, bringing a competitive advantage to the exploration and exploitation of biological processes. It also speeds up access to new potential therapeutic agents. Furthermore, it will prime and efficiently catalyse collaborations at the interface of chemistry and biology between university laboratories both in France and abroad, as well as collaborations between universities and industry. In this book, the questions dealing more specifically with molecular diversity are discussed in chapters 10, 11, 12, 13 and 16; the question of the choice of solvent is covered in chapters 1, 3 and 8; the question of the choice of chemical library is dealt with in chapters 8 and 16. This short presentation underlines, in brief, the huge effort in terms of organisation, the quality procedures (see chapter 7) and the contractual framework necessary for such a collaboration between laboratories to be able to succeed in enhancing the chemical heritage.

2.5. REFERENCES

[*Chimiothèque Nationale*] <http://chimiotheque-nationale.enscm.fr>

DE CLERCQ E. (2004) HIV-chemotherapy and -prophylaxis: new drugs, leads and approaches. *Int. J. Biochem. Cell Biol.* **36**: 1800-1822

HIBERT M (2009) French/European academic compound library initiate. *Drug Discov. Today* **14**:723-5.

Chapter 3

THE MINIATURISED BIOLOGICAL ASSAY: CONSTRAINTS AND LIMITATIONS

Martine KNIBIEHLER

3.1. INTRODUCTION

The aim of this chapter is to sensitise the reader to the precautions to take when miniaturising a screening assay and when interpreting the results.

to miniaturise = to adapt

The miniaturisation of a pharmacological screen represents a key step in the process of the discovery of bioactive molecules. It has to permit the development of the most powerful automated assay possible, which will then allow the selection of quality hits. The necessary steps for the **adaptation** of a biological assay to the format and pace of the screening platform are referred to by the term **miniaturisation**. However, whereas this term may suggest foremost a reduction in format and volume, the concept of miniaturisation is in fact much more complex. It comprises both the design aspects (choice of assay in terms of the biological reaction to be evaluated) and the technical and practical aspects (choice of a suitable technology for signal detection) (fig. 3.1).

The therapeutic targets currently listed (DREWS, 2000) are essentially proteins, the majority of which are enzymes and receptors. These targets can be classified into large families: **kinases** (enzymes catalysing the transfer of phosphate groups to other proteins, either to their serine or threonine residues, called Ser/Thr kinases, or to tyrosine residues, called Tyr kinases), **receptors** (the large majority being G-protein-coupled receptors, GPCR), **ion channels** and **transcription factors** (see chapter 14). This idea of **target families** can come into play, as we shall see further on, in the choice of equipment for screening platforms and/or the choice of biological assay: we may refer to these as ‘dedicated’ platforms.

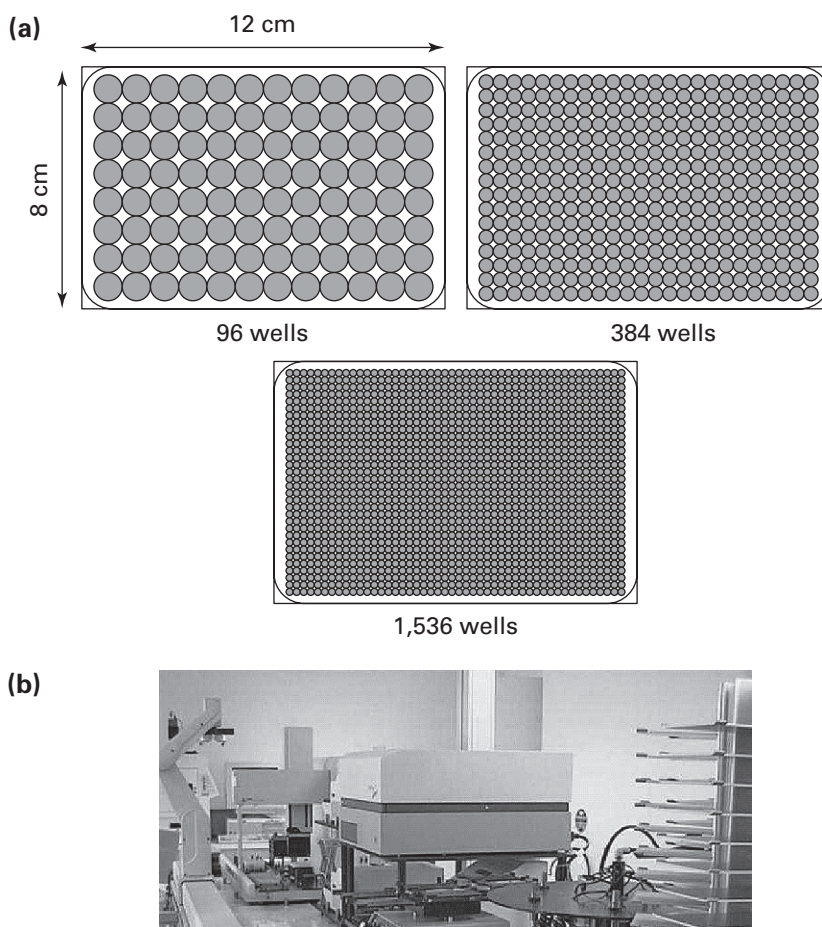


Fig. 3.1 - The material constraints in designing a test

(a) Adapt to the microplate format (form, dimensions, rigidity, planarity etc.). The test has to be able to be performed in the most commonly used plates, having 96, 384 or 1536 wells. The available liquid-dispenser heads comprise 8, 96 or 384 tips or needles. **(b)** Adapt to the specification of the platform (shown here, the IPBS platform, Toulouse, France). The test must be operational using robotic modules, permitting different operations for dispensing, transfer, washing, filtration, centrifugation, incubation etc. It has to be possible to track the progress of the test using an available means for measurement (signal detection).

3.2. GENERAL PROCEDURE FOR THE DESIGN AND VALIDATION OF AN ASSAY

In the long and costly process of the discovery of bioactive molecules (fig. 3.2), the factors leading to failure must be eliminated as early as possible (REVAH, 2002). The choices of target and biological assay that permit automated screening are thus the determining parameters with respect to the quality of the results.

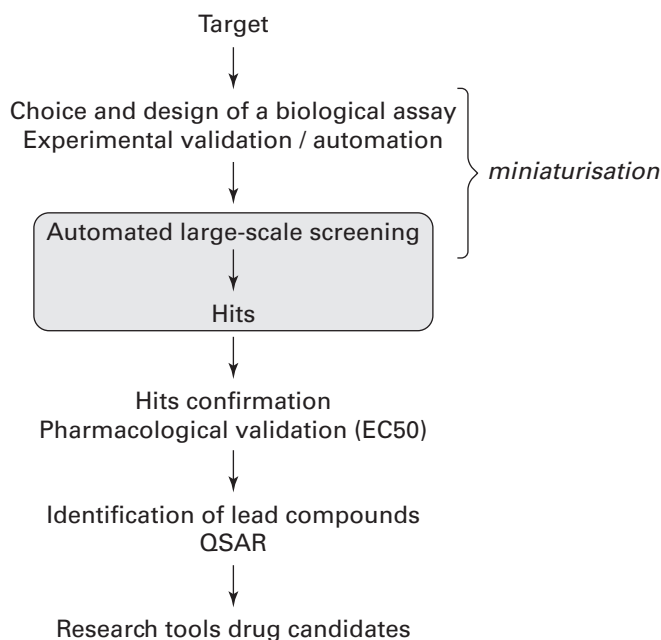


Fig 3.2 - Outline of the different stages in the miniaturisation of the process of bioactive molecule discovery for a pharmacological target of interest
For EC50, see chapter 5; for QSAR, see chapters 12, 13 and 15.

Consequently, it is important **to take the time to ask oneself questions in order to make appropriate choices**. All large-scale screening endeavours require from the outset a **precise assessment of the knowledge relating to the target**: its degree of pharmacological validation for a given pathology, the state of structural data (for the selection of molecular banks to screen or the optimisation of hits) and functional data (for the optimisation of the test), without forgetting the aspects concerning intellectual property (numerous assays are patented).

3.2.1. CHOICE OF ASSAY

Choice based on biological criteria

The first choice facing the experimenter is to determine whether screening should be carried out on the isolated target or in its cellular context; this question quite obviously depends on the target of interest, but remains widely debated (MOORE and REES, 2001; JOHNSTON and JOHNSTON, 2002). The goal is to put into practice the most **pertinent** and **informative** assay possible. An assay is considered pertinent in biological terms if the phenomenon measured permits answering most precisely the question asked. An assay is considered informative if it delivers a wealth of data regarding the molecules, which is sufficient to allow them to be sorted and selected (*i.e.* by efficacy, selectivity, toxicity etc.). In this context, the principal

advantage of the cellular assay is that it constitutes a predictive model of the expected physiological response.

- » **Screening of an isolated target** may be chosen, for example, to search for molecules modulating an enzymatic activity *in vitro* (most commonly, a search for inhibitors) (KNOCKAERT *et al.*, 2002a, 2002b). This approach first of all permits the identification of molecules that act well on the chosen target at the molecular level, but it does not enable any judgement to be made about what effects might take place within cells or tissues. It is therefore necessary, in a second step, to characterise the targeting of the molecule at the cellular level, with all the difficulties and surprises that this may reveal (lack of selectivity, poor bioavailability, metabolising, rejection etc.).
- » **Cellular screening** can be approached in a variety of ways. A **cell model is said to be homologous** when the cells experimented with have not been genetically modified. The screening is thus based on the detection of particular cellular properties, for example, the level of intracellular calcium by using fluorescent probes (SULLIVAN *et al.*, 1999). **Screening in a recombinant cellular model** (containing genetic constructs) also called **heterologous**, rests on the exploitation of the gene(s) introduced into the cell (for example, the yeast two-hybrid technique, which allows detection of interactions between pairs of proteins, or the use of reporter genes in transfected cells). Screening relies therefore on indirect measurements of biological activity, ‘reported’ by the proteins introduced by genetic engineering. The processes intermediate between biological activity and the measurement (interacting proteins, heterologous gene expression systems) can be **affected by the molecules present**. A thorough analysis of the results is consequently necessary so as to identify any artefacts generated.
- » **Phenotypic screening**, practised on cells in culture or on whole organisms (chapters 8 and 9), permits the selection of molecules capable of interfering with a given biological process, by observing a phenotype linked to the perturbation that one wishes to elicit (STOCKWELL *et al.*, 1999; STOCKWELL, 2000). In this case, the complementary steps will involve identifying the molecular target of the active substance.

Choice based on technological criteria

One could state as a primary principle that a ‘good’ assay must fulfil a certain number of criteria: precision, simplicity, rapidity, robustness and reliability. We shall see further on that there is actually a way of evaluating some of these criteria, by calculating a statistical factor. At the technological level, the principal choice concerns whether to use an homogenous or heterogenous phase assay.

- » **Homogenous phase assay (mix and read or mix and measure)**, consists of directly measuring the reaction product in the reaction mix, without any separating step. This procedure is ideally suited to high-throughput screening since it is both simple and fast. Homogenous phase assays generally require labelled molecules,

the cost, preparation and reliability of which must be taken into consideration (ex. 3.1).

Example 3.1 - homogenous phase enzymatic assays

The principle of an assay may involve measuring the hydrolysis of a substrate that displays a characteristic absorbance of the fluorescence signal only after hydrolysis (for example, para-nitro-phenyl-phosphate for a phosphatase, or a peptide labelled with amino-methyl-coumarine for a protease). Technologies exist, such as polarisation or fluorescence, scintillation proximity or energy transfer, that are particularly suited to the homogenous phase.

- » **Heterogenous phase assay** involves several steps for the separation of the reagents (filtration, centrifugation, washing etc.) thus making the assay longer and more complex, but sometimes more reliable. The best example to illustrate this procedure is the ELISA test (Enzyme-Linked Immuno Sorbent Assay, see glossary), which requires numerous washing steps and changes of reagent (ex. 3.2).

Example 3.2 - targets for which it is possible to perform homogenous and heterogenous phase assays

- ▶ To screen for molecules acting upon a **kinase**, it is possible to design an **homogenous phase assay**, using a fluorescent substrate, the phosphorylation of which generates an epitope recognised by a specific antibody: subsequent detection by fluorescence polarisation of the phosphorylated antibody-substrate complex presents different characteristics from the non-phosphorylated fluorescent substrate. Alternatively, **in the heterogenous phase assay**, it is possible to use the natural substrate of the kinase, adenosine triphosphate (ATP), whose phosphate is radioactively labelled: the phosphorylated substrate is detected, after filtration, by measuring the radioactive count.
- ▶ To screen for molecules acting upon the **binding of a ligand to its receptor**, it is possible to design an **homogenous phase assay**, by immobilising membranes containing the receptor of interest at the bottom of a microplate well containing wheatgerm agglutinin, and then incubating this preparation with the radioactively labelled ligand. Detection by proximity scintillation permits measurement of ligand binding. Alternatively, **in the heterogenous phase assay**, the radioligand can be detected after filtration, by counting the radioactivity.

3.2.2. SETTING UP THE ASSAY

This step consists of validating experimentally the choices that have been made. The assay is carried out manually in the format which has been selected for performing the screen (in general 96 or 384 wells), in conditions as close as possible to those to be done with the platform (volumes, order of dispensing and mixing of the reagents, reaction temperatures, incubation times etc.).

The preparation of the biological material necessary for setting up and carrying out the automated screening must be done with extreme care in terms of its homogeneity, traceability and quality (see chapter 7). It is impossible to list all of the practical advice that is completely generally applicable. Below, however, we do outline some important aspects to be considered during the development of an assay.

» **The preparation of an isolated protein target**

Most often the proteins are produced in cellular systems in the laboratory (bacteria, yeast) after introduction of the corresponding gene by genetic engineering. It is possible to add extra peptide segments to the extremities of the natural protein sequence: these segments are called tags, compatible with the detection methods that one wishes to employ. In the case of membrane receptors, the assays are frequently carried out with membrane preparations from cells over-expressing the receptor of interest.

» **The specificity of a substrate or an enzymatic activity**

For different enzyme families (kinases, proteases, phosphatases) commercially available generic substrates exist (often sold in kits). In all cases it is better to work with a **specific substrate**, which permits selection, *a priori*, of more specific hits.

» **The specificity of an antibody**

The question of specificity is also critical in the choice of antibodies as detection tools in assays like ELISA or Cytoblot (see glossary). These restrictions apply and require even more acuity than when carrying out an assay in which the target is not purified (cell extracts, whole cells).

» **The relevance of the cellular model**

This point is to connect to the pharmacological aspect of the procedure; it is indispensable to have a cellular model suited to the biological, physiological or physiopathological question that is being asked (HORROCKS *et al.*, 2003). The model used for the primary screening can serve, for example, for the first pharmacological tests immediately following the screening, for the determination of the EC₅₀, for instance, or for testing the specificity of molecules.

» **The experimental conditions** (WU *et al.*, 2003)

Depending on the **equipment** in the screening platform, it is necessary to determine carefully:

- › the most appropriate material: for example, with microplates, it is important to test different makes in order to find the best signal-to-background noise ratio (there are very large differences in the quality of materials on offer); the choice must be suitable for the apparatus to be used for measuring the signal in the platform,
- › the volumes to be transferred, respecting the buffer conditions and the reagent concentrations suited to the kinetic parameters of the reaction (chapter 5),
- › the incubation times and temperatures compatible with the sequence of operations in the robot.

The experimenter must never hesitate to explore several leads in parallel, different genetic constructions for the expression of recombinant proteins, different labels, different cell models, several substrates, several differentially labelled antibodies and so on, as **each novel target represents a unique case for which appropriate conditions must be established.**

The **solvent** for solubilising molecules in a chemical library, in general DMSO (dimethylsulphoxide, see chapter 1), may interfere with the assay. The tolerance of the assay to DMSO must therefore be evaluated at the concentration envisaged for screening, and indeed adjusted to a minimum if the assay is not sufficiently robust. Furthermore, the **nature of the chemical library** to be screened must be taken into account (see chapter 2). It is necessary to know the number of compounds for testing, and whether or not these compounds are likely to interfere with one or more steps in the designed protocol (*i.e.* with the detection method or the biological assay itself).

Once all of these parameters have been carefully examined, the biological assay can be set up to perform the screening under the best conditions.

3.2.3. VALIDATION OF THE ASSAY AND AUTOMATION

- » **The validation of a biological assay requires the use of reference molecules.** In order to be sure of working in the right conditions for observing the expected effect (for example, inhibition of activity) it is **indispensable** to have reference values. These reference values could be obtained with known molecules, less specific than those under investigation, but allowing the assay to be calibrated (HERTZBERG and POPE, 2000; FERNANDES, 2001).
- » **The statistical reliability of a biological assay is evaluated by calculating the Z' factor.** Miniaturisation aims for the economy of time (by speeding up the work rate) and material costs (by a reduction of the products and reactants). These objectives do not permit duplication of each assay. The calculation of the **Z' factor** was proposed for measuring the performance of assays in microplates (ZHANG *et al.*, 1999). This factor takes into account at least 30 values from the minimum (conditions without enzyme, for example) and 30 values from the maximum (activity determined in the screen's buffer and solvent conditions), which serve to determine the 100% activity level and consequently permit calculation of the percentage of inhibition, or possibly activation, of the molecules screened (see the definition of the controls for bioactivity and bio-inactivity in chapter 1). The Z' factor takes into account the standard deviation (σ) and the means (μ) of the maxima (h) and minima (l). It assumes that these minima and maxima values obey the Normal distribution law:

$$Z' = \frac{1 - (3\sigma_h + 3\sigma_l)}{\mu_h - \mu_l}$$

The value of Z' lies between 0 and 1. An assay is considered to be reliable only if Z' is **greater than 0.5**.

Beware, the Z' factor is indicative of experimental quality, of the reproducibility of the test and of its robustness, but provides no indication of the biological relevance of the assay. A 'good' test according to the criterion of the Z' factor, with an unsuitable cellular model, using a less specific substrate, with

poorly chosen reference molecules will lead to ‘bad hits’. The quality of the hits selected during screening is evaluated by the confirmation rate (see below).

- » **The cost and feasibility on a large scale** must be taken into account very early on in the process paying attention notably to the possibilities for the supply of materials and biological reactants (recombinant proteins; cell lines to be established and/or amplified), chemicals (substrates to be synthesised) and consumables. It is therefore important, on the one hand, to ensure the availability of batches of homogenous reagents and materials for the entire screening project, without neglecting the confirmation experiments. On the other hand, it is necessary to explore the stability of the reagents under the screening conditions (while taking into account the time, delays, and temperatures compatible with the programming of the automated assay as a whole).

3.3. THE CLASSIC DETECTION METHODS

Practically all of the current detection methods, from absorbance measurements to confocal microscopy, exist in microplate format and are therefore compatible with high-throughput work. However, techniques such as surface plasmon resonance (SPR) or nuclear magnetic resonance (NMR) still remain for the time being a little bit separate.

The principal qualities required for detection are sensitivity of the method and robustness of the signal (to limit positive or negative interference by the compounds of the chemical library).

We shall present here a non-exhaustive list of the principal detection methods more particularly dedicated to high-throughput screening (table 3.1), notably for homogenous phase assays. The reader may also like to refer to the book edited by Ramakrishna SEETHALA and Prabhavathi FERNANDES (2001) and to the reviews by EGDELING *et al.* (2003) and JÄGER *et al.* (2003) for fluorescence-based techniques. The *References* section of this chapter provides a number of other citations that the interested reader may consult for the detail of measurement methods.

3.4. THE RESULTS

3.4.1. THE SIGNAL MEASURED: INCREASE OR DECREASE?

During a random search for active molecules, on a large scale, the results of the screening provide the first indication of any activity. At this point, the analysis will reveal whether **the effect sought exhibits a reduction rather than a gain in signal strength, or vice versa**. The most commonly studied effects are with inhibitors; in this case, if the bioactivity manifests as a drop in signal, **false positives** (see chapter 1) may result, for example, from an undispensed reagent.

When the search for a gain of function proves to be biologically more relevant, the assay undertaken will most often aim to observe a rise in the signal. This is the case for example when searching for agonists of G-protein coupled receptor (GPCR) or ion channels, very widely explored by the pharmaceutical industry, by exploiting intracellular probes sensitive to the level of cyclic adenosine monophosphate (cAMP), of calcium or to the membrane potential. Even when aiming for a gain of function, false positives resulting for example from the intrinsic fluorescence of molecules in the chemical library, can also be generated.

3.4.2. THE INFORMATION FROM SCREENING IS MANAGED ON THREE LEVELS

- » **On the scale of the well**, it is necessary on the one hand to be capable of **identifying the active wells**: thanks to barcodes, the test plates are all identified, each active well must be able to be related to the wells of the plate containing the compounds. Besides, it is important to **quantify the activity** (results of the biological test) from the signal obtained. The activity is measured in relation to the value of the signal with maximal amplitude, and is thus expressed as a **percentage**.
- » **On the scale of the plate**, it is necessary to be able to control the statistical reliability of the assay, with the Z' factor calculated using at least 30 points for the minimal signal value (obtained for example without the enzyme if the test is an enzymatic assay with the isolated target) and as many for the maximal signal value (obtained with all the reactants in the assay and the solvent for the library molecules). It must be possible to establish EC50 curves with the reference molecules eliciting a biological effect similar to that being researched (for example staurosporin or roscovitine for calibrating kinase assays – see chapter 5).
- » **On the scale of a campaign**, the results obtained over several days have to be normalised. The selection of hits must be harmonised by considering **the set of results** (*i.e.* taking into account potential drifts in the signal from one day to the next, therefore by normalising using the reference molecules and controls).

This point is not trivial and the statistical model for the standardisation operation must be established according to the principle of the test and the type of signal variations potentially observed (see also chapter 4).

To select the hits, an **activity threshold** (or **cut-off**) is defined, expressed as a percentage of the activity. This concept of cut-off is fundamental since it **directly conditions the number of hits selected**. In practice, the procedure generally consists of setting beforehand a maximum number of hits (0.1 to 0.5% most often) depending on the facilities available for working with these molecules with respect to the **identification of the active compound** (if screening had been carried out with a mixture), and to the chemical **confirmation** (new preparation of the identified molecule, tested again in the same assay as in the primary screening). This last step can be manually performed if it involves a restricted number of molecules.

Table 3.1 - The most frequently used detection methods, suited to high-throughput screening

Method	Principle	Advantages	Disadvantages
Fluorescence			
Total fluorescence <i>Molecular Probes</i>	excitation at a wavelength (WL), the excitation wavelength, and measurement at a 2 nd WL, the emission wavelength, higher than the first	▶ sensitive ▶ easy to carry out ▶ very wide choice of fluorescent molecules	▶ not very robust with respect to interference
FP (Fluorescence Polarisation) <i>Panvera</i>	anisotropy : when a fluorescent molecule is excited by polarised light, the degree of polarisation of the emitted light corresponds to its rotation (proportional to its mass)	▶ sensitive ▶ easy to carry out ▶ compatible with the homogenous phase ▶ particularly well suited to 'small ligand/macromolecule' interactions	▶ cost of reactants and need for dedicated equipment ▶ auto-fluorescence of the compounds can interfere
FRET (Förster Resonance Energy Transfer)	2 fluorescent molecules having spectral characteristics such that emission of the first is quenched by the second (e.g., the CFP-YFP pair)	▶ measure the quenching of the emission of the donor ▶ low interference ▶ compatible with the homogenous phase ▶ study of protein/protein interactions having a proximity of 10 to 100 Å	▶ not very robust with respect to interference if the acceptor's emission is measured
HTRF (Homogenous Time- Resolved Fluorescence) <i>CISBIO</i>	same principle as before, but the fluorescent markers (rare earth metals and allophycocyanines) permit measurements spaced apart in time	▶ measurements spaced out in time (100 to 1000 µs) due to a longer half-life of the emission ▶ permits the elimination of the natural fluorescence of compounds with a short half-life (nanosecond)	▶ tagging of molecules (costly reagents) and need for dedicated equipment
TR-FRET (Time-Resolved Fluorescence Resonance Energy Transfer)			
FMAT™ (Fluorimetric Microvolume Assay Technology) <i>Biosystems</i>	quantitative measurement of fluorescence suited to functional assays with cells, determination of cytotoxicity	▶ high sensitivity ▶ possible to use in homogenous phase assays with beads	▶ acquisition time of several minutes for 384-well plates ▶ very large data files

Method	Principle	Advantages	Disadvantages
Radioactivity			
SPA (Scintillation Proximity Assay) <i>Perkin Elmer</i>	beads coated with scintillant permit the amplification of the radioactivity, which is thus detected over a very short distance	▶ high sensitivity ▶ exist in microplates or beads in suspension	▶ disadvantages linked to the use of radioelements ▶ detection time (10 to 40 minutes for 96- or 384-well plates)
CytostarT <i>GE Healthcare</i>	same principle, suited to cells cultured in transparent-bottomed plates	▶ measurement of the incorporation of a radiomarker, metabolic tagging ▶ possible to use in the homogenous phase with soft beta emitters	▶ disadvantages linked to the use of radioelements ▶ detection time (10 to 40 minutes for 96- or 384-well plates)
Luminescence			
Chemoluminescence	use of a chemical substrate generating a signal and its auto-amplification, allowing high sensitivity	▶ no excitation (therefore no interference with compounds from the chemical library)	▶ detection by luminometer or scintillation counter
Bioluminescence <i>Perkin Elmer</i>	use of biological substrate generating a signal and its auto-amplification, allowing high sensitivity	▶ quantitative ▶ can be used in luciferase 'reporter gene' systems ▶ cellular assays	▶ requires molecular engineering ▶ use of costly reagents
BRET (Bioluminescence Resonance Energy Transfer)	Based on the transfer of bioluminescence by using 'coupled' enzymes (e.g., <i>Renilla</i> & firefly luciferases)	▶ high sensitivity ▶ no interference ▶ cellular assays	▶ requires molecular engineering ▶ use of costly reagents
ALPHAscreen (Amplified Luminescence Proximity Homogenous Assay) <i>Perkin Elmer</i>	excitation at 680 nm of donor/acceptor beads permitting the transfer of singlet oxygen (half-life of 4 μ s) at a proximity of < 200 nm and a measurement of the WL emitted at 520-620 nm (0.3 μ s)	▶ measurement of proximity in homogenous phase assay (protein/protein interactions, detection of epitopes by specific antibodies) ▶ no interference with the natural fluorescence since WL2 < WL1	▶ high sensitivity, but costly reagents ▶ need for dedicated equipment ▶ interference with singlet oxygen from the compounds in the chemical library

A point that we have not discussed in this chapter is the concentration of molecules. This concentration is perhaps unknown (for extracts from natural substances, for example) or controlled with more or less variability (in mass or molarity). The assays are conducted in a constant volume of added molecules. The comparison of molecules implies that the assay can be reproduced *a posteriori* with variable concentrations of molecules, permitting evaluation of the EC₅₀ (chapter 5).

3.4.3. PHARMACOLOGICAL VALIDATION

Pharmacological validation consists of determining the EC₅₀ value (see chapter 5) of each active molecule. Only those molecules presenting a **dose-effect** with an efficacy more or less comparable to that of the reference molecules will be kept.

Only after all of these steps can there be **confirmed hits**, which are potentially interesting if the determined EC₅₀ values are compatible with later studies (*in vivo* assays, possible optimisations, QSAR – see chapters 12 and 13).

3.5. DISCUSSION AND CONCLUSION

The time-frame for establishing an assessment in terms of high-throughput screening for the discovery of candidate drugs is extremely long and many laboratories in this field have existed for no more than 5 to 10 years. One study involving 44 laboratories employing high-throughput screening has generated an increasing number of **lead** molecules (FOX *et al.*, 2002). A lead is defined as a hit that is confirmed by more than one *in vitro* assay and if possible *in vivo*, which proves a significant biological activity in relation to the target; to be a lead, a compound must permit a structure-function relationship to be established. On average, over one year is required to know whether or not a hit could become a drug candidate.

Since 2002 numerous comparative studies have been published. They address a concern, which is to evaluate the **potential bias** introduced into the results by screening **methods**, as well as the following question: **do different versions of the same screening assay enable identification of the same compounds?** Large pharmaceutical groups have set about the task of answering this, by testing a significant sample of their chemical libraries (several tens of thousand molecules) in different conditions. The results are quite surprising, they sometimes reveal great consistency (HUBERT *et al.*, 2003), and other times, in contrast, significant divergence (SILLS *et al.*, 2002). Nevertheless, in all cases the chemical families identified by different methods are the same. This type of study has the advantage of eliminating **false-positives** which are most often directly linked to the technology used (interference, attenuation or quenching, intrinsic fluorescence of the compounds etc.) and **false-negatives**. But artefacts are not merely present at the detection stage, and the miniaturisation protocols and the work by MC GOVERN *et al.* (2002) signal caution about the nature of small molecules arising from screens (using enzymes). The hits are often not very specific, displaying EC₅₀ values of the order of micromolar and their development into medicines may be compromised by their propensity to form micellar or vesicular aggregates.

Two important messages should be remembered:

- it is necessary to remain prudent in the evaluation of results, as long as the pharmacological results are not convincing,

- › the methodological and technological problems presented by the miniaturisation of an assay ought never to obscure the biological question.

3.6. REFERENCES

- DREWS J. (2000) Drug discovery: a historical perspective. *Science* **287**: 1960-1964
- EGGELING C., BRAND L., ULLMANN D., JAGER S. (2003) Highly sensitive fluorescence detection technology currently available for HTS. *Drug Discov. Today* **8**: 632-641
- SEETHALA R., FERNANDES P.B. (2001) *Handbook of drug screening*. New York-Basel, Marcel Dekker inc.
- FERNANDES P.B. (1998) Technological advances in high throughput screening. *Curr. Opin. Chem. Biol.* **2**: 597-603
- FOX S., WANG H., SOPCHAK L., FARR-JONES S. (2002) High throughput screening 2002: moving toward increased success rates. *J. Biomol. Screen.* **7**: 313-316
- GOPALAKRISHNAN S.M., MAMMEN B., SCHMIDT M., OTTERSTAETTER B., AMBERG W., WERNET W., KOFRON J.L., BURNS D.J., WARRIOR U. (2005) An offline-addition format for identifying GPCR modulators by screening 384-well mixed compounds in the FLIPR. *J. Biomol. Screen.* **10**: 46-55
- HAGGARTY S.J., MAYER T.U., MIYAMOTO D.T., FATHI R., KING R.W., MITCHISON T.J., SCHREIBER S.L. (2000) Dissecting cellular processes using small molecules: identification of colchicine-like, taxol-like and other small molecules that perturb mitosis. *Chem. Biol.* **7**: 275-286
- HERTZBERG R.P., POPE A.J. (2000) High throughput screening: new technology for the 21st century. *Curr. Opin. Chem. Biol.* **4**: 445-451
- HORROCKS C., HALSE R., SUZUKI R., SHEPHERD P.R. (2003) Human cell systems for drug discovery. *Curr. Opin. Drug Discov. Dev.* **6**(4): 570-575.
- HUBERT C.L., SHERLING S.E., JOHNSTON P.A., STANCATO L.F. (2003) Data concordance from a comparison between filter binding and fluorescence polarization assay formats for identification of ROCK-II inhibitors. *J. Biomol. Screen.* **8**: 399-409
- JAGER S., BRAND L., EGGELING C. (2003) New fluorescence techniques for high-throughput drug discovery. *Curr. Pharm. Biotechnol.* **4**: 463-76.
- JOHNSTON P.A., JOHNSTON P.A. (2002) Cellular platforms for HTS: three case studies. *Drug Discov. Today* **7**: 353-363
- KEMP D.M., GEORGE S.E., KENT T.C., BUNGAY P.J., NAYLOR L.H. (2002) The effect of ICER on screening methods involving CRE-mediated reporter gene expression. *J. Biomol. Screen.* **7**: 141-148
- KNOCKAERT M., GREENGARD P., MEIJER L. (2002b) Pharmacological inhibitors of cyclin-dependent kinases. *Trends Pharmacol. Sci.* **23**: 417-425

- KNOCKAERT M., WIEKING K., SCHMITT S., LEOST M., GRANT K.M., MOTTRAM J.C., KUNICK C. MEIJER L. (2002a) Intracellular targets of paullones: identification following affinity purification on immobilized inhibitor. *J. Biol. Chem.* **277**: 25493-25501
- KNOCKAERT M., MEIJER L. (2002) Identifying in vivo targets of cyclin-dependent kinase inhibitors by affinity chromatography. *Biochem. Pharmacol.* **64**: 819-25
- MCGOVERN S.L., CASELLI E., GRIGORIEFF N., SHOICHET B.K. (2002) A common mechanism underlying promiscuous inhibition from virtual and high-throughput screening. *J. Med. Chem.* **45**: 1712-1722
- MOORE K., REES S. (2001) Cell-based versus isolated target screening: how lucky do you feel? *J. Biomol. Screen.* **6**: 69-74
- REVAH F. (2002) La révolution du médicament: de 1040 à 10 molécules. *Sciences et Vie* **218**: 18-27
- SILLS M.A., WEISS D., PHAM Q., SCHWEITZER R., WU X, WU J.J. (2002) Comparison of assay technologies for a tyrosine kinase assay generates different results in high throughput screening. *J. Biomol. Screen.* **7**: 191-214
- STOCKWELL B.R. (2000) Frontiers in chemical genetics. *Trends Biotechnol.* **18**: 449-455
- STOCKWELL B.R., HAGGARTY S.J., SCHREIBER S.L. (1999) High throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications. *Chem. Biol.* **6**: 71-83
- SULLIVAN E., TUCKER E.M., DALE I.L. (1999) Measurement of $[Ca^{2+}]$ using the Fluorometric Imaging Plate Reader (FLIPR). *Methods Mol. Biol.* **114**: 125-133
- VON LEOPRECHTING A., KUMPF R., MENZEL S., REULLE D., GRIEBEL R., VALLER M.J., BUTTNER F.H. (2004) Miniaturization and validation of a high-throughput serine kinase assay using the AlphaScreen platform. *J. Biomol. Screen.* **9**: 719-725
- WILLIAMS C. (2004) cAMP detection methods in HTS: selecting the best from the rest. *Nat. Rev. Drug Discov.* **3**: 125-135
- WU G., YUAN Y., HODGE C.N. (2003) Determining appropriate substrate conversion for enzymatic assays in high-throughput screening. *J. Biomol. Screen.* **8**: 694-700
- YOUNG K.H., WANG Y., BENDER C., AJIT S., RAMIREZ F., GILBERT A., NIEUWENHUIJSEN B.W. (2004) Yeast-based screening for inhibitors of RGS proteins. *Methods Enzymol.* **389**: 277-301
- YOUNG K., LIN S., SUN L., LEE E., MODI M., HELLINGS S., HUSBANDS M., OZENBERGER B., FRANCO R. (1998) Identification of a calcium channel modulator using a high throughput yeast two-hybrid screen. *Nat. Biotechnol.* **16**: 946-950
- ZHANG J.H., CHUNG T.D., OLDENBURG K.R. (1999) A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screen.* **4**: 67-73

Chapter 4

THE SIGNAL: STATISTICAL ASPECTS, NORMALISATION, ELEMENTARY ANALYSIS

Samuel WIECZOREK

4.1. INTRODUCTION

The **elementary analysis of raw data** coming from automated pharmacological screening (*i.e.* the bioactivity signals) aims to identify bioactive molecules (called candidate hits) that will then be subjected to more in-depth testing. This selection is made by setting a bioactivity threshold and the interesting molecules are therefore identified purely on the basis of the bioactivity signal. This measure represents the most concise **information** about the bioactivity of compounds in a chemical library and is as such particularly precious.

During automated screening, the bioactivity signals are characterised by variability and uncertainty due to measurement errors (fig. 4.1), which may have a biological, chemical or technological origin. These errors give rise to **false-positives** (molecules wrongly identified as bioactive) as well as **false-negatives** (molecules identified as bio-inactive despite having actual bioactivity). These phenomena degrade the quality of the selection of bioactive molecules.

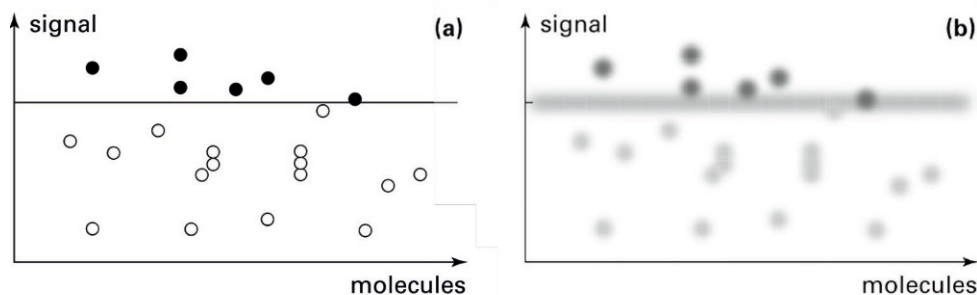


Fig. 4.1 - A threshold for the measured signal permits selecting the molecules of interest

(a) Ideal case - Measurements without errors: the signals and the bioactivity threshold are precise. **(b)** Real case - Measurements marred by errors: the signals as well as the bioactivity threshold are imprecise.

The validity of the conclusions drawn from the elementary analysis depends on the quality of the underlying raw data. Would pre-processing of the raw signals help to improve the precision of the information and to limit the influence of errors on the results?

4.2. NORMALISATION OF THE SIGNALS BASED ON CONTROLS

The variability within the data arising from screening complicates the identification of bioactive molecules. Considering the whole set of data for a given screening, the selection is carried out by using a **cut-off** for the raw signal, which is not always comparable from one plate to another. To overcome this difficulty, the **traditional approach of normalisation** (by the **percentage of inhibition**), based on the means of the control values for bioactivity and bio-inactivity, functions correctly and remains widely used. If the side effects are not too widespread and if the controls are inspected for discrepancies and **aberrant values**, then normalisation by the percentage of inhibition is often valid (BRIDEAU *et al.*, 2003).

4.2.1. NORMALISATION BY THE PERCENTAGE INHIBITION

The Percentage of Inhibition (PI) scales the raw bioactivity signal to a value lying between 0 and 1 (and multiplied by 100 to put it on a percentage scale). For a plate p , the **percentage inhibition** PI_p^i of the signal measured in a well with index i represents its relative distance from the mean of a set of control bioactivity values. Let I^{act} and I^{inact} be the respective means of a set of controls for bioactivity and bio-inactivity and I_p^i , the signal from a molecule measured in a well with index i in a plate p , the normalised signal is defined as follows:

$$PI_p^i = \frac{I_p^i - I^{inact}}{I^{act} - I^{inact}} \quad (\text{eq. 4.1})$$

The normalised signal is interpreted thus: the closer the raw signal measured is to the mean of the controls used for bio-inactivity, the more the percentage inhibition approaches 0; conversely, the closer the signal approaches the mean of the controls used for bioactivity, the more the normalised signal value tends towards unity.

Note that it is entirely possible to observe molecules for which the raw signal exceeds that of the controls (percentage inhibition < 0 and > 100).

4.2.2. NORMALISATION RESOLUTION

The normalisation presented in the preceding section is based on a set of controls. This particular set, termed the **normalisation window** (fig. 4.2), defines the controls for bioactivity and bio-inactivity whose means permit the calculation of the percentage inhibition (eq. 4.1). The width of this window, *i.e.* the number of con-

trols taken into consideration, is directly linked to the idea of **resolution of the normalisation**, which defines, in a sense, the level of detail that will be smoothed arising from normalisation.

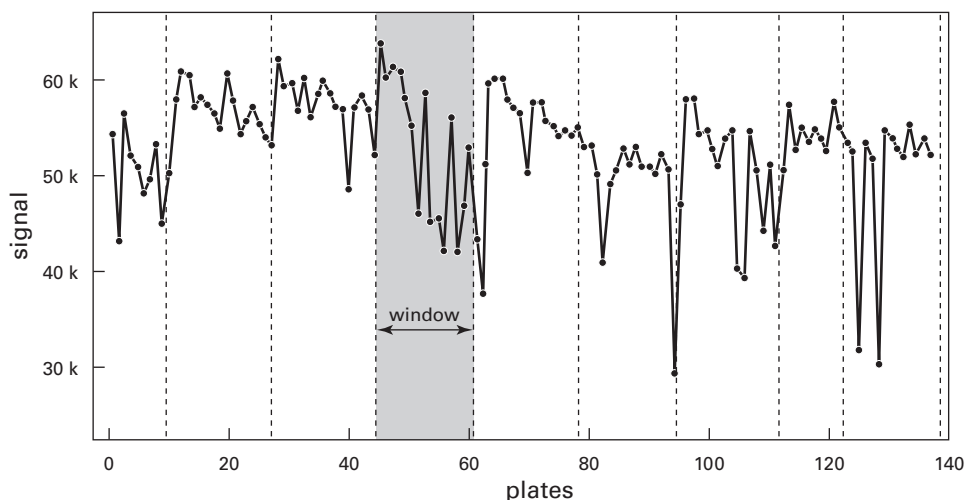


Fig. 4.2 - The normalisation window allows the resolution of the normalisation to be controlled

The choice of window is guided by observing the phenomena that perturb the results of screening. For example, if the signals are perturbed in a similar manner on each day of screening, then the signal of a molecule measured on day d will have to be normalised with a resolution equivalent to a whole day of screening. In other words, the controls considered will be those measured on day d . Routinely, one would choose a window equal to the size of a plate (fig. 4.3).

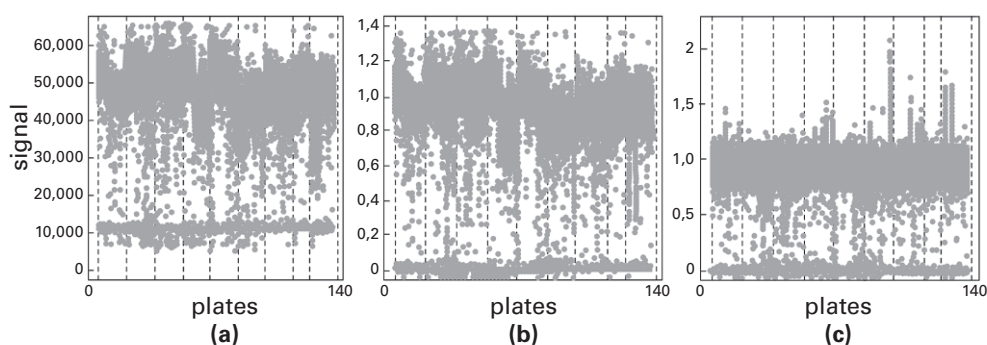


Fig. 4.3 - Example of a normalisation taking into account signal discrepancies at the level of plates

The vertical dashed lines delimit the days of screening. **(a)** Raw signals measured over the course of screening 140 plates. **(b)** The signals were normalised with a window equal to the size of the totality of the screening controls. Here, the normalisation is of poor quality: daily drifts can still be distinguished. **(c)** The normalisation window is equal to a plate, the normalisation that follows from this no longer shows large-scale drifts.

4.2.3. ABERRANT VALUES

Over the course of measuring signals, it is possible to observe some values that deviate significantly from the majority of the other signals: these particular measurements are designated as being **aberrant**; they are likely to arise from measurement errors. The presence of such signals skews the calculation of the mean and as a consequence, that of the previously described normalisation.

In order to surmount this problem, aside from **manual suppression**, it is recommended to use **robust estimators** that behave in a constant manner even when they are subjected to non-standard conditions. This means that, in spite of the presence of data relatively removed from the ideal case, the response of the system remains hardly disturbed. The **median** and the **α -censored mean** are examples of robust estimators; they are less influenced by aberrant values than the **mean**. The class of L-estimators (RAPACCHI, 1994; WONNACOTT *et al.*, 1998) is defined as follows:

» **Definition 1 (weighted mean)**

Let $x_0, x_1, \dots, x_n$ be the values in a sample of size n where x_i is the i^{th} value in increasing order from the sample (we have $x_1 \leq x_2 \leq \dots \leq x_n$). Let $a_1, a_2, \dots, a_n$ be real numbers where $0 \leq a_i \leq 1$, for $i = 1, 2, \dots, n$ and $\sum a_i = 1$, the weighted mean

is defined by:

$$T = \sum_{i=1}^n a_i x_i \quad (\text{eq. 4.2})$$

This definition characterises a class of estimators called *L-estimators* that are distinguished by the values of the coefficients a_i .

» **Definition 1 (median)**

The median is the L-estimator that includes the central value if n is odd, or the mean of the two central values if n is even.

$$\begin{aligned} \text{If } n = 2p + 1 \quad \text{so } a = & \begin{cases} 1 & \text{if } i = p + 1 \\ 0 & \text{if } i \neq p \end{cases} \\ \text{If } n = 2p \quad \text{so } a = & \begin{cases} 1/2 & \text{if } i = p \text{ or } i = p + 1 \\ 0 & \text{if not} \end{cases} \end{aligned} \quad (\text{eq. 4.3})$$

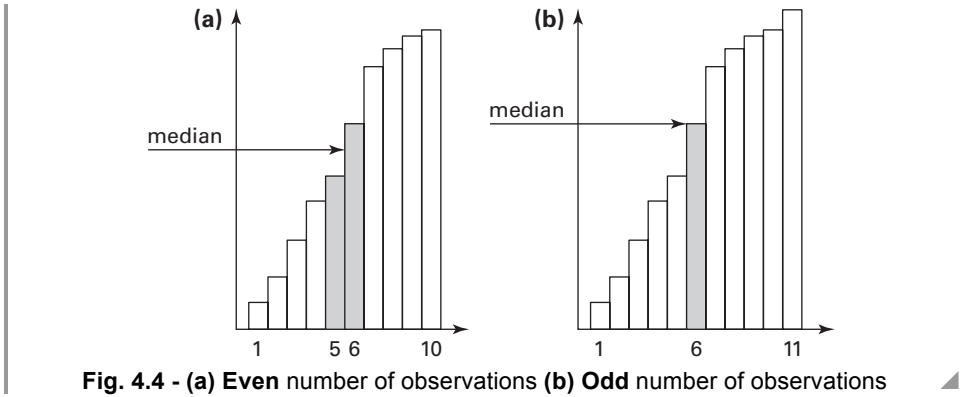
The median is the point that divides the distribution of a series of observations (ordered from the smallest to the largest) into two equal parts.

Example 4.1 - the median

Taking a sample of size $n = 10$, where $x_i = 1, \dots, 10$ (fig. 4.4a), the median value $m(10)$ is given by:

$$m(10) = \frac{1}{2}(x_5 + x_6)$$

With a sample of size 11 (fig. 4.4b), the median value is: $m(11) = x_{(6)}$.



» **Definition 3 (α -censored mean)**

Let α be a real number where $0 \leq \alpha \leq 0.5$, the α -censored mean, $T(\alpha)$, is a weighted mean that automatically neglects the extreme values. The weights a_i are such that:

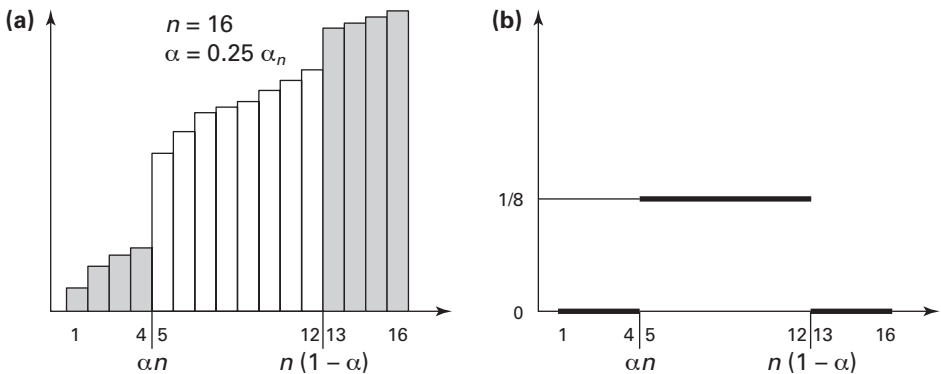
$$a_i = \begin{cases} 0 & \text{if } i \leq g \text{ or } i \geq r \\ \frac{1}{n(1-2\alpha)} & \text{if } g+1 \leq i \leq r-1 \text{ where } g = \alpha n \text{ and } r = n(1-\alpha) \end{cases} \quad (\text{eq. 4.4})$$

It is calculated for a sample of the data by omitting a proportion α of the smallest values and another proportion α of the largest values and then calculating the mean of the remaining data. The parameter α indicates the number of extreme points in the sample to leave out. The smaller the value of α , the fewer the points left out. For $\alpha = 0$, the α -censored mean is equivalent to the 'classical' mean.

Example 4.2 - the α -censored mean

Let there be a sample of size 16, where $x_i = 1, \dots, 16$. By choosing $\alpha = 0.25$, we remove half of this sample (one quarter at the beginning and one quarter at the end of the distribution).

Thus, $\alpha_n = 0.25 \times 16 = 4$ and: $T(0.25) = \frac{1}{8} \sum_{i=5}^{12} x_i$



- (a) Grey bars, the ordered statistics excluded from the calculation of the mean.
 (b) Values of the coefficient α_i . This coefficient is nought (= 0) for extreme values.

4.3. DETECTION AND CORRECTION OF MEASUREMENT ERRORS

As with all physical measurements, the value of the signal measured is generally different from the true value of the signal emitted. This difference, termed the **measurement error**, is never precisely known, and so it is nearly impossible to correct for measurement error in order to find the real value. In the context of the identification of bioactive molecules by using cut-offs, these errors significantly increase the rate of false positives and negatives. One possible method to deal with this problem consists of lowering the cut-off value for the bioactivity threshold with the aim of reducing the rate of false negatives, which tends, however, to increase the rate of false positives in an unquantifiable manner. Another solution involves rather the analysis of measurement errors and then to limit their effects.

In general, these errors can be classified into two categories depending on their origin: *systematic errors* (or bias) and *random errors* (or statistical errors). This classification can be extended to the context of HTS signals by *semi-systematic errors*.

» **Random errors** crop up in a totally random way and even if their origin is known, it is not possible to know either their value or their sign. Random error, E_a , is the difference between the result of a measurement M_i and the mean $\bar{M}$ of the n measurements repeated when n tends to infinity and when these measurements are obtained under reproducible conditions:

$$E_a = M_i - \bar{M} \quad (\text{eq. 4.5})$$

Repetition of the experiments enables reduction of these errors but can in no case eliminate them.

Example 4.3 - random errors

A random error can result from the chemical instability of molecules, from the state of the biological material used (e.g. cells in different stages of their cycle), from reading a perturbed signal (e.g. heterogenous mixture). ▲

» **Systematic errors** are constant errors that overlap random errors and introduce systematically the same shift. The systematic error E_S is the difference between the mean $\bar{M}$ of n repeated measurements, where n tends to infinity (measurements obtained under reproducible conditions), and the measured quantity M_0 :

$$E_S = \bar{M} - M_0 \quad (\text{eq. 4.6})$$

Unlike a random error, a systematic error cannot be reduced by repeating the experiments. However, a careful examination of the series of measurements enables, more often than not, discovery of the source of error and thus its reduction by improving the sequence of processing or by a suitable procedure post-measurement of the signals.

Example 4.4 - systematic errors

Sources of systematic errors include recurring problems with pipetting (e.g. blocked tips) or more generally problems linked to the automation of the platform. ▲

- » **Semi-systematic errors** can appear in the context of automated screening due to the complexity of the experimental protocols. The source of these errors typically seems to be systematic but their behaviour (*i.e.* their values and signs) remains random.

Example 4.5 - semi-systematic errors

The phenomenon of signal *gradients* provoked by certain factors such as the filling of the wells column by column can be observed for each plate (fig. 4.6). To correct for this bias, several approaches may be envisaged (HEYSE, 2002). Among these, we can simply insert during the screening several plates containing only controls in order to model the observed gradients. They are then corrected by standardising the controls.

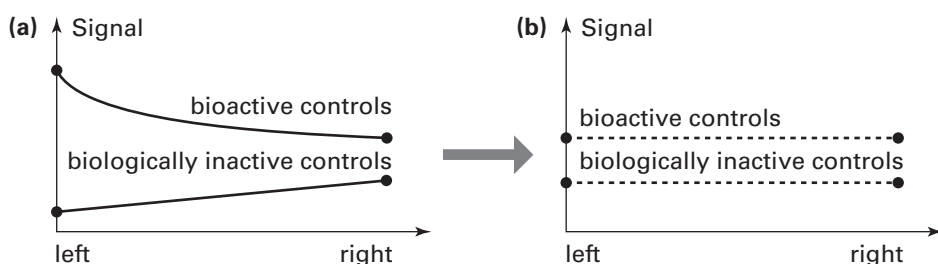


Fig. 4.6 - Representation of an increasing linear signal gradient for the controls for bio-inactivity and a decreasing exponential for the controls for bioactivity between the left and right columns of a plate. The signals for the controls on the left of the plate are more intense. **(a)** Linear gradient before correction. **(b)** Gradient corrected on the basis of the controls for the column C.

Modelling the gradients is a complex task, so the problem can be simplified with the help of hypotheses based on the form of these functions (linear, exponential, etc.) by looking for the possible source of these variations. ▲

4.4. AUTOMATIC IDENTIFICATION OF POTENTIAL ARTEFACTS

Generally, the identification of bioactive molecules is carried out without the experimenter necessarily knowing the logic, if any, behind the distribution of molecules in the plates containing the chemical libraries. In other words, without knowing if a given family of molecules is grouped together in a particular place, for example. By default, one may assume that the molecules are distributed randomly in the plates.

4.4.1. SINGULARITIES

The observation of the position of bioactive molecules in the plates shows that they are not distributed in a uniform manner: some plates contain them, others do not.

Furthermore, within a single plate, some molecules seem to be isolated whereas others are grouped in the same zone (fig. 4.7).

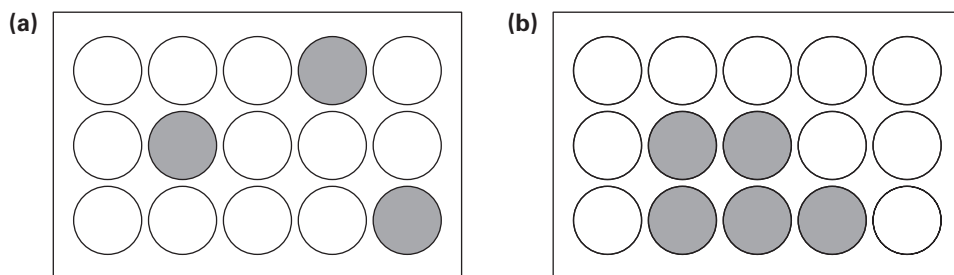


Fig. 4.7 - Positions of the bioactive molecules in one plate
(a) The molecules are distanced from each other in the plate.
(b) The molecules are grouped together in the same zone of the plate.

According to the assumption proposed in the previous paragraph, it could be interesting to study these particular groupings, termed *singularities*, due to the fact that they may be linked to experimental artefacts. Indeed, the probability of observing such singularities in screening plates (calculated using BAYES' rule) shows that this phenomenon would not seem to be due only to chance. The two following hypotheses can explain this:

» **Hypothesis 1 - a localised experimental artefact**

Several artefacts can give rise to these singularities, such as the 'contamination of a well' by a foreign bioactive molecule (a leak from one well into other wells, fig. 4.8a) or indeed 'heterogeneous experimental conditions' in a plate.

» **Hypothesis 2 - the presence of a chemical family**

The presence of structurally similar bioactive molecules in neighbouring wells can also be the reason for singularities. Based on the assumption that the biological activity of a molecule results largely from its structure, one might expect molecules from the same chemical family to display a similar biological activity (fig. 4.8b).

These singularities can be detected automatically with the help of **clustering** techniques (or **non-supervised classification**). The underlying algorithms seek to group together neighbouring wells according to different criteria. Classical approaches use partitioning algorithms, or those based on the notion of density or indeed hierarchical classification techniques. For more detailed information the reader should refer to BERKHIN (2002) and CORNUÉJOLS *et al.* (2002).

4.4.2. AUTOMATIC DETECTION OF POTENTIAL ARTEFACTS

Having detected singularities, a simple solution permits discrimination as a function of their origin. By calculating the **average structural similarity of the molecules in a singularity**, it is possible to evaluate the probability that a group is due to a **local artefact** or to the grouping of a **chemical family**.

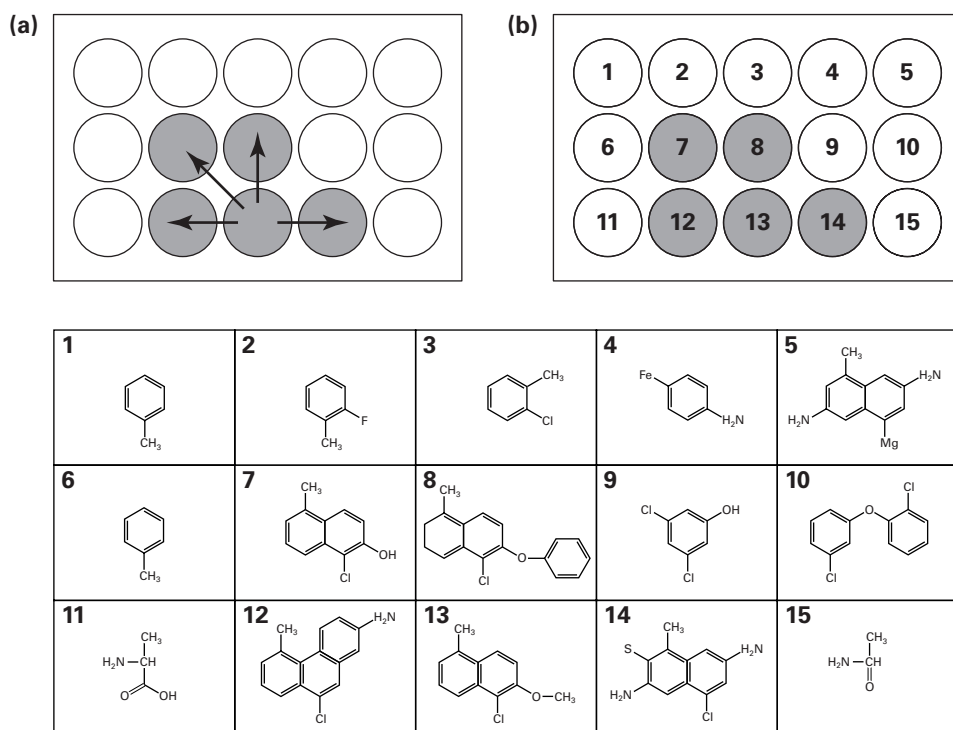


Fig. 4.8 - (a) Contamination between neighbouring wells. Grey wells indicate the supposed presence of a molecule identified as bioactive; the real bioactive molecule having perhaps leaked from the sides. **(b)** Grouping of a family of bioactive molecules. The molecules identified in the grey wells share the same sub-structure, which may explain the bioactivity of the molecules possessing it.

How can the structural similarity between two molecules be measured? This is an important point regarding the exploration of chemical space and is described in more detail in the third part of this book. Here we shall consider in a first simple approach a representation of the molecular structures in two dimensions to be translated into vector form, of which each element having a Boolean value, marks the presence (bit value equal to 1) or the absence (bit value equal to 0) of a particular substructure (also termed *structural keys*, ALLEN *et al.*, 2001). Numerous *distances* (WILLETT *et al.*, 1998; GILLET *et al.*, 1998) permitting measurement of the similarity of the Boolean vectors can be employed. One of these is the **TANIMOTO index**. Letting m_i and m_j be the structural keys of length L of two molecules i and j , and $m_i(l)$ and $m_j(l)$ the values of the Boolean elements with respective indices i and j in each of the keys, the Tanimoto index $S_t(m_i, m_j)$ is defined as:

$$S_t(m_i, m_j) = \frac{\sum_{l=1}^L m_i(l) \cdot m_j(l)}{\sum_{l=1}^L m_i(l) + \sum_{l=1}^L m_j(l) - \sum_{l=1}^L m_i(l) \cdot m_j(l)} \quad (\text{eq. 4.7})$$

This index represents the relationship between the number of bits with a value of 1 (*i.e.* the number of substructures present in the molecules, and carried over in their respective structural keys) common to the two keys and the total number of bits of value 1 in each of the two keys.

The mean similarity SM_C of all the molecules (*i.e.* the number of substructures common to the two molecules) within a singularity is thus defined as the mean of the similarities of each pair of molecules in this cluster. Letting m_i be the structural key of the i^{th} molecule in the cluster C of size M, the mean similarity is written:

$$SM_C = \frac{2}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^M S_i(m_i, m_j) \quad (\text{eq. 4.8})$$

From this notion, it can be deduced that, if SM_C has a high value, then it is more probable that the singularity is due to a family of bioactive molecules being grouped in the plate; conversely, with a low SM_C value, it is more probable that the singularity is due to a local artefact (for example, a bioactive molecule leaking into neighbouring wells).

4.5. CONCLUSION

Automated pharmacological screening concluding with the measurement of bioactivity signals involves complex experimental protocols, which can produce errors in signal measurement. These measurement errors can significantly affect the identification of bioactive molecules because a number of false positives and false negatives are generated. Despite the simplicity or the obviousness of some approaches, the detection and correction of errors are too often neglected. It is very important to be aware of them and to attempt to limit their number, particularly so as not to miss potentially important molecules and to limit the cost of analysis of the wrongly identified molecules. This chapter has highlighted a few approaches that permit an improvement in data precision and an increase in the confidence in the identification of candidate molecules as well as in the interpretation that follows from their analysis. Some of these methods are included in commercially available software for the analysis of screening results. However, a universal method does not exist: the sources of measurement error are different for each screen, and so a careful examination of the results and statistical expertise will orient the experimenter towards the best correction method.

4.6. REFERENCES

- ALLEN B.C.P., GRANT G.H., RICHARD W.G. (2001) Similarity calculations using two-dimensional molecular representations. *J. Chem. Inf. Comput. Sci.* **41**: 330-337
- BERKHIN P. (2002) *Survey of clustering data mining techniques*.
Technical Report Accrue Software, San Jose, California

BRIDEAU C., GUNTER B., PIKOUNIS B., LIAW A. (2003) Improved statistical methods for hit selection in high-throughput screening. *J. Biomol. Screen.* **8**: 634-647

CORNUÉJOLS A., MICLET L. (2002) *Apprentissage artificiel, concepts et algorithmes*. Editions Eyrolles, Paris

GILLET V.J., WILD D.J., WILLETT P., BRADSHAW J. (1998) Similarity and dissimilarity methods for processing chemical structure databases. *Computer J.* **41**: 547-558

HEYSE S. (2002) Comprehensive analysis of high-throughput screening data, *Proceedings of SPIE*, Vol. 4626: 535-547

RAPACCHI B. (1994) *Une introduction à la notion de robustesse*. Centre Interuniversitaire de Calcul de Grenoble

WILLETT P., BARNARD J.M., DOWNS G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**: 983-997

WONNACOTT T.H., WONNACOTT R.J. (1998) *Statistique, Economie - Gestion - Sciences - Médecine*. Editions Economica

Chapter 5

MEASURING BIOACTIVITY: K_i , IC_{50} AND EC_{50}

Eric MARÉCHAL

5.1. INTRODUCTION

Which quantity permits a characterisation of the performance of a bioactive molecule? How can a test be created so as to detect the effect of a molecule on a given target? Are there any general rules to respect? The design of a test is a complex problem dealt with in chapter 3; here we emphasise that, above all, **the target must be a limiting factor in the reaction system**.

How do we know whether a molecule that affects several targets (for example, an inhibitor of different kinases) has a preferred target? Is it more bioactive in one case and less in another? Let us initially discuss the problem for an enzyme or receptor inhibitor. For a molecule interfering for example with an **enzyme, termed Michaelian in ideal conditions**, which is discussed further below, the biochemist makes use of K_i , the **inhibition constant**. When the inhibitor is a competitor of a ligand binding to a **receptor**, the biochemist uses the **IC_{50} (concentration of inhibitor at 50% of the total inhibition)**. Practically, if it is possible to measure the variation in a signal corresponding to the effect of a molecule (at the molecular, functional or phenotypic level), the experimenter will be able to define on a **dose-effect curve** the concentration of molecule for which 50% of the bioactivity is observed. We refer to this as the **EC_{50} (the effective concentration at 50% of the total effect)**. What is the difference between K_i and IC_{50} ? Is the EC_{50} an absolute value? Can we rely on the EC_{50} to qualify a molecule as bioactive? We shall deal with this set of questions in this chapter.

5.2. PREREQUISITE FOR ASSAYING THE POSSIBLE BIOACTIVITY OF A MOLECULE: THE TARGET MUST BE A LIMITING FACTOR

Let us suppose that the target is very abundant and very active. It functions at its maximum capacity in a medium *consuming, for example, all of its substrate in a few minutes*. Let us now suppose that, in these conditions, the target is inhibited and that its intrinsic activity drops by one half. *It is possible that the affected target*

is still sufficiently active to consume all of the substrate from the medium in a few minutes. Thus, we see no difference between the normally active target and the inhibited target! The study of the activity associated with a target is a classic problem in biochemistry, when analyzing an enzyme (which catalyses a chemical reaction) or a receptor (which binds to its natural ligand): the target must be a limiting factor in the system. More precisely, the dynamic phenomenon associated with the target (e.g. enzymatic catalysis or ligand binding to a receptor) must, in the conditions of a given test, be a linear function of the target's concentration (fig. 5.1).

Once the test confirms this condition of linearity, it is possible to measure the concentration of bioactive molecule which alters the activity to 50% of the total effect sought (the EC50). Besides the practical measurement, the EC50 value can possess a theoretical meaning if the test respects certain additional constraints.

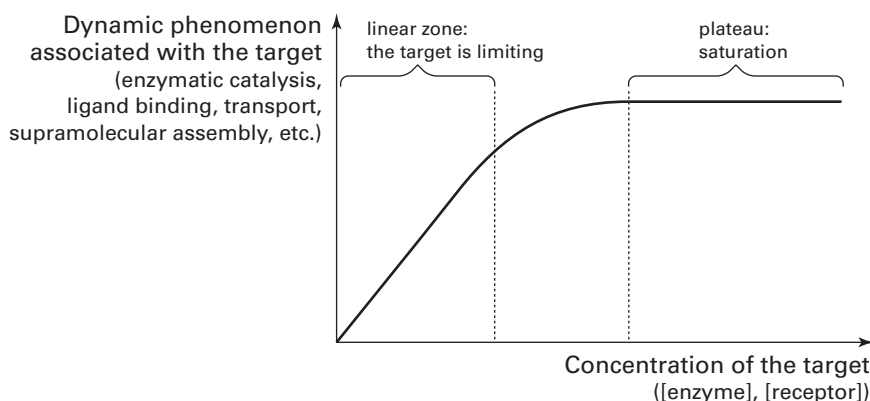


Fig. 5.1 - At low target concentrations, the dynamic phenomenon that is associated to it (enzymatic catalysis, ligand binding) is a linear function of the concentration (linear zone). In this case, when the target is inhibited the measurement drops proportionately. It is therefore important to measure within this limiting zone. If the measurement is carried out in the plateau phase, the target (although affected and less active) continues to function at saturation. It is thus not possible to detect any potential bioactivity.

5.3. ASSAYING THE ACTION OF AN INHIBITOR ON AN ENZYME UNDER MICHAELIAN CONDITIONS: K_i

The purpose of this paragraph is to give the main theoretical and practical aspects of Michaelian enzymology and inhibition in this context. The Michaelian constants are briefly explained theoretically and practically. There will be no mention of enzymes having several substrates for which the Michaelian model is a generalisation, nor allosteric enzymes which deviate from this. The reader is invited to consult works on enzymology or general biochemistry (CORNISH-BOWDEN, 2004; PELMONT, 2005).

5.3.1. AN ENZYME IS A BIOLOGICAL CATALYST

Let us take for example a reaction at equilibrium $A \rightleftharpoons B$, which occurs very slowly because the energetic barrier (activation energy) to overcome in order for the reactions $A \rightarrow B$ and $B \rightarrow A$ to proceed is very high (fig. 5.2, grey curve). These very slow reactions are consequently non-existent on the biological time-scale! The molecular soup which constitutes a biological system is in theory capable of undergoing all energetically possible reactions, but very slowly. An enzyme can associate with a reactant, for example A , and lower the general activation energy thanks to transition states that are less difficult to attain (fig. 5.2, black curve).

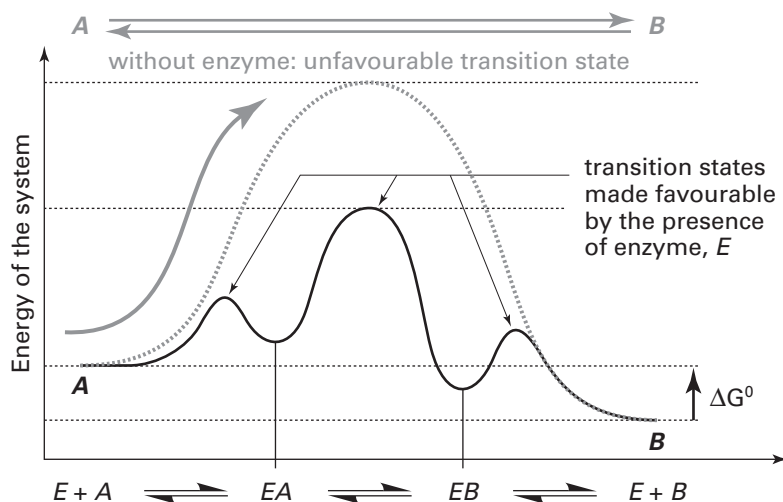


Fig. 5.2 - In this simple uncatalysed reaction (in grey), the reaction rate of $A \rightarrow B$ depends on the difference between the energetic level of A and that of the transition state. *Vice-versa*, the rate of $B \rightarrow A$, will be slower as the energy difference between B and the transition state is greater still. The addition of an enzyme (in black) creates more favourable transition states. Despite these differences in reaction rates, the equilibrated reaction $A \rightleftharpoons B$ only depends on the energy difference between A and B (termed ΔG^0).

An enzyme therefore does not permit the progress of initially impossible reactions, it merely lowers the energy necessary for the reaction to proceed. The reaction is accelerated and thus takes place on the biological time-scale. An enzyme often accelerates a reaction more than 1 billion (10^9) times!

5.3.2. ENZYMATIC CATALYSIS IS REVERSIBLE

In the majority of textbooks, enzymatic reactions feature as oriented reactions, $A \rightarrow B$, as though they are complete and not reversible ($A \rightleftharpoons B$). This misleading representation corresponds to a sequential vision of metabolism in which each catalysed reaction is considered individually, as if it were produced from a pure substrate. Let us take, for example, A in solution, chosen as the *substrate*. The spontaneous reaction leading to the *production* of B takes place very

slowly (fig. 5.3a). In the presence of an enzyme, this reaction $A \longrightarrow B$ is accelerated (fig. 5.3c). If we now take B as the *substrate* in solution, the spontaneous reaction with the *production* of A takes place very slowly (fig. 5.3b). In the presence of the same enzyme, this reaction $B \longrightarrow A$ is also accelerated (fig. 5.3d).

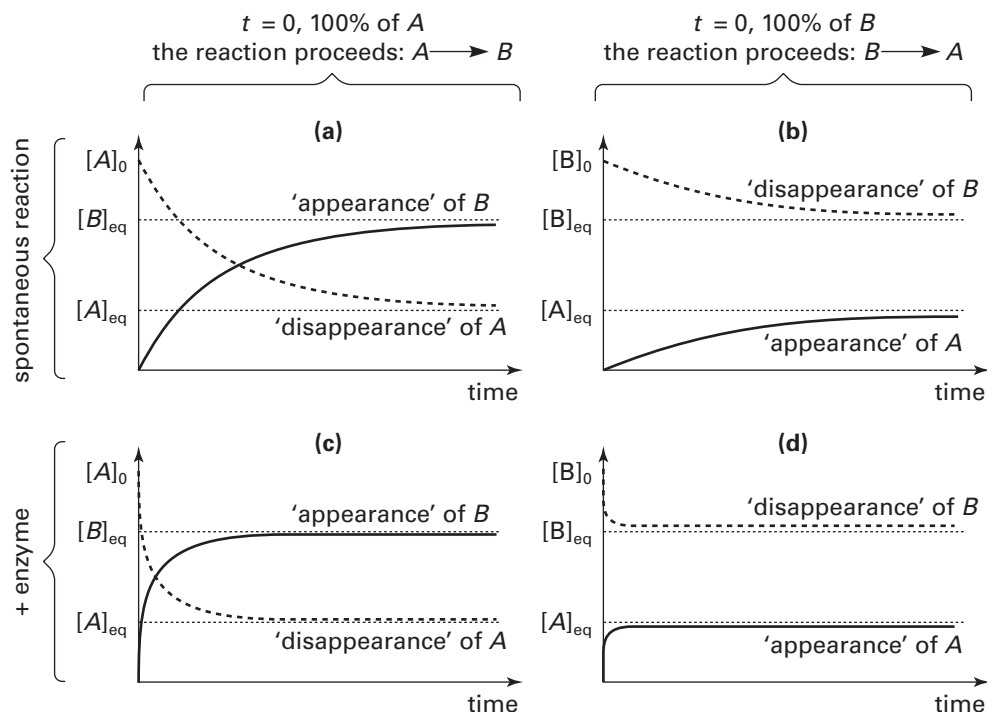


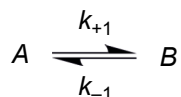
Fig. 5.3 - Concentrations of A and B over time. The reaction $A \rightleftharpoons B$ is produced spontaneously and very slowly, (a) and (b). If the solution initially only contains A (a), then the reaction initially produced is $A \longrightarrow B$. Vice-versa, if the solution initially only contains B (b), then the reaction initially produced is $B \longrightarrow A$. Biochemists often incorrectly depict the reactions as irreversible. However, in all cases, the reaction ends up at equilibrium. In the presence of an enzyme, the two mixtures converge more quickly towards equilibrium, (c) and (d).

Biologically speaking, choosing a direction for the reaction is logical. Indeed, from metabolism in a cell, a molecule is produced by certain reactions, and then becomes a substrate for others. At no moment has an individual reaction the time to reach its theoretical equilibrium. Enzymatic catalysis with a substrate A , can produce B which is immediately used in another catalytic step, producing C , and so on. Dynamically, the reactions $A \longrightarrow B \longrightarrow C$ follow sequentially in a process called *channelling*. Besides, B can be extracted very rapidly from the reaction medium by pumping it into another biological compartment. Lastly, some reactions, for which the ΔG^0 (see fig. 5.2) is unfavourable, are coupled to other reactions that liberate the necessary energy. All of these biological phenomena *orient* the reactions 'independently' away from their theoretical chemical equilibrium and justify this representation.

In a test automated for screening (conceived to measure $A \longrightarrow B$ in the biological direction), it is possible that the reaction *in vitro* may not consume all of its substrate as it has quite simply reached its equilibrium ($A \longleftrightarrow B$)!

5.3.3. THE INITIAL RATE, A MEANS TO CHARACTERISE A REACTION

For spontaneous reactions, the rates of production of A and B are linked to their concentrations according to the law of mass action:



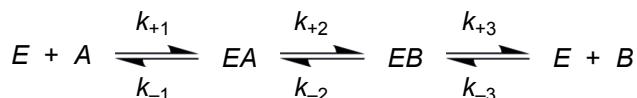
$$\frac{d[B]}{dt} = -\frac{d[A]}{dt} = k_{+1}[A] - k_{-1}[B] \quad (\text{eq. 5.1})$$

where $d[B]/dt$ is the rate of appearance of B .

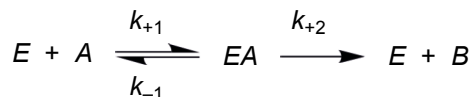
When equilibrium is reached, $[B]_{\text{eq}}/[A]_{\text{eq}} = k_{+1}/k_{-1}$, which is a constant ratio that is termed K_{eq} , the equilibrium constant for the reaction. Consequently, there are two ways to characterise this spontaneous reaction (in particular K_{eq}): either by waiting for equilibrium to be reached and then measuring $[B]_{\text{eq}}/[A]_{\text{eq}}$, or by starting at $t = 0$, measuring the **initial rate** and then simply deducing k_{+1}/k_{-1} from the above equation (eq. 5.1).

5.3.4. MICHAELIAN CONDITIONS

When enzyme catalysis takes place, the reaction $A \rightleftharpoons B$ can be studied by way of the following sequence:



In the initial conditions, in the absence of product B , and assuming that the complex EB dissociates faster than it is formed, this sequence can be simplified to:



MICHAELIS and MENTEN (1913) deduced from this simplified version a relationship between the initial reaction rate ($v_i = d[B]/dt = -d[A]/dt$) and the concentration of the substrate A :

$$v_i = V_{\text{max}} \frac{[A]}{[A] + K_m} \quad (\text{eq. 5.2})$$

where V_{max} is the maximal value that the initial rate v_i can take, and K_m is a constant, known as the MICHAELIS-MENTEN constant.

This relation is represented in [figure 5.4a](#). In double reciprocal form the plot becomes linear:

$$\frac{1}{v_i} = \frac{1}{V_{max}} + \frac{K_m}{V_{max}} \frac{1}{[A]} \quad (\text{eq. 5.3})$$

This equation, proposed by LINEWEAVER and BURK (1934), enables an extremely simple graphical determination of the constants K_m and V_{max} ([fig. 5.4b](#)).

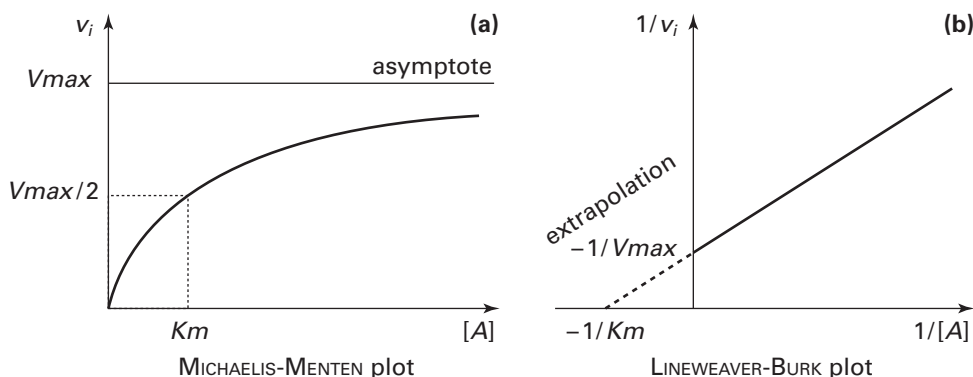


Fig. 5.4 - Effects of the concentration of substrate $[A]$ on the enzymatic reaction rate; for the majority of enzymatic catalyses the initial reaction rate (v_i) is a function of the concentration of substrate $[A]$ which confirms the MICHAELIS-MENTEN equation **(a)**. It is thus possible to deduce V_{max} and K_m graphically. However, since V_{max} is measured by an asymptote, this type of graphical determination is less reliable. LINEWEAVER and BURK greatly simplified this graphical determination by extrapolating the double-reciprocal plot ($1/v_i$ as a function of $1/[A]$) **(b)**.

5.3.5. THE SIGNIFICANCE OF K_m AND V_{max} IN QUALIFYING THE FUNCTION OF AN ENZYME

V_{max} is the maximal theoretical initial rate that an enzyme-catalysed reaction can reach, when all the enzyme is saturated in the form EA . It is therefore a value proportional to the enzyme concentration. This parameter is thus linked to the intrinsic dynamic functioning of the catalyst and can therefore be considered to **measure the activity of the enzyme**.

K_m is the substrate concentration that saturates one half of the enzyme population. The smaller the K_m , the less the substrate needs to be concentrated. This parameter is thus linked to the **affinity of the enzyme** for the substrate; *the smaller the K_m , the greater the affinity*.

5.3.6. THE INHIBITED ENZYME: K_i

Under Michaelian conditions, an inhibitor may affect an enzyme in several ways. Here we shall explore only two simple instances. The power of the Michaelian model resides in the significance of the parameters V_{max} and K_m , which have just

been covered above. When the inhibitor is a structural analogue of the substrate, which occupies the substrate's site in the enzyme, we speak of a **competitive inhibitor**. The affinity of the enzyme for its natural substrate is reduced, and so the K_m increases. When saturated by its substrate, the activity of the enzyme is not modified, thus the V_{max} is unchanged (fig. 5.5a). In contrast, where the inhibitor acts at a distinct site in the enzyme, rendering it less active, we term this a **non-competitive inhibitor**. The activity of the enzyme (V_{max}) drops, whereas the affinity for the substrate (K_m) is unchanged (fig. 5.5b).

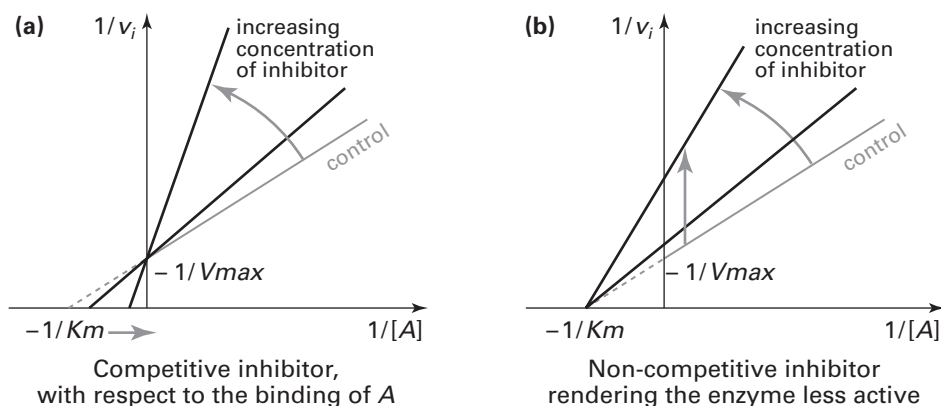


Fig. 5.5 - When the inhibitor is a structural analogue of the substrate, occupying the substrate's site in the enzyme, we refer to a **competitive inhibitor**. The affinity of the enzyme for its natural substrate falls, and thus K_m increases (a). Once saturated by its substrate, the activity of the enzyme, i.e. V_{max} , remains unchanged. In contrast, (b) if the inhibitor acts at a distinct site in the enzyme, rendering it less active, we refer to a **non-competitive inhibitor**. The activity of the enzyme (reflected by the value of V_{max}) drops, whereas the affinity for the substrate (K_m) is unchanged.

An inhibitor I binds to the enzyme E without being converted, according to a reaction whose **dissociation constant at equilibrium** is called K_i .



In the case of a non-competitive inhibitor, the inhibitor I can bind to the enzyme E which is already associated to the substrate A , according to a reaction having the equilibrium constant K_i' :



In its simplest expression, i.e. for a **competitive inhibitor**, K_i corresponds to the **inhibitor concentration at which one half of the enzyme sites are occupied**. In general, the smaller the K_i , the less concentrated the inhibitor needs to be in order to inhibit the enzyme. In a similar way that K_m is a measure of the affinity for the substrate, so K_i is a measure of the affinity for the inhibitor. In practice, we compare K_i to K_m . When K_i is very small relative to K_m ($K_i \ll K_m$), the inhibitor is particularly strong. The analysis of an inhibitor is generally focussed on the reaction of its binding to the enzyme alone (giving K_i). We draw the attention of

the reader to the interest of doing a more in-depth study. For example, in the case of enzymes with several substrates, the analysis of the kinetics by varying only one of the co-substrates, permits the problem to be reduced to one of the cases shown in [figure 5.5](#). It is thus possible to deduce rich structural information, such as the inhibitor binding site (competitive, for example, for only one of the co-substrates).

5.4. ASSAYING THE ACTION OF A COMPETITIVE INHIBITOR UPON A RECEPTOR: IC₅₀

The binding of a ligand L to its receptor R is a biological phenomenon similar in nature to that of a substrate binding to an enzyme.



Whereas it is simple to detect binding of an enzyme to a substrate, since it is followed by enzymatic catalysis measurable with various tests, the direct measurement of a substrate binding to its receptor is tricky (for instance, using surface plasmon resonance; SCHUCK, 1997). Classically, the reaction of a ligand binding to its receptor is characterised by competition with a radiolabelled ligand maintained at a fixed concentration, yielding a **competitive binding curve**. As the concentration of unlabelled ligand increases, the lower the measured amount of radioligand on the receptor. Thus, the binding of a radioactive ligand L^* is **displaced** by the unlabelled ligand L , as would be the case with an antagonistic inhibitor I .



The concentration of I or L , for which the binding of radioligand L^* is at half of the maximal inhibition, corresponds to an inflexion point; this is termed IC₅₀ ([fig. 5.6](#)).

The binding of a ligand to its receptor may seem to be a simple problem to analyse and describe by equations. However, the concentrations of *free* ligand and *bound* ligand vary in this experiment and can have an influence on the measured binding. *Cooperativity* is possible, meaning that ligand binding stimulates further binding of ligand.

The IC₅₀ depends on three theoretical factors:

- the affinity of the receptor for the inhibitor; the higher the affinity for I , the lower the IC₅₀ becomes,
- the concentration of radioligand L^* ; the greater the concentration of L^* , the more inhibitor I will be required to reach half-maximal inhibition and the higher the IC₅₀,

- the affinity of the radioligand L^* for the receptor, expressed by the dissociation constant, K_d . More inhibitor will be necessary to compete with a tightly bound ligand (low K_d) than to displace a weakly bound ligand (high K_d).

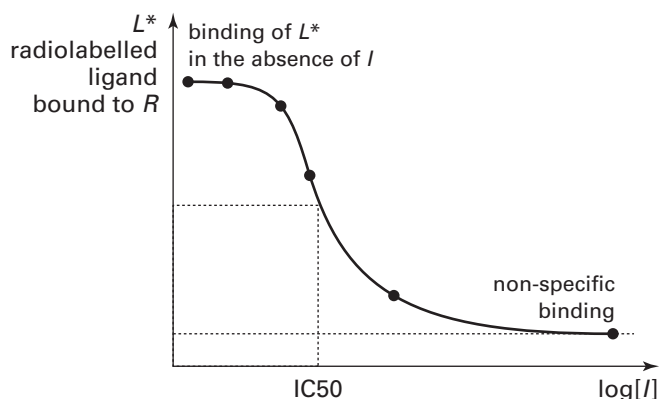


Fig. 5.6 - Competitive binding experiments measure the binding of a radioligand L^* to a receptor R at varying concentrations of unlabelled ligand L or inhibitor I . The concentration of unlabelled ligand is plotted in logarithms. Initially, a plateau phase measures the binding of L^* in the absence of I . At high antagonist concentrations the plateau measures non-specific binding. The antagonist concentration at which the binding of L^* marks the halfway point between the two plateaux is called IC_{50} , the concentration at 50% of the maximal inhibition.

5.5. RELATIONSHIP BETWEEN K_i AND IC_{50} : THE CHENG-PRUSOFF EQUATION

Taking the simple case of reversible binding at equilibrium with no cooperativity, according to the law of mass action, K_i – the dissociation constant at equilibrium – can be deduced from the IC_{50} according to the equation of CHENG and PRUSOFF (1973).

$$K_i = \frac{IC_{50}}{1 + \frac{[L]}{K_d}} \quad (\text{eq. 5.4})$$

Similarly, for an enzyme, the CHENG-PRUSOFF equation is expressed:

$$K_i = \frac{IC_{50}}{1 + \frac{[A]}{K_m}} \quad (\text{eq. 5.5})$$

The IC_{50} measures, at a fixed concentration of substrate A , the concentration of I eliciting 50% inhibition. The reader is invited to read the articles by CHENG (2002, 2004) for a discussion of this equation in the particular case of *cooperativity*.

5.6. EC50: A GENERALISATION FOR ALL MOLECULES GENERATING A BIOLOGICAL EFFECT (BIOACTIVITY)

The experimenter can define with any **dose-effect curve** the concentration of molecules for which 50% of the bioactivity is observed, whether with respect to the inhibition of a chemical activity as we have described in this chapter, or activation, or indeed any measurable effect. We refer to this as **EC50** (the **effective concentration at 50% of the total effect**). The definition of EC50 grew out of the need to compare molecules to each other for their effects on a biological component. More so than with the IC50 measured in the context of enzymatic activity, or of a receptor, the EC50 measured for complex cellular processes or on the scale of whole organisms can say no more than what it measures in the given experimental conditions.

5.7. CONCLUSION

K_i is a measure of the *affinity* of an inhibitor for a biological object (an enzyme, a receptor). This measure is representative of an ideal chemical reaction carried out under Michaelian conditions, established *in vitro*. K_i is the most informative parameter for characterising inhibition. It allows amongst other things a comparison of different bioactive molecules and different molecular targets. It is important to remember that Michaelian conditions do not necessarily reflect the situation *in vivo*. (A famous example is that of catalase dismuting hydrogen peroxide. Catalase is so abundant in some cells that it is observable in crystalline form. This enzyme is therefore not limiting and functions outside of the zone of validity for MICHAELIS-MENTEN conditions.) Besides, MICHAELIS-MENTEN conditions are not always possible to set up during screening. For instance, for reasons relating to cost or to scale-up, a substrate may be used in a limiting amount, so it is important to ensure that the target is always limiting in this system.

The IC50 for inhibition, and more generally EC50 for all types of bioactivity, is a pragmatic measure, which depends on the test. The power of this measure is that, on the one hand, it can be established under Michaelian conditions and linked to K_i by the CHENG-PRUSOFF equation. On the other hand, EC50 can be established in test conditions that are closer to those *in vivo*. Furthermore, this measure can be generalised beyond biochemistry, at the physiological and phenotypic levels (for pharmacological effects on cells, tissues and organisms). The EC50 thus permits a comparison of the dose-effect relationships over a very varied range of biological components. It is, however, necessary to be prudent when comparing the EC50 values from different tests, or when desiring to refer to this value outside of its domain of validity, *i.e.* the test in which EC50 is measured.

5.8. REFERENCES

BADET B. (2002) *La catalyse enzymatique*.

Numéro spécial de *L'Actualité Chimique* **256**: 5-30, EDP Sciences

CHENG H.C. (2002) The power issue: determination of K_B or K_i from IC50.

A closer look at the Cheng-Prusoff equation, the Schild plot and related power equations.

J. Pharmacol. Toxicol. Methods **46**: 61-71

CHENG H.C. (2004) The influence of cooperativity on the determination of dissociation constants: examination of the Cheng-Prusoff equation, the Scatchard analysis, the Schild analysis and related power equations. *Pharmacol. Res.* **50**: 21-40

CHENG Y.C., PRUSOFF W.H. (1973) Relationship between the inhibition constant (K_i) and the concentration of inhibitor which causes 50 percent inhibition (IC50) of an enzymatic reaction. *Biochem. Pharmacol.* **22**: 3099-3108

CORNISH-BOWDEN A. (2004) *Fundamentals of Enzyme Kinetics* (3rd edition).

Portland Press, 438 p.

LINEWEAVER H., BURK D. (1934) The determination of enzyme dissociation constants.

J. Am. Chem. Soc. **56**: 658-666.

MICHAELIS M., MENTEN M.L. (1913) Die Kinetik der Invertinwirkung.

Biochem. Z. **49**: 333-369

PELMONT J. (2005) *Enzymes. Catalyseurs du monde vivant*. Collection Grenoble Sciences. PUG, 1039 p.

SCHUCK P. (1997) Use of surface plasmon resonance to probe the equilibrium and dynamic aspects of interactions between biological macromolecules.

Annu. Rev. Biophys. Biomol. Struct. **26**: 541-566

Chapter 6

MODELLING THE PHARMACOLOGICAL SCREENING: CONTROLLING THE PROCESSES AND THE CHEMICAL, BIOLOGICAL AND EXPERIMENTAL INFORMATION

Sylvaine Roy

6.1. INTRODUCTION

Why undertake **modelling** of the screening process? Although the screening of bioactive molecules does not always fall into the category of being high-throughput ('only' several tens of thousands of molecules screened at a time in pharmacological screens within academic laboratories), the quantity of data generated is significant. Does one wish to save the raw results and the parameters of these experiments? On the **scale of a single screen**, the analysis of the raw data and the identification of potential hits can potentially be carried out with the help of a simple **spreadsheet** (the most popular being supplied as part of office software along with word processing programs). Such a 'manual' solution is already more difficult to establish when the user wishes to understand particular results by correlating them with different parameters such as the molecular structures or the experimental conditions. The process becomes tedious and difficult to manage when the user decides to carry out statistical analyses, standardisations and syntheses of several screens relating to the same test, which soon becomes the case in classic screening projects. Furthermore, these data, already precious for a screen, potentially represent on the **scale of several screens** a considerable body of information from which it will be possible to extract **knowledge**. In the end, this information represents a large part of the resources/capital of the platform or of the team having performed the screens.

The **data** (results and parameters) must therefore be conserved in such a way that they can be exploited and analysed both immediately and in the longer term. Manual organisation often proves to be inappropriate and it is necessary to envisage using an **information system** that will allow the backup, management and analysis of the data. How do we choose such a system? Should existing software (commercial or public) be used? How do we evaluate the suitability of software

developed in-house and the means necessary to set this up? How, in this case, should quality software be developed to meet the given needs? How, incidentally, do we express these needs? What should be done in order that biologists or chemists on the one hand and informaticians on the other speak the same language and understand one another? How might the biologists or chemists control and validate the future organisation and management of their data? The results of screening, for example, are presented in several forms (raw data files, normalised signals, notion of hit-molecules): how should the relationships between these objects or concepts be expressed and implemented? In order to answer these questions, **modelling the ‘screening business’** and/or the necessary information is essential. In this chapter, after having specified the framework and the goals of such a procedure, its principles will be illustrated.

6.2. NEEDS ANALYSIS BY MODELLING

In any computerisation procedure a crucial step is needs analysis. It is critical for continuation and errors at this stage will prove to be very costly afterwards. It is also one of the most difficult steps as it requires precise and efficient communication between those involved from differing cultures, who may not necessarily use the same terminology. This concerns more particularly the future users unfamiliar with computational techniques and the informaticians who are not specialists in the given field.

It is appropriate here to properly distinguish the three following steps:

- › **the analysis step** involves an investigation of the problem and the needs;
- › **the design step** aims to come up with a design solution satisfying the needs (for example, a scheme for a database). These two steps can be summarised by the phrases “*do the right thing*” (for the analysis) and “*do it well*” (for the design);
- › **the implementation step** entails actually implementing the solution (e.g. coding in a defined programming language, about precise materials etc.).

In this chapter, we shall deal essentially with the analysis step. The stages of design and implementation, more independent of the field considered, are mentioned only at the end of the chapter. **The analysis** itself can be subdivided into three parts:

- › **capture of the needs**, *i.e.* the collection of information in order to understand, on the one hand, the domain (for example, the screen), and on the other, the problem(s) (for example, the analysis and the management of data) to be treated by the information system,
- › **definition of the needs**, *i.e.* the review and synthesis of the information gathered in a language understandable by future users,
- › **specification of the needs**, *i.e.* a detailed and more formal expression of their essential characteristics, useful both for the users and for the people who will then devise the information system.

Different methodologies for needs analysis exist, some oriented towards a decomposition into functionality (for example: managing a chemical library, analysing the results of a screen), others more centred towards a decomposition into objects or concepts (concept of hit molecule, mother plate or test plate). The procedure outlined here blends the two types of decomposition and employs a language that has established itself as the standard modelling script, known as **UML (Unified Modelling Language)**. In the context of screening, it is possible to model totally the screening business in order to analyse the needs in terms of information systems. It is important to note that UML is a modelling language (essentially graphical) and not a method. The majority of methods for program development offer both a modelling language and a development process.

6.3. CAPTURE OF THE NEEDS

This step is often conducted by a computer modeller, as the aim is to understand the domain (here, screening for bioactive molecules) and the problem (here, the needs of the management program and data analysis). This step must involve all of the actors, in particular robot operators, biologists and chemists, concerned by these needs. In order to collect the maximum amount of information, the coordinator gathers the existing documents and organises interviews or even workshops during which any needs are expressed. He or she also proposes a study of the existing systems. From our experience, an excellent formula for a workshop for the expression of needs can be a full day simulating the screening process, from the proposal of a target, to the creation of synthesis reports identifying the bioactive molecules.

6.4. DEFINITION OF THE NEEDS AND NECESSITY OF A VOCABULARY COMMON TO BIOLOGISTS, CHEMISTS AND INFORMATICIANS

The information gathered must be reproduced in a synthetic form. The draft can be informal, in the form of **an initial list of specifications** in natural language, **specifying what the system has to do but also what it must not do and what are the priorities**. Immediately apparent is the necessity for developing a common language, a vocabulary (see chapter 1). Once this vocabulary is well defined (e.g. control of bioactivity, hit candidate) it is then used not only in informal documents, but also in the UML documents and in program development (designation of the program objects and user interface).

6.5. SPECIFICATION OF THE NEEDS

The principal interest in using UML resides in the possibilities for communication that it offers. It permits the communication of certain concepts with more clarity

than natural language or than other modelling tools. It also helps to acquire a global vision of the system.

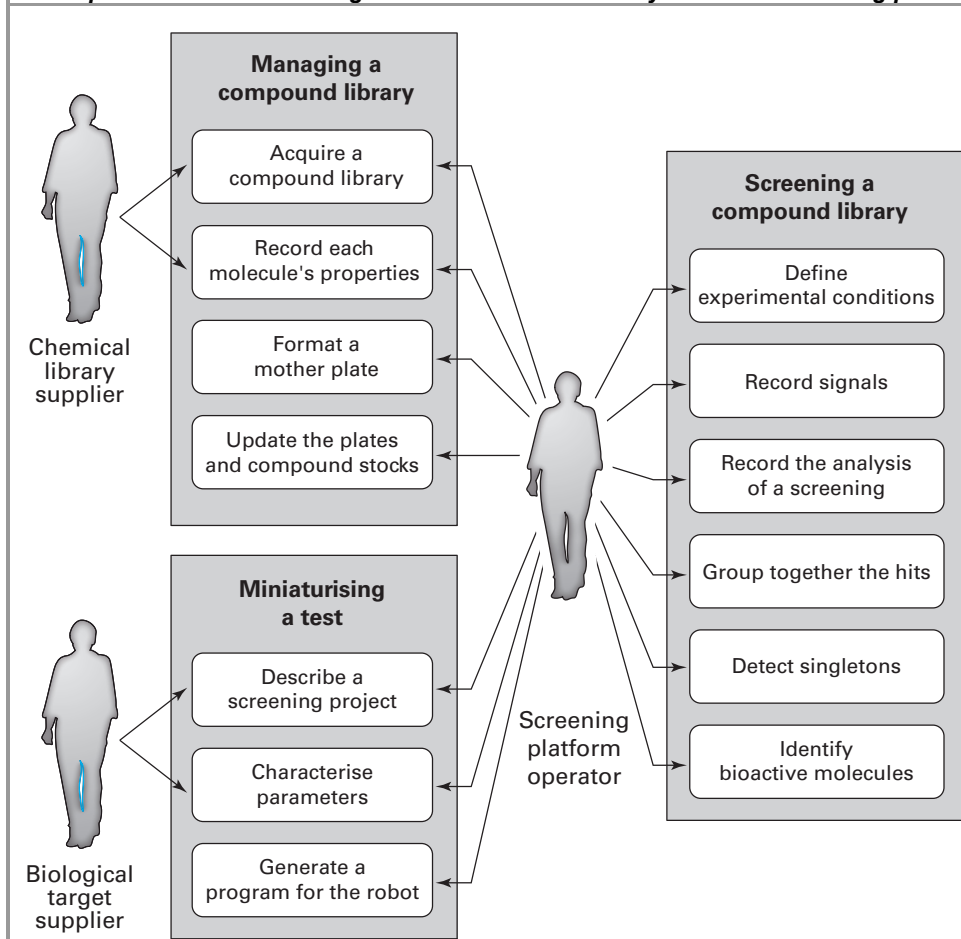
UML offers a number of **diagrams** and particular **notation** usable in the analysis, design and implementation phases. The principal diagrams for the analysis phase are:

- › use-case diagrams,
- › activity diagrams,
- › class diagrams.

6.5.1. USE CASES AND THEIR DIAGRAMS

Use-case diagrams serve to define both the limits of the system and its functions from the users' point of view. They allow structuring of the needs expressed in the list of specifications, as illustrated in example 6.1.

Example 6.1 - a use-case diagram for the information system of a screening platform



In order to construct use-case diagrams, it is necessary to ask the following questions: “*who is interested in the system?*” and “*in what?*”

The first question enables a definition of the **actors**, for example the ‘screening platform operator’, the ‘chemical library supplier’ or the ‘supplier of the biological target’. These actors represent roles with respect to the system and not physical people. The same physical person can fulfil several roles (in a project being, for example, both the ‘screening platform operator’ and the ‘supplier of the biological target’), or the same role can concern several people (for example there will be several laboratories that have the role of ‘supplier of the biological target’). A precise name, which uses terms of the screening business, must be given to each actor. If this is difficult to find, then it signifies that the actor is poorly identified, and that he plays several roles which ought to be distinguished. Each actor must have a brief description, as example 6.2 illustrates.

Example 6.2 - description of the actors

- ▶ The chemical library supplier identifies the physical entity (university laboratory, private firm, research institute) who supplies a chemical library to the platform. It supplies the chemicals in multi-well plates as well as a file listing the molecules with structural information.
- ▶ The supplier of a **biological target** is a physical person who wishes to perform screening with the platform in order to study a biological phenomenon specific to his research topics. He knows well the ‘full-size’ test and takes part in its miniaturisation and automation.
- ▶ The platform **operator is a physical person who carries out the screens with the platform** and controls the information system; he analyses the results and identifies the bioactive molecules. ▲

The second question, “*in what?*” permits the identification of activities, which are named by a verb of action, for example, “*identify bioactive molecules*”, “*take delivery of a chemical library*”. Each activity is also described by text, called the **use case**. The use case describes the future way in which the system will be used in order to reach a precise objective. The description can be relatively formalised (concept of objective, reference or precondition etc.) as in example 6.3.

Example 6.3 - use case: “record the analysis of screening”

- | | |
|---------------|--|
| Objective | ➔ record a snapshot of the screening analysis |
| Actors | ➔ the operator of the platform (initiator) |
| References | ➔ list of specifications |
| Preconditions | ➔ a threshold value for bioactivity has been validated |

The actor enters the filename destined for saving, as text, the information generated in the business process *Screening a chemical library*:

- ▶ statistical parameters of the different functions of the wells: controls for bioactivity and bio-inactivity, sample,
- ▶ list of the rejected candidate hits,
- ▶ hit list.

For each element of the preceding lists, the following data are recorded:

- ▶ barcode of the test plate,
- ▶ position of the well in the test plate,

- chemical formula and molecular mass,
- concentration of the molecule,
- values of the different raw, standardised signals.

After validation by the user, the results file is created; the information previously described is automatically saved in the database. ▲

These use cases are written or validated in close collaboration between the system modeller, the expert chemists, biologists, and robot operators and the potential developer(s).

6.5.2. ACTIVITY DIAGRAMS

This type of diagram is very useful for the dialogue between experts and informaticians, notably to articulate the sequence of different tasks, relative to one another. These diagrams describe the organisation of activities by supporting both conditional behaviour (“*if such a condition is met, do this, else do that...*”) and parallel processes; the contribution of such a diagram is essential for **modelling the process** as it allows a representation not only of the processing to be done but also of the actors involved and the use of information.

The **activities** are represented by bubbles and are linked by transitions. These transitions are triggered by events that can be the culmination of the activity preceding the transition, the availability of an object in a particular state, or satisfying a condition (ex. 6.4).

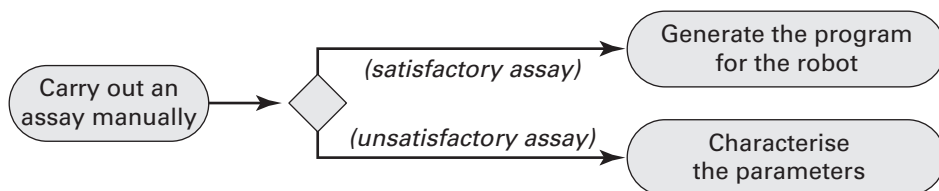
Example 6.4 - an activity diagram



The activities: *Build a pool of molecules for screening* and *Perform screening* are related by a transition; the initial point, represented by a black dot, precedes the activity *Build a pool of molecules for screening* and the final point, represented by a black spot surrounded by a circle, follows *Perform screening*. ▲

An activity diagram can contain alternative paths. Each branch is subjected to a guard condition as shown in example 6.5.

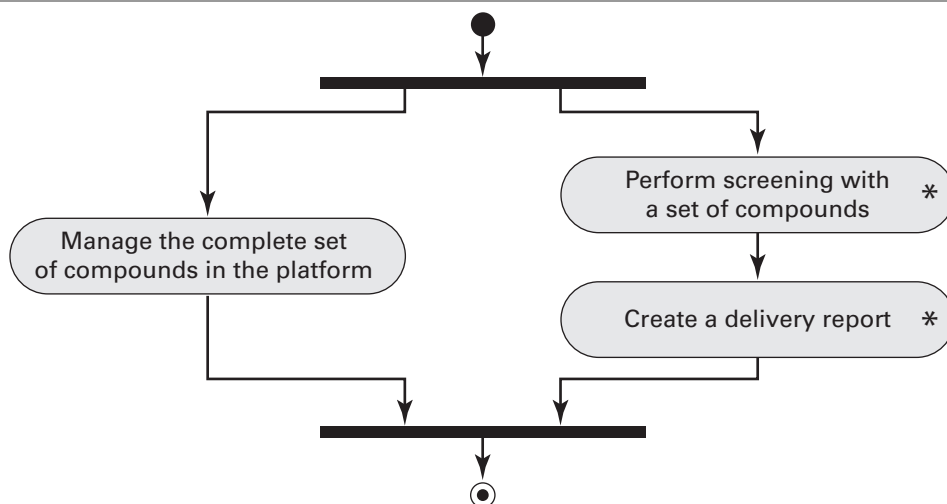
Example 6.5 - control flow with conditional branching



If the manual test is carried out satisfactorily, the program for the robot can be generated; in the opposite case, the parameters are characterised all over again. ▲

Sometimes an activity diagram can comprise two parallel paths. These two paths begin at a fork and rejoin each other by synchronising again later, as depicted in example 6.6.

Example 6.6 - an activity diagram for a screening project on an automated platform.



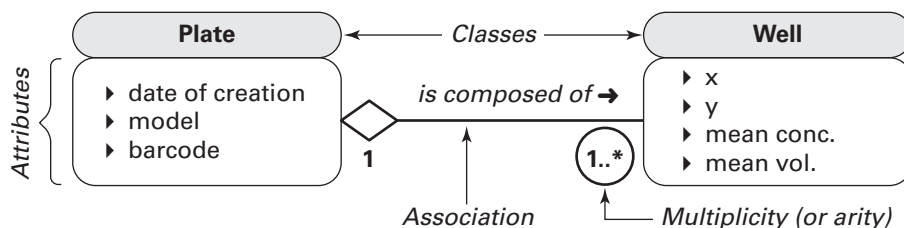
The realisation of a project can involve the implementation of several screens that end up in the development of one or several delivery reports (the stars located by the activities *Perform screening of a group of molecules* and *Draft a delivery report* signify that these stages are iterative). In parallel, the experimenters must focus on another task, the management of each chemical compound.

6.5.3. CLASS DIAGRAMS AND THE DOMAIN MODEL

Besides the processes that we have just mentioned, it is necessary to represent the significant **objects** or concepts, also called **conceptual classes** of the screening. For this, a **domain model** has to be created, which is a group of class diagrams (ex. 6.7, 6.8 and 6.10). A **class diagram** permits visualisation of:

- › the **conceptual classes or objects** of the domain (represented by rectangles),
- › the **associations** or relationships between classes (represented by lines),
- › the **attributes**, *i.e.* the characteristics or properties of each class.

Example 6.7 - what a class diagram shows



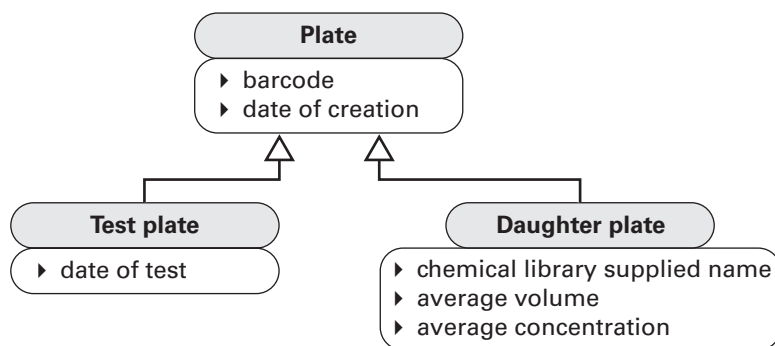
Here, two *objects* or *conceptual classes* have been identified: the *plate* and the *well*. The plate is composed of several wells; we shall note the importance of the idea of *multiplicity*

(the number at each extremity of the link). A plate contains several wells and always at least one well (multiplicity denoted '1..*'). A well can only belong to a single plate (multiplicity denoted '1').

A plate is characterised by several *attributes*: its date of creation; the model (plate with a transparent base, black base etc.) and its barcodes, which is a unique identifier. A well is characterised by several attributes: its coordinates (number of rows x, number of columns y) allowing it to be uniquely identified within a plate; the concentration of the molecule; the fill volume. ▴

The links between the classes can be of different types. In addition to the simple link symbolised by '—', or the composition symbolised by '◊—', the specialisation link ('◄—') is often used. It allows identification, among objects of a generic class, the object sub-groups (specialised classes) having specific characteristics, as shown in example 6.8.

Example 6.8 - specialisation



Here, the class *Plate* is a generic class. It has attributes (*barcodes* and *creation date*) common to the two sub-classes *Test plate* and *Daughter plate*. These latter two are specialised classes which have attributes specific to their plate type. ▴

How can the classes be identified? Each conceptual class is an idea or a thing, an object from the *real* world (never a program component). These classes can notably be found by carrying out a linguistic analysis of the text documents already produced (minutes from meetings, list of specifications, use case) and picking out terms or terminology (see ex. 6.9).

Example 6.9 - analysis of an extract from a list of specifications (identification of important terms)

(...) Receipt of a chemical library

The chemical libraries (collection of chemical molecules) received by the platform come from different suppliers: academic laboratories, research institutes, private firms and so on. A chemical library consists of a group of multi-well plates termed mother plates. Each well of these plates contains a molecule at very high concentration. The wells of a mother plate are characterised by an equal average concentration of molecules and equal average volume (...) ▴

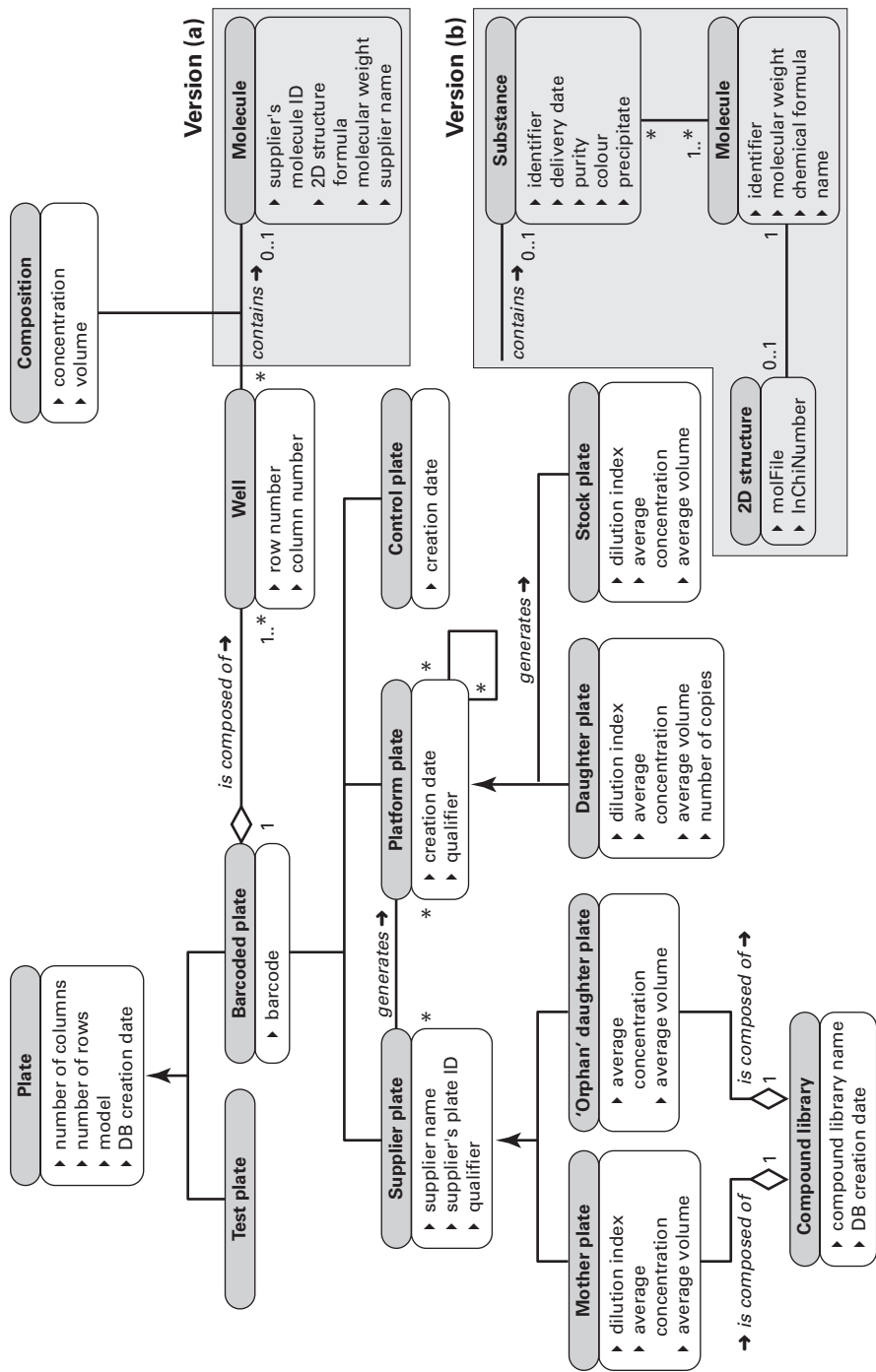
Some authors also recommend drawing up lists of objects or possible concepts according to identified categories (for example, the category of physical objects: *wells, molecule*; of processes: *analysis of the signals*; or of documents: *file containing the signal readout*, for instance).

It is important to note that there is no ‘correct’ list; similarly, a domain model is not absolutely true or false. It is merely a communication tool. Frequently, it must be modified or even entirely reconstructed several times, before arriving at a consensus, created in close collaboration between the experts (biologists/chemists) and the modeller. It must be annotated by text explaining the choices made, as illustrated in example 6.10 – an extract from a model developed for a screening platform.

This diagram (and its associated text) has highlighted numerous points which were not noticed only by reading the list of specifications, for example:

- » The notion of ‘genealogy’ of the different plates (associations ‘generate’ between plates) was introduced due to a request for a precise follow-up of the plates, set out during the construction of the domain model.
- » It did arise during the validation of this schema that a chemical library is not necessarily composed solely of plates but that compounds are sometimes proposed for screening by the biologists themselves, and that they could be brought as samples contained in EPPENDORF tubes, a notion that the diagram has to include. This point also highlights the need to specify the chemical library concept for a platform, and in particular the need for harmonisation in the management of different chemical libraries coming from varied public or private suppliers (chapter 2).
- » The concept of well contents has hugely evolved through the course of modelling, as shown by the comparison between the two versions **(a)** and **(b)** of the representation of these ‘contents’. The version **(a)** of this diagram shows that each well contains a molecule. There is therefore a comparison between the ‘molecule’ concept (which has certain properties: a structure, a formula, a molecular mass etc.) and the substance or ‘sample’ concept synthesised on a precise date, with a certain purity, by a particular chemist, diluted in a preservative on a precise date. This ‘confusion’ may have no repercussions on the functioning of a screening platform, and there are not necessarily grounds for a distinction between the two notions. However, such a task of formalisation makes one aware that the information system to be based on such a diagram will not make this distinction and that, notably, it will not cope with the notions of ‘shelf life’ or ‘expiry date’ of each substance. If this is important and those in charge of the platform also wish, for example, to record the physical aspect of the well contents (colour, presence of a precipitate etc.), the appearance of the substance concept (version **(b)**) is important.

Example 6.10 - a class diagram modelling the management of molecules in the platform with two possible versions for modelling the well contents.



Example of associated text (extract)***The plate concept***

All of the plates used in the platform are multi-well plates characterised by a certain number of lines and columns but also by a type of well (e.g. wells with black or transparent bases). It has therefore been decided, at the level of the domain model, to represent the concept of a plate characterised by a number of lines and columns, and by a plate model.

The barcoded plate concept

When one looks more closely at the different types of plate used in the platform, two plate categories may be distinguished:

- ▶ test plates, in which the biological tests are conducted and which contain in their wells a screening molecule or a control but also a given volume of the reaction mix and the biological target. Test plates are generated in non-barcoded plates.
- ▶ other plates that are barcoded and that only contain the small molecules screened.

The plate concept has thus specialised into two other concepts: that of test plate and that of barcoded plate. A barcoded plate is a plate characterised by barcodes. It is composed of wells.

The contents of a well

Version (a): a well can contain one molecule.

Version (b): a well can contain a substance which has certain physical characteristics. This substance corresponds to a molecule whose supplier can or cannot supply the two-dimensional structure. ▽

Consequently, if biologists or chemists wish to compare bioactivity data with the results of virtual screening on the molecules, it will be necessary to expound further the concept of a 'molecule': thus, the calculation of molecular descriptors (chapter 11) and their uses (chapters 12, 13 and 15) might require considering the inclusion in the model of multiple representations of the same molecule (tautomers, three-dimensional conformations etc.).

The moments of awareness that a class diagram stimulates cost no more than a 'pencil stroke' when they arise during needs analysis. Changing or adding a functionality is, however, extremely costly, indeed impossible to consider if the user becomes aware of particular needs while using the software, when already designed or bought. Such modelling, both out of the need for quality in the specifications and an active involvement of the user, therefore avoids substantial costs in the middle or long term.

The model constructed allows a formalisation of the domain concepts and the relationships between them, but it also imparts to users **knowledge and control of the future organisation of the data**. The design and implementation of the data management structures are indeed derived from the domain model. Rules exist for the translation of classes into tables (components of relational databases), precise rules to such a point that **certain tools for editing UML diagrams offer functions for the automatic translation of class diagrams into schemas for relational databases, in the language SQL (Structured Query Language)**, a language directly interpretable by an information system for data management. **The biologist or chemist fully participates here in the modelling and the organisation of his or her data and thus is equipped with a command tool for screening projects.** The

set of diagrams also permits evaluation of the cost of developing an information system corresponding to the desired functions. The number of use cases, their size, the scope of the domain model and the complexity of the diagrams can often be good indicators of what workload to expect.

6.6. CONCLUSION

The job of modelling the process of screening for bioactive molecules is extremely useful. It enables a precise formalisation of the needs in a language common to both users and informaticians, and it gives preliminary answers relating to the acquisition or development of an appropriate information system. Such a study permits evaluation of the adequacy of existing systems (commercial or public) for the needs; it also gives a good estimation of the effort required in terms of the design and development of a custom system, and allows judgement to be made of the suitability of developing it in house.

The **command of the processes and the data** that modelling permits is also a point not to be ignored in the set-up of quality procedures for the technological tool (the platform) and the screening projects as a whole (chapter 7). Besides the results of this modelling process, its benefit lies as well in the effort of explaining the screening business, the identification and organisation of objects, the data and the knowledge handled. This effort can of course be capitalised on at a later stage of data use. Indeed, computing must go beyond being a simple tool for storing data and calculating results (MORGAT and RECHENMANN, 2002). The mass of data accumulated from screening platforms concerning the interaction of molecules with various targets potentially represents a mine of knowledge well worth extracting (chapters 12, 13 and 14). It could be useful to move from databases to knowledge-bases in chemogenomics, as has been widely initiated in genomics.

6.7. REFERENCES

- FOWLER M. (2001) *UML*. Le Tout en Poche, Campus Press, Paris
- LARMAN C. (2002) *UML et les Design Patterns*. Campus Press, Paris
- MORGAT A. and RECHENMANN F. (2002) Bio-informatique. Modélisation des données biologiques. *Médecine Sciences* **18**: 366-374
- MULLER P.A. and GAERTNER N. (2000) *Modélisation objet avec UML*. Eyrolles, Paris
- NAIBURG E. and MAKSIMCHUK R. (2002) *Bases de données avec UML*. Campus Press, Paris

QUALITY PROCEDURES IN AUTOMATED SCREENING

Caroline BARETTE

7.1. INTRODUCTION

As with all technological platforms for which the number of samples and data exceeds the capacity for human handling, and the activity of which may be of industrial value, it is indispensable to conduct the activity with an **ordered mode of functioning, aiming to control and to monitor** what takes place.

To become caught up in the mode of functioning rather than the science may seem tedious, nevertheless, the upstream investment in traceability allows any doubt to be reduced about the reliability of the results obtained. It is certainly possible to select a bioactive molecule after conducting experiments without traceability, however, this result will not be of good quality and it will be impossible to justify the proper progression of experiments in case of doubt. Conversely, it is possible to screen a collection of molecules without succeeding in selecting a bioactive component, by following controlled protocols and modes of operation, with good traceability of the samples and data, and this result, while possibly disappointing in other respects, will be of good quality. Quality refers, therefore, to the means put in place to obtain a result and not the result itself. The work involved to establish the procedures – which are not at all quibbling, indiscriminate and obscure rules, but constructed in cooperation with the actors of the platform – is in addition an opportunity to understand the role of each person participating in a screening project. This short chapter provides several fundamental facts in order to embark upon this approach, referred to as **quality procedures**.

7.2. THE CHALLENGES OF QUALITY PROCEDURES

The implementation of quality procedures within the framework of a screening platform's activities is justified in several respects:

» Scientific issues

Recourse to a molecular screening platform implies the screening of a large number of molecules and consequently the generation of a large number of data. The screening in itself only constitutes the first step in a long-term project, but this is a decisive step where the reliability and traceability of the results determine the success of the project and consequently its value.

» Financial issues

The screening of a large number of molecules involves proportional costs: the cost of the infrastructure to acquire and to maintain the platform, material costs linked to the reagents and consumables, the human cost associated with the qualified personnel necessary for running the platform, and the overall costs of the post-screening research. Together these costs add up and so it is important to limit them as much as possible in the context of research, by avoiding, through prevention, any dysfunction. More broadly, the organisations that finance the research would then exert pressure to ensure that the research satisfies the quality criteria in order to avoid waste.

» Competitiveness

The term ‘platform’ suggests a possible access to conducting screening projects brought in by external teams, in the form of collaborations or provision of a service. It is therefore fundamental to instil confidence in the collaborators or clients while continually proving that the project will be carried out with a sustained quality practice permitting reliable data to be obtained within a controlled budget.

Engaging in a quality approach by setting up optimal procedures is therefore necessary to guarantee the **reliability** and **traceability** of screening project results, two of its principal aims.

As each platform is unique in its personnel, its technological material and its scientific aims, there is no absolute, well-informed solution that may then be indiscriminately imposed. Quality procedures do not come from an external source but are established by everyone. Although, to date, a set of priority criteria and the need to obtain certification for academic platforms have not been clearly defined, an assessment of the workplace and a proposal for solutions to be put into effect regarding quality management can draw on an international reference guide, the ISO 9001 Standard 2000 version.

7.3. A REFERENCE GUIDE: THE ISO 9001 STANDARD

This reference guide to ‘quality’ reputed worldwide specifies the requirements relative to all quality management systems, centred on the control of the processes and their efficacy to the client’s satisfaction. By **client**, we mean **any person benefiting from the results** of the platform. This may for example be a **biologist** desiring

to identify a bioactive molecule in a given process, a **chemist** bringing a chemical library and interested in the response profile of a group of biological processes, or an **computer scientist** wishing to analyse the set of measured experimental data. In the context of a collaboration or the provision of a service the term ‘client’ can designate the initiator of a project for which screening is required. The ISO standard is the only reference guide used to obtain **certification** that is internationally recognised, annually revised and valid for three years, and proves the actual implementation of all the requirements. It is part of a series of three standards (ISO 9000, ISO 9001, ISO 9004) dedicated to quality management.

The essential principles of the ISO 9001 Standard, inherited from the ISO 9000 Standard, are the following:

- › to determine the **needs** and **expectations** of the collaborator or client and to establish a **quality policy** of the corresponding objectives,
- › to determine the **processes** and **objectives** for attaining the **quality objectives**,
- › to define the **means to measure the efficacy** of each process for attaining the quality objectives,
- › to measure the **efficacy of each process**,
- › to determine the means **to prevent any non-conformity** and to eliminate the causes,
- › to improve, through identification of any **optimisation** to be carried out, by planning the means to bring about these improvements, by putting into effect this planning, by following the effects by analysis of the results, and by revising the actions for improvement.

Furthermore, the ISO 9001 Standard (example 7.1) contains several significant contributions:

- › **client orientation**: listening to and taking into account the expectations of the client,
- › the **process approach**: operational and functional processes must be articulated in order to guarantee satisfying the expectations,
- › the **involvement of all the actors** in the realisation of the objectives by giving responsibility to each,
- › **continual improvement**, by the permanent follow-up of the results and the measures for improvement, leading to development and excellence in the activities.

Example 7.1 - extract of the key requirements taken from the ISO 9001 Standard 2008 version

- › **Quality management system** - general requirements; requirements relative to the documentation (quality manual, document control).
- › **Responsibility of the management** - engagement of the management; client listening; quality policies; planning (quality objectives, planning of the quality management system); responsibility, authority and communication (responsibility and authority, management representative, internal communication); management analysis (analysis input, analysis output).

- ▶ **Resource management** - making resources available; human resources (competence, raising awareness and training); infrastructures; work environment.
- ▶ **Product implementation** - planning of product **implementation**; processes relating to the clients (determination and analysis of the requirements relating to the product, communication with the client); design and development (planning the design and development; design and development: input and output aspects, review, checking, validation and control of modifications); purchases (purchasing processes, information relating to purchases, verification of the product bought); production and preparation of the service (control, process validation, identification and traceability, client property, preservation of the product); control of monitoring and measurement devices.
- ▶ **Measurement, analysis and improvement** - monitoring and measurements (client satisfaction, internal audit, monitoring and measurement of the processes, monitoring and measurement of the product); control of non-conforming products; data analysis; improvement (continuous, corrective, preventive). ▽

Finally, the most recent ISO 9001 Standard is the 2008 version (November 2008). Compared to previous recommendations, this version includes the qualification of human resources (by training if necessary); this new orientation can simplify and reduce the documentation, which was initially based on the following principle: *“write what you have to do and do what you wrote”*.

7.4. QUALITY PROCEDURES IN FIVE STEPS

Each chapter of the ISO 9001 Standard must be analysed in order for the platform to be aligned with the criteria of the standard, while knowing that the domains dealt with concern as much the functional as the technical organisation. An evaluation of the project setup must first of all be organised to ascertain some crucial points, which are: the traceability of the raw data and the **analytic parameters**; the **readability** and the **reliability of the working methods**. The **notion of process** is established by drawing on the **different steps of the theoretical running of a project**.

In practice, establishing quality procedures involves five successive steps:

7.4.1. ASSESSMENT

The first step consists of carrying out an assessment of all the activities of the screening platform; at this point, a person in charge of quality management is designated. This assessment is made through individual interviews with all members of the team and audits in the workplace, in relation to the different chapters developed in the ISO 9001 Standard. The resources are evaluated and the different processes identified; practically, these processes (such as processes for: the management of screening projects, the automated screening of molecules, the management of chemical libraries) can be formalised in the form of schemas; this schematisation allows a clear definition of the different steps to be developed throughout the process. Lastly, pilot processes are designed. This stage obviously benefits from modelling the processes described in chapter 6.

The evaluation report from this assessment serves as a basis for reflection and discussion in the following step, notably to identify the critical points and from there, the priorities to be defined by the platform manager in cooperation with his or her team.

7.4.2. ACTION PLAN - PLANNING

Having established a policy and quality objectives for the platform, the roles of each person can thus be defined (e.g. management, pilot committee, pilot processes, person in charge of quality control). Outlining the different processes and articulating them must respond in a measurable way to the requirements of the collaborators or clients. Next, quality planning has to describe the manner in which the quality objectives will be achieved, by action plans specifying notably the responsibilities, the time-scale and the associated log-keeping.

7.4.3. PREPARATION

This essential phase, not to be neglected, aims to raise awareness and to train each individual to ensure a total and necessary adherence to the quality procedures. One of the obstacles to setting up quality procedures is the mistrust with respect to a set of practices which can be perceived as being constraining if they have not been established in cooperation.

Besides, this preparatory step must enable the setup of a system for documentation, permitting the assembly of documents, the preparation of document folders and the organisation of the procedures, operating modes and forms. The definition of the contribution of each, during the implementation stage, enables allocation of the different tasks of writing, verification, revision and updating of documents.

Finally, in the 2008 version, the mandatory documentation is limited to the Quality Manual and few written procedures (related to document controls, recording controls, controls of non-conforming products, preventive actions, corrective actions and internal audit).

7.4.4. IMPLEMENTATION

Once a documentation system has been set up, it must next be applied to obtain operational procedures and other work instructions. It is necessary to ensure that all documents are consistent, suitable, easily available, accepted and usable by all, and lastly scalable.

7.4.5. MONITORING

Monitoring the procedures is indispensable in order that the Quality Management System remains permanently suitable for the policies and the quality objectives, which may evolve. This monitoring starts off with internal audits, based on the

procedures or other existing documents, as well as on reviews by the management, which will be able to decide about any adjustments needed. Eventually, the quality procedures can lead to an audit certificate, for official recognition of the level of the platform's quality with respect to the ISO 9001 Standard.

7.5. CONCLUSION

Quality procedures are a collective approach requiring the mobilisation of everyone, but at the same time developing the engagement of each individual, and it can unite a team. It must not be neglected that the will and support of the person in charge of the platform implementing this approach, is a prerequisite for motivating and mobilising the entire team. Quality procedures immediately involve forethought within the team about the screening activity, its objectives and the work methods. This consideration is fed by the formalisation linked to the writing of procedures and by applying them. It contributes greatly to increasing the reliability of the results. Thus, in spite of the heavy investment of human time and energy which is required for setting it up, the quality approach improves the visibility of the role and responsibility of each person, sheds light on the complex functioning of the platform and in this respect it makes the working methods easier.

7.6. REFERENCES

NAVEH, E., MARCUS A. (2004). When does ISO 9000 Quality Assurance standard lead to performance improvement? *IEEE Transactions on Engineering Management* **51**: 352–363

International Organization for Standardization (ISO): Quality management and quality assurance: <http://www.iso.org>

ISO 9001:2008(E), Quality management systems – Requirements:
http://www.iso.org/iso/catalogue_detail.htm?csnumber=46486

SECOND PART

HIGH-CONTENT SCREENING AND STRATEGIES IN CHEMICAL GENETICS

PHENOTYPIC SCREENING WITH CELLS AND FORWARD CHEMICAL GENETICS STRATEGIES

Laurence LAFANECHÈRE

8.1. INTRODUCTION

A commonly used method to understand the role of complex biological systems and how they function is **to disrupt** them and then **to observe** the result of this disruption. A classic way to create such disruptions is to generate **genetic mutations** and then to observe the effect of these mutations on the cell or the organism. **Small organic molecules** can also cause disruption to the functioning of biological systems and can be employed to understand the role of the protein with which they interact. The history of biology is full of examples of complex systems whose **molecular functioning can be understood thanks to the use of drugs or ligands as well as to the characterisation of the protein targets of these ligands**. One such example is the role of colchicine in the discovery of tubulin, a component protein of microtubules (SHELANSKI and TAYLOR, 1967; PETERSON and MITCHISON, 2002).

The development of automated screening methods in an academic setting and the access to large collections of organic molecules has permitted systematising the exploration of the chemical world's diversity for isolating molecules active on biological systems. In parallel, the concept of 'chemical genetics' or 'pharmacological genetics' was born. This term may seem to be a corruption of language because in chemical genetics approaches we are not dealing with the gene but, most of the time, with the gene product. In fact, this concept designates a group of approaches that aims to use small molecules to interfere with proteins systematically and therefore to determine their function, in the same way as mutations are utilised in actual genetics. Conceptually therefore, chemical genetics and genetics are rather analogous. More recently the term **chemogenomics** was proposed to designate the multidisciplinary approaches aiming to dissect biological functions with the help of small molecules (BREDEL and JACOBY, 2004). Genetics has been a formidable engine for discovery in the biomedical sciences. Small molecules offer advantages compared to genetic technologies: they are versatile research tools, which can be quickly adopted by different laboratories and used for the precise control of the

function of certain proteins in a cellular context, particularly in the case of difficult, indeed impossible, genetic manipulation. Furthermore, these small molecules can be a first step towards the development of therapeutic agents. Indeed, on the one hand, they offer a means to test the possible involvement of a given protein in a pathology; on the other hand, their chemical structure can be the starting point for drug development.

This chapter will be devoted more specifically to the approaches for **phenotypic screening with cells**, in the context of **forward chemical genetics**. This approach, which allows the identification and characterisation of new proteins, has taken an increasingly important place in basic research and has proved to be complementary of biochemistry and genetics. For applied research, forward chemical genetics also offers the advantage of selecting drug candidates capable of penetrating cells straightaway and being active in a cellular context.

8.2. THE TRADITIONAL GENETICS APPROACH: FROM PHENOTYPE TO GENE AND FROM GENE TO PHENOTYPE

8.2.1. PHENOTYPE

Phenotype refers to the set of physical and physiological characteristics of an individual resulting from its genome and its environment. A phenotype can be defined on different levels, from the cell (fig. 8.1) to the whole organism (chapter 9).

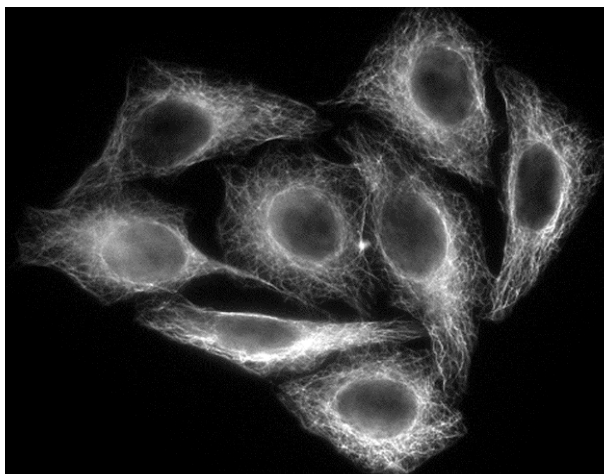


Fig. 8.1 - The cellular phenotype is complex. It can be in part analysed with the help of molecular markers enabling certain structures to be observed. In this image fluorescent probes allow visualisation of the microtubular cytoskeleton.

By using mutagenic agents or radiation, it is possible to cause random mutations in the whole genome of model organisms. With the techniques of molecular biology it is also possible to mutate specifically a given gene. **Genetic mutations** can induce in this way a modification of quite a specific nature at the level of the cell or the whole organism, such as, for example, retarded growth, or an altered appearance or behaviour. We refer to this as a **mutant phenotype**.

8.2.2. FORWARD AND REVERSE GENETICS

Genetics stemming from the works of MENDEL initially consisted of elucidating the nature of those genes responsible for the appearance of a phenotype (DARLAND and DOWLING, 2001). This historically earlier approach is known as **forward genetics** and does not imply any prior knowledge of the genes involved.

A second type of genetic approach, put into practice more recently thanks to genomic sequencing, is **reverse genetics**. In this approach, a previously identified gene is selectively mutated in an organism, or deleted. The total deletion of a gene is called a **knock-out**. The resulting phenotype of this mutation is subsequently studied so as to understand precisely the role played by the gene within the organism.

8.3. CHEMICAL GENETICS

Chemical genetics aims to reproduce these forward and reverse genetic approaches by replacing mutations with the use of small molecules.

In the **reverse chemical genetic** approach, the goal is to isolate small molecules capable of interacting with purified proteins and interfering with their function. The molecules thus identified are then tested for their effects of disabling the function of these proteins in the cell or the whole organism (fig. 8.2). This is the traditional approach to searching for drug candidates.

In **forward chemical genetic** screens, small molecules capable of eliciting specific phenotypes at the cellular level are sought. These molecules can thereafter be used as bait in order to ‘fish’ out their protein targets (fig. 8.2).

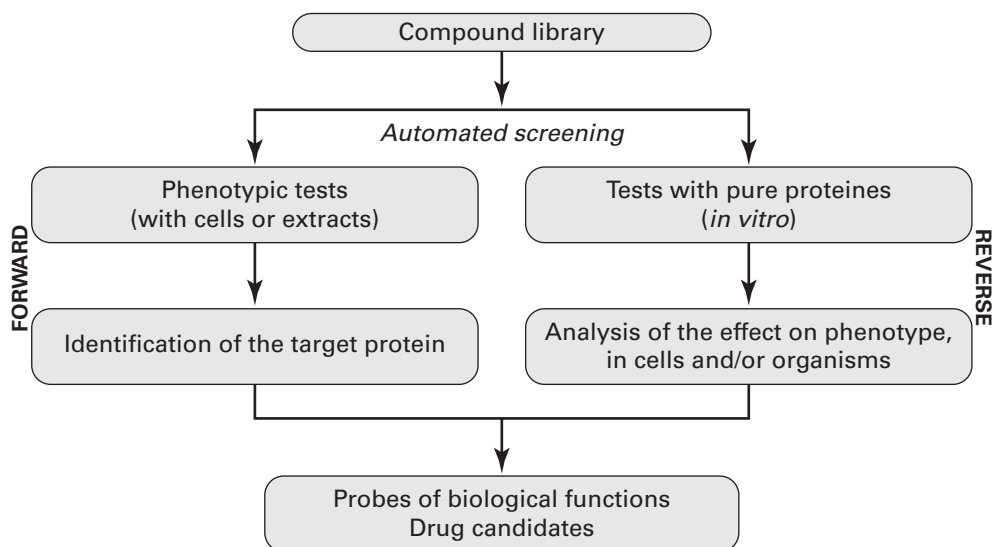


Fig. 8.2 - Schematic representation of chemical genetic strategies

- » The reverse approach has allowed selection of molecules active on known proteins. The action of these molecules on whole organisms has led to expected phenotypes, as with protease inhibitors used in the treatment of AIDS (DE CLERQ, 2004), or even to unexpected phenotypes, as with an inhibitor of cGMP-dependent phosphodiesterase, proving itself capable of restoring erectile function, whereas the initial aim of its development was the treatment of cardiac problems, including chest pains (angina pectoris) (MCCULLOUGH, 2004).
- » The forward approach has led to the discovery of important proteins, targets of pre-existing drugs such as cyclooxygenase, for example, the target of aspirin (TARNAWSKI and CAVES, 2004).

Chemical genetics is in a sense a new outlook upon, and a change of scale from, classical pharmacology, combining recent technological developments such as combinatorial chemistry, miniaturisation and automation of biological tests to isolate new bioactive components with potential therapeutic properties.

Forward chemical genetics, which is of particular interest to us here, requires three principal elements: a collection of chemical compounds, a test that can be adapted to robotic handling, and some methods permitting the identification of the compound's biological target.

8.4. CHEMICAL LIBRARIES FOR CHEMICAL GENETICS

Different types of chemical library are available for screening (chapters 1 and 2) including collections of natural substances, as well as different libraries assembled from molecules arising from synthetic chemistry. These libraries may be composed of existing molecules or molecules synthesised *de novo*. Depending on the chemical strategies adopted during the molecular syntheses, two types of chemical library produced by synthetic chemistry can be distinguished: targeted libraries and 'diversity-oriented' libraries (chapters 1 and 2). In practice, it is not always simple to assemble a collection of molecules that completely answers all of the needs of chemical genetic approaches. However, certain characteristics should be taken into account in order to choose a chemical library suitable for the approaches in forward chemical genetics.

Does an ideal chemical library exist for forward chemical genetics? The question of the size and the quality of a chemical library for forward chemical genetics is important. We shall deal here with some questions relating to the suitability of a chemical library for phenotypic screening, in terms of:

- › chemical library size,
- › concentration suitable for tests with cells,
- › physicochemical properties (diversity, complexity, accessibility in the cell),
- › potential as a molecular tool for research (abundance, capacity for functionalisation).

8.4.1. CHEMICAL LIBRARY SIZE

The larger the size of a chemical library, the greater the chance will be of isolating a bioactive molecule. The pharmaceutical industry has frequently used collections of several hundred thousand compounds during screening campaigns on isolated protein targets. The tests employed in phenotypic screening are often longer and more complex. The multiplication of the steps raises the unit cost of reagents and consumables, and the use of cells requires cell-culture methods which increase the bill. It is therefore necessary to find a practical compromise. An analysis of the literature in this field indicates that the size of chemical libraries employed in forward chemical genetic approaches varies from 5,000 to 30,000 compounds.

8.4.2. CONCENTRATION OF MOLECULES

One of the parameters to be set when screening is the concentration of the chemical library molecules to be tested. In theory, the higher the concentration of the molecules to be screened, the more likely it will be to obtain hits, at the expense of specificity (WALTER and NAMCHUK, 2003). The relative specificity of the hits can thereafter be evaluated by carrying out, for example, secondary screening at two concentrations, and then by plotting dose-effect curves (chapter 5).

The question of molecular concentration in phenotypic tests is a tricky one, since in contrast to an *in vitro* test it is difficult to know the intracellular concentrations of the compounds. On the one hand, the solubility of a compound in a biological environment is rarely known (LIPINSKI and HOPKINS, 2004); on the other hand, the intracellular compound concentration depends on the equilibria between the entrance and exit of the compounds across the plasma membrane. These equilibria are very variable depending on the molecule type, its affinity for the target, and the cell type (JORDAN and WILSON, 1999).

Another factor to take into consideration is the DMSO concentration. DMSO is a very frequently used solvent in chemical libraries (chapter 1; see also the study of compound stability in DMSO by CHENG *et al.*, 2003). Now, DMSO at a concentration higher than 1% exhibits toxic effects on cells. The dilution of library molecules must therefore accommodate both a sufficiently high concentration of molecules and a sufficiently low DMSO concentration. In practice, an analysis of the literature indicates that screening with chemical libraries at a concentration of 10 to 50 μM has proved to be a good compromise.

8.4.3. CHEMICAL STRUCTURE DIVERSITY

In undertaking phenotypic screening with cells, for example, one has absolutely no prior idea in principle about the nature of the protein targeted. Not only membrane receptors, but in fact every cellular protein is a potential target. An ideal library should therefore contain molecules able to interact specifically and selectively with each protein in the proteome. Intuitively, we would expect that the molecules

comprising this ideal library ought to display great structural diversity. An analysis of chemical diversity can be achieved by different numerical methods (chapters 11, 12 and 13). However, although a correlation often exists between the diversity of chemical structures and the diversity of biological activities, there is not necessarily a direct correspondence:

» **Very diverse structures can interact with the same target**

For example, it is possible to assemble a collection of about ten very different compounds based on their chemical structures but all of which bind to tubulin (fig. 8.3). This is a case of a collection exhibiting a large structural diversity but a poor diversity of biological effects (STOCKWELL, 2004).

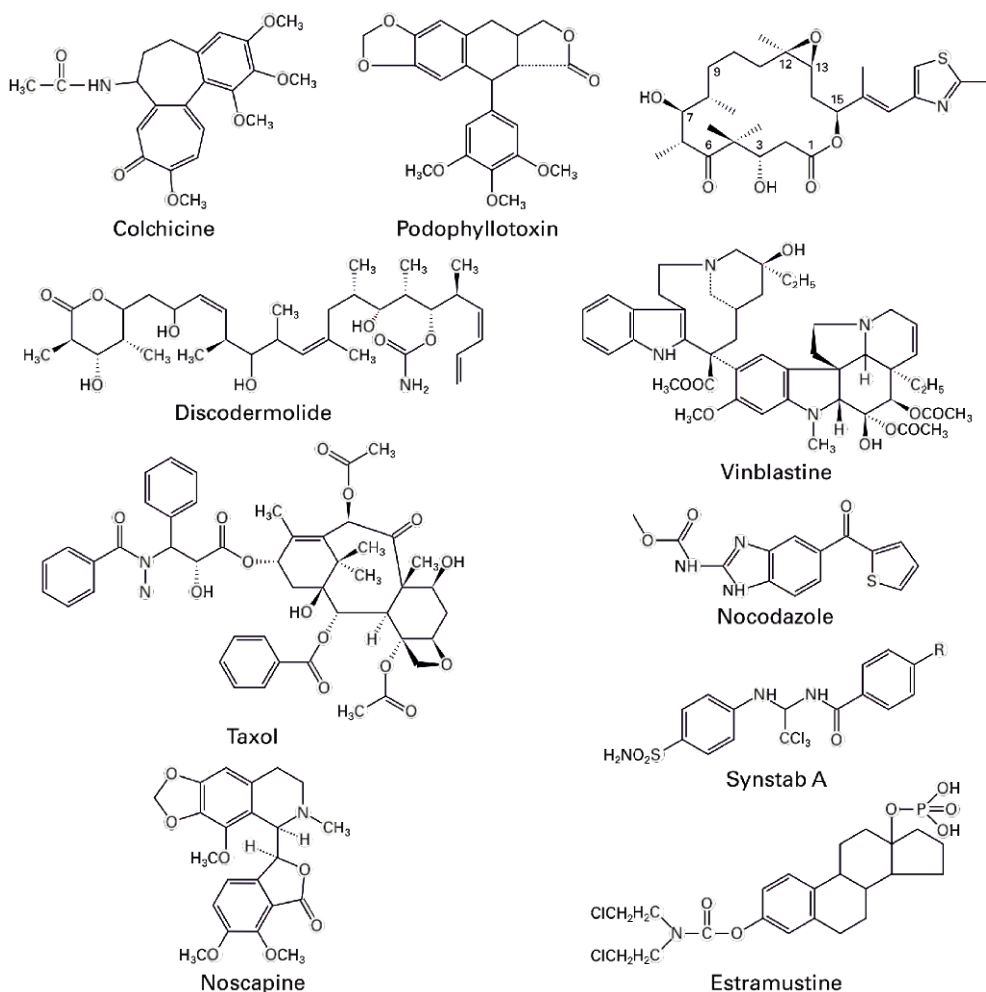


Fig. 8.3 - Molecules with very varied structures are able to bind to tubulin

» **Very similar structures can act upon numerous targets**

Certain targeted collections are thus of particular interest in forward chemical genetics. For example, the synthesis, using combinatorial chemistry, of a library

of guanine-mimetic molecules has permitted the different biological processes that depend physiologically on small molecules derived from guanosine and its different phosphorylated forms (cGMP, GMP, GDP, GTP) to be probed, as well as cofactors that are derivatives of guanine, such as folate (MILLER *et al.*, 2004).

Although a correlation between the diversity of chemical structures and the diversity of biological activities permits definition of a chemical library's potential based on information about molecular structures, the history of successful screens using a given library in phenotypic tests is an even more precious source of information. This may seem surprising, but one does not yet know the optimal type of chemical structure able to interact with the biological systems (STOCKWELL, 2000). The current developments in chemistry, combined with a comparative analysis of the results of different screenings by powerful computational methods, should in the future enable a better understanding of the fundamental principles governing the relationship between a molecule's chemical structure, its efficacy and its biological selectivity (PETERSON *et al.*, 2001; see also chapter 15).

8.4.4. COMPLEXITY OF MOLECULES

Are simple or really complex molecules preferred? Numerous biological processes depend upon protein-protein interactions. It is generally viewed as essential to increase the size and number of rigidifying and protein-binding elements in small molecules for close and specific bonds at protein-protein interaction sites (SCHREIBER, 2000). One of the interests of phenotypic screening is to perturb the system in all its complexity, in particular at the level of protein-protein interactions. This justifies why, in phenotypic screening, a fraction of the library molecules used should be endowed with great structural complexity.

Aside from these considerations, it is important to bear in mind that by searching for compounds active towards the whole cell, numerous potential targets, known or unknown, are thus probed, and finally the cell itself will guide the screening towards the most suitable and accessible target.

8.4.5. ACCESSIBILITY OF MOLECULES TO CELLULAR COMPARTMENTS

The molecules in an ideal chemical library, suitable for phenotypic screening with cells, should be capable of crossing biological membranes, be stable in the cellular environment and be soluble in water (the biological medium) and in dimethylsulfoxide (DMSO), the organic solvent used in most chemical libraries (chapter 1). Some of these properties can be calculated or measured. For example, the lipophilicity is measured by the log *P* (see also chapter 12), which represents the equilibrium between a polar (aqueous) phase and an apolar phase (often *n*-octanol, appropriate for the simulation of biological membranes).

Christopher LIPINSKY and co-workers (LIPINSKI *et al.*, 2001) proposed a set of rules for the design of an optimised chemical library in terms of its accessibility, in particular in relation to drug candidates destined to be absorbed orally. These well-known rules, known as **LIPINSKI's rule of 5** (or **LIPINSKI's rules**) were established from a comparison of the properties of marketed drugs and of 'non-drug' compounds.

LIPINSKI's rule of 5:

- › a molecular mass less than 500 daltons
- › a number of hydrogen donors less than 5
- › a number of hydrogen acceptors less than 10
- › a theoretical octanol/water partition coefficient ($\log P$) less than 5

However, these rules were established with the aim of helping the design of drugs, which for the most part, are destined to be absorbed after oral administration. The requirements would not be the same if the aim of screening were the isolation of molecules to serve as 'research tools', without administration to animals or man (LIPINSKI and HOPKINS, 2004).

8.4.6. THE ABUNDANCE OF MOLECULES

A key step in forward chemical genetic strategies is the identification of the target from the ligands selected by screening. Often during this step the ligand itself is utilised as bait in order to catch the protein target (see further on). In order to conduct these approaches, it is therefore indispensable to have available an abundant source of active molecules. This is an important criterion to take into consideration when assembling the chemical library.

8.4.7. THE POSSIBILITY OF FUNCTIONALIZING THE MOLECULES

Targeted collections exist where, from the moment of synthesis, the systematic introduction of a motif was planned, permitting both the subsequent binding of a molecule on a support and its labelling. This type of library, although not frequent, is very convenient in forward chemical genetic strategies as it enables the protein target to be more readily isolated from the molecules selected during screening (KHERSONSKY *et al.*, 2003; MITSOPOULOS *et al.*, 2004).

8.5. PHENOTYPIC TESTS WITH CELLS

Phenotypic screening is generally conducted on living cells or on complex extracts, such as those from *Xenopus* eggs (example 8.1; [fig. 8.4](#)).

Example 8.1 - *Xenopus* cell screening to identify molecules active on the cytoskeleton

Xenopus is a species of batrachian and the female lays oocytes from 1 to 1.3 mm in diameter. Acellular extracts obtained from these oocytes are used as a model system by biologists. Indeed, different steps of the cell cycle – such as the formation of the mitotic

spindle – can be reproduced and observed in these extracts. The mitotic spindle is a structure composed of microtubules, which forms at the time of cell division, and is responsible for the correct separation of chromosomes into the two daughter cells.

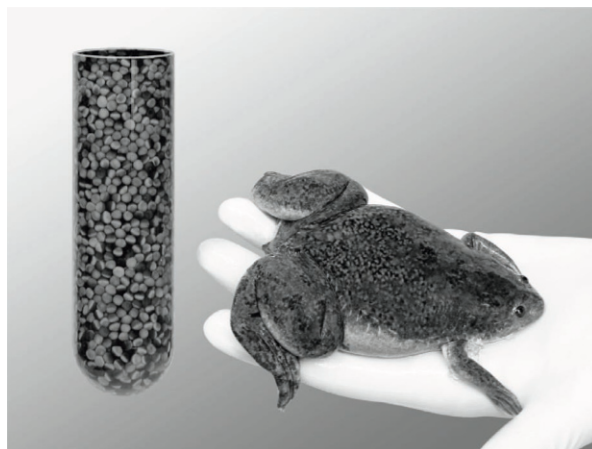


Fig. 8.4 - *Xenopus* (a) and a test-tube containing eggs (b)
Each egg measures approximately 1.2 mm in diameter.

Xenopus egg extracts were used in the search for inhibitors of signalling pathways regulating the assembly and the polymerisation of the actin cytoskeleton, by screening a library of more than 26 000 compounds (PETERSON *et al.*, 2001). While screening for inhibitors of the entire signalling pathway numerous potential targets, both known and unknown, were probed by ultimately letting biology guide the screening towards the most suitable target (PETERSON and MITCHISON, 2002). This study enabled the identification of an inhibitor of a protein known for its role in the integration of signals, the protein N-WASP (Neural Wiskott Aldrich Syndrome Protein). The inhibitor acts by preventing the conformational modification of N-WASP towards its active form.

More recently, *Xenopus* egg extracts have been used in the search for novel regulatory proteins of **microtubule** stability using a test involving mitotic spindle formation. The formation of a bipolar mitotic spindle was visualised under the microscope. This phenotypic screening permitted the selection of a molecule called diminutol. **The protein target of this molecule has since been identified:** an NADP-dependent oxidoreductase. The investigators subsequently confirmed that this enzyme does indeed regulate the state of microtubule polymerisation (WIGNALL *et al.*, 2004). ▀

In cells, some proteins or post-translational modifications can be characteristic of a given functional state (FONROSE *et al.*, 2007; LAFANECHÈRE, 2008). This has been exploited in the setup of phenotypic screens with whole cells, based on the use of antibodies specifically directed against the protein of interest with a method derived from ELISA (Enzyme-Linked Immunosorbent Assay): the **cytoblot** (the antibody is used to identify a given cellular state on fixed, whole cells; example 8.2).

Example 8.2 - screening of cells by the 'cytoblot' technique

The '**cytoblot**' technique was employed in a famous study aiming to identify compounds capable of disrupting the mitotic machinery without targeting tubulin (see example 8.1; MAYER *et al.*, 1999). In an initial screen of a library of 16 320 compounds, the molecules blocking mitosis were identified by detecting a protein uniquely phosphorylated during mitosis: nucleolin. To do this, after incubation with each library compound the cells were fixed ▀

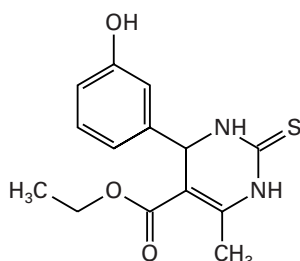


Fig. 8.5 - Monastrol

and the wells containing the cells undergoing mitosis were identified thanks to a specific antibody to phosphonucleolin. The hit **molecules were subsequently subjected to a secondary round of screening where their activity on tubulin** assembly in vitro was evaluated, permitting elimination of the compounds acting directly on tubulin. This strategy enabled the identification of **monastrol**, a specific inhibitor of a molecular motor (the kinesin Eg5). Monastrol represents the lead compound of a novel class of antimetotics, potentially harbouring anti-cancerous properties. ▀

Since these first studies, others have used the cytoblot technique to search for inhibitors of DNA synthesis (STOCKWELL *et al.*, 2004), of tubulin deacetylation (HAGGARTY *et al.*, 2003), of a pathway involving Janus kinase (BLASKOVICH *et al.*, 2003) or to search for agents that could stabilize or depolymerise microtubules (VASSAL *et al.*, 2006; LAFANECHÈRE, 2008).

The development of **automated microscopes** capable of acquiring images of cells in microplates now allows the screening of chemical libraries acting on biological mechanisms such as nuclear translocation (DING *et al.*, 1998), cell migration (YARROW *et al.*, 2004) or parasitic invasion (WARD *et al.*, 2002; CAREY *et al.*, 2004). The results can be analysed directly or else with the help of form-recognition software capable of distinguishing the phenotypes of interest. The speed of screening with imaging is much slower due to the longer image acquisition and analysis times. However, their **high information content** (simultaneous analysis of several parameters) justifies their development (YARROW *et al.*, 2003; SOLEILHAC *et al.*, 2010).

Multiparametric approaches (fig. 8.6) can be exploited to establish a **quantitative cytological profile** of the effect of a given compound. It is thus possible to create a database in which the effects of known compounds on a set of physiological processes are stored along with suitable descriptors. This database can then be used to predict the mode of action of unknown compounds from their cytological profiles with the help of appropriate algorithms (PERLMAN *et al.*, 2004).

8.6. METHODS TO IDENTIFY THE TARGET

The identification of protein targets of the molecules selected by phenotypic screening is one of the challenges of chemical genetics and often constitutes the most difficult step (BURDINE and KODADEK, 2004).

An approach often taken is the purification of the target from cell extracts, by means of a **labelled ligand** (either radioactive, or coupled to biotin) or a ligand **immobilised on an affinity support** (WIGNALL *et al.*, 2004; BURDINE and KODADEK, 2004; KNOCKAERT *et al.*, 2000). Different methods such as sequencing, immunological tools or proteomics can thereafter be used to characterise the molecular nature of the target.

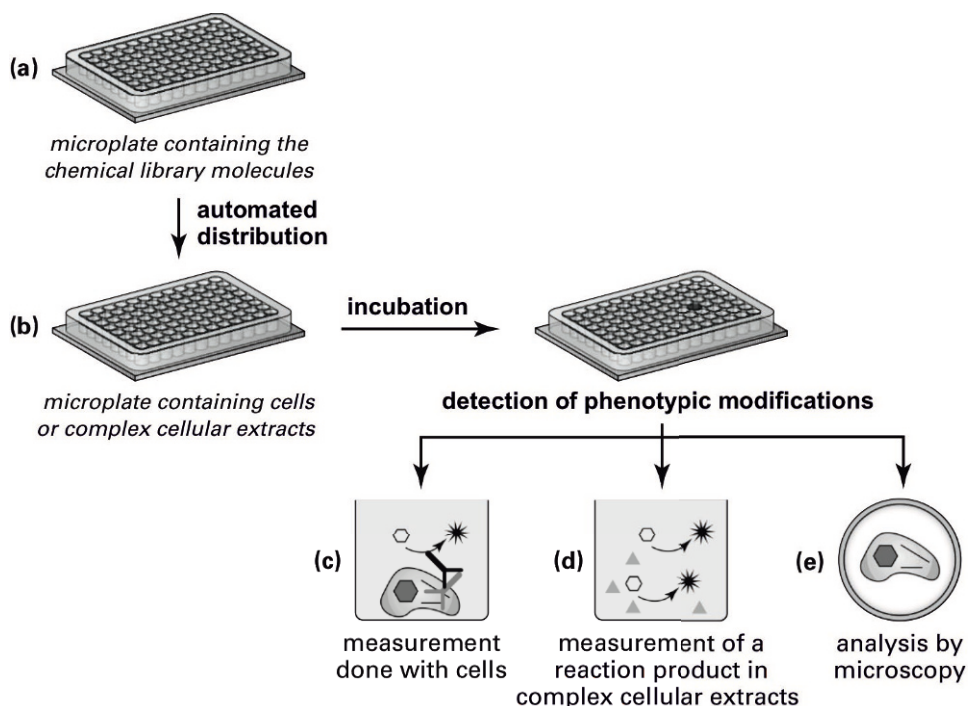


Fig. 8.6 - Principle of automated pharmacological screening

The chemical library is formatted in microplates **(a)**. Assisted by liquid-handling robots, the compounds are distributed into microplates containing cells or complex extracts **(b)**. After an incubation period allowing the compounds to act, different methods permit an analysis of the effects of each compound on the biological phenomenon of interest: **(c)** measurements on cells are carried out, for example, with specific antibodies or reporter genes; **(d)** the product of a reaction is measured in complex cellular extracts; **(e)** cell morphological changes, modifications in the subcellular localisation of a given protein or other modifications are visualised under the microscope.

The search for the target implies, of course, the synthesis of the modified ligand and the verification that the derivatives conserve their activity. It is important to underline that purification does not always allow a distinction to be made between an abundant cellular target of weak affinity and a target presenting stronger affinity but represented in small amounts. A comparative or differential analysis of the profiles of those proteins retained on affinity supports, comprising either the active molecule or an inactive structural analogue, can partially resolve these problems.

Another simple approach consists of testing one presumed target after another. Using this approach the target of monastrol was identified (example 8.2; *MAYER et al.*, 1999). Indeed, the addition of monastrol leads to the formation of a monopolar (monoastral) mitotic spindle instead of the normal bipolar spindle. Such a phenotype had already been described for a family of mitotic kinesins. This observation led the investigators to test *in vitro* the effect of monastrol on the kinesin Eg5. These studies demonstrated that monastrol specifically inhibited this motor

whereas a structural analogue, without any effect on the cellular phenotype, proved to be inactive in an activity test with Eg5. This point highlights the **importance of negative controls** when studying the specificity of a molecule.

Generally speaking, the **specificity of a molecule** with respect to a protein target is difficult to establish. For example, it has been shown that a complex molecule of appreciable size, like taxol, which binds with high affinity to tubulin, can also have other cellular targets such as the anti-apoptotic protein Bcl-2 (RODI *et al.*, 1999). These interactions, often termed secondary, could contribute to the anti-proliferative effect. For other molecules, this type of secondary interaction may be at the origin of phenotypic effects that remain unexplained. Due to this, the link between a target and a phenotype established only from an analysis of *in vitro* data must be considered carefully. A comparison between cellular phenotypes elicited by small molecules with the phenotypes obtained from screening with RNA interference (RNAi) libraries (KIGER *et al.*, 2003) can give a good indication about the targets involved (EGGERT *et al.*, 2004).

There is a powerful method permitting the entire proteome to be scanned in order to isolate all of the targets of a chemical compound (BECKER *et al.*, 2004): the chemical three-hybrid system. This method, derived from the yeast two-hybrid system, first requires coupling of the active compound to a compound like methotrexate (or dexamethazone; BAKER *et al.*, 2003). The chemical hybrid compound ‘active molecule – methotrexate’ (chemical dimeriser) can then be used in a two-hybrid system, to associate to the protein targets of the bioactive molecule synthesised from an expression bank with the glucocorticoid receptor fused to the DNA-binding domain. This three-hybrid strategy has the following advantages:

- › helping to identify the target of a chemical molecule, even when poorly represented,
- › accessing directly the protein’s gene,
- › identifying other possible targets (prediction of secondary effects).

In practice, this method requires that the molecules of interest cross the yeast cell wall and that the affinity between the target and the molecule be strong (KLEY, 2004).

Target identification may in some cases rely on the generation of **mutant cells (or organisms) resistant to the molecule’s action**, and then on the identification of the gene and its expression product responsible for this resistance (WARD *et al.*, 2002; DOBROWOLSKI and SIBLEY, 1996).

Lastly, thanks to the development of screening by imaging, more recent strategies have appeared that aim to predict the pathway targeted by a new compound, by comparing its cellular response profile – by measuring different parameters – with those obtained for a set of reference compounds (PERLMAN *et al.*, 2004; EGGERT and MITCHISON, 2006).

8.7. CONCLUSIONS

The contribution of traditional genetics to dissecting complex cellular processes needs no further proof. Nevertheless, numerous situations exist in which chemical genetics can prove to be very useful, in particular for organisms in which genetic manipulation is difficult (WARD *et al.*, 2002). Small molecules are particularly useful for knocking out the function of a vital protein, the expression of which is not possible to suppress. Finally, even in cases where cellular production of the protein of interest can be deleted, either by conditional or non-conditional gene knockout or by RNA interference (RNAi), the genetic approaches are not suitable for the study of dynamic processes that occur on a time-scale of the order of seconds or minutes. Small molecules directly act on the gene product and, if their action is reversible, they can be added to and then removed from the medium in order to disrupt the function of the protein of interest in a time-dependent manner. Coupled with methods such as videomicroscopy of living cells, this approach can prove to be rich in information content.

So as to gain the maximum benefit from phenotypic screening, the tests and the hit selection strategies must be designed carefully and cleverly in order to limit the selection of molecules of low specificity and interest. Furthermore, the validation of hits and the analysis of their cellular effects needs a whole set of techniques and technological developments in biochemistry and in molecular and cellular biology. This is why this sort of approach is inseparable from a research environment having expertise in all of these areas.

8.8. REFERENCES

- BAKER K., SENGUPTA D., SALAZAR-JIMENEZ G., CORNISH V.W. (2003)
An optimized dexamethasone-methotrexate yeast 3-hybrid system for high-throughput screening of small molecule-protein interactions. *Anal. Biochem.* **315**: 134-137
- BECKER F., MURTHI K., SMITH C., COME J., COSTA-ROLDAN N., KAUFMANN C., HANKE U., DEGENHART C., BAUMANN S., WALLNER W., HUBER A., DEDIER S., DILL S., KINSMAN D., HEDIGER M., BOCKOVICH N., MEIER-EWERT S., KLUGE A.F., KLEY N. (2004)
A three-hybrid approach to scanning the proteome for targets of small molecule kinase inhibitors. *Chem. Biol.* **2**: 211-223
- BLASKOVICH M.A., SUN J., CANTOR A., TURKSON J., JOVE R., SEBTI S.M. (2003)
Discovery of JSI-124 (cucurbitacin I), a selective Janus kinase/signal transducer and activator of transcription 3 signaling pathway inhibitor with potent antitumor activity against human and murine cancer cells in mice. *Cancer Res.* **63**: 1270-1279
- BREDEL M., JACOBY E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **4**: 262-275
- BURDINE L., KODADEK T. (2004) Target identification in chemical genetics: the (often) missing link. *Chem. Biol.* **5**: 593-597

- CAREY K.L., WESTWOOD N.J., MITCHISON T.J., WARD G.E. (2004) A small-molecule approach to studying invasive mechanisms of *Toxoplasma gondii*. *Proc. Natl Acad. Sci. U.S.A.* **101**: 7433-7438
- CHENG X., HOCHLOWSKI J., TANG H., HEPP D., BECKNER C., KANTOR S., SCHMITT R. (2003) Studies on repository compounds stability in DMSO under various conditions. *J. Biomol. Screen.* **8**: 292-304
- DARLAND T., DOWLING J.E. (2001) Behavioral screening for cocaine sensitivity in mutagenized zebrafish. *Proc. Natl Acad. Sci.* **98**: 11691-11696
- DE CLERQ E. (2004) Antiviral drugs in current clinical use. *J. Clin. Virol.* **30**: 115-133
- DING G.J., FISCHER P.A., BOLTZ R.C., SCHMIDT J.A., COLAIANNE J.J., GOUGH A., RUBIN R.A., MILLER D.K. (1998) Characterization and quantitation of NF-kappaB nuclear translocation induced by interleukin-1 and tumor necrosis factor-alpha. Development and use of a high capacity fluorescence cytometric system. *J. Biol. Chem.* **273**: 28897-28905
- DOBROWOLSKI J.M., SIBLEY L.D. (1996) *Toxoplasma* invasion of mammalian cells is powered by the actin cytoskeleton of the parasite. *Cell* **84**: 933-939
- EGGERT U.S., KIGER, A.A., RICHTER C., PERLMAN Z.E., PERRIMON N., MITCHISON T.J., FIELD C.M. (2004) Parallel chemical genetics and genome-wide RNAi screens identify cytokinesis inhibitors and targets. *PLoS Biol.* **2**: e379
- EGGERT U.S., MITCHISON T.J. (2006) Small molecule screening by imaging. *Curr. Opin. Chem. Biol.* **10**: 232-237
- FONROSE X., AUSSEIL F., SOLEILHAC E., MASSON V., DAVID B., POUNY I., CINTRAT J-C., ROUSSEAU B., BARETTE C., MASSIOT G., LAFANECHÈRE L. (2007) Parthenolide inhibits tubulin carboxypeptidase activity. *Cancer Res.* **67**:3371-3378
- HAGGARTY S.J., KOELLER K.M., WONG J.C., GROZINGER C.M., SCHREIBER S.L. (2003) Domain-selective small molecule inhibitor of histone deacetylase 6 (HDAC6)-mediated tubulin deacetylation. *Proc. Natl Acad. Sci. U.S.A.* **100**: 4389-4394
- JORDAN, M.A., WILSON L. (1999) The use and action of drugs in analyzing mitosis. In *Methods in Cell Biology*, RIEDER L. Ed., New York, Vol. **61**: 267-295
- KHERSONSKY S.M., JUNG D.W., KANG T.W., WALSH D.P., MOON H.S., JO H., JACOBSON E.M., SHETTY V., NEUBERT T.A., CHANG Y.T. (2003) Facilitated forward chemical genetics using a tagged triazine library and zebrafish embryo screening. *J. Am. Chem. Soc.* **125**: 11804-11805
- KIGER A.A., BAUM B., JONES S., JONES M.R., COULSON A., ECHEVERRI C., PERRIMON N. (2003) A functional genomic analysis of cell morphology using RNA interference. *J. Biol.* **2**: 27
- KLEY N. (2004) Chemical dimerizers and three-hybrid systems: scanning the proteome for targets of organic small molecules. *Chem. Biol.* **5**: 599-608
- KNOCKAERT T.M., GRAY N., DAMIENS E., CHANG Y.T., GRELLIER P., GRANT K., FERGUSSON D., MOTTRAM J., SOETE M., DUBREMETZ J.F., LE ROCH K., DOERIC C., SCHULTZ P., MEIJER L. (2000) Intracellular targets of cyclin-dependent kinase inhibitors:

identification by affinity chromatography using immobilised inhibitors.

Chem. Biol. **7**: 411-422

LAFANECHÈRE L. (2008) Chemogenomics and cancer chemotherapy: cell-based assays to screen for small molecules that impair microtubule dynamics.

Comb. Chem. High Throughput Screen. **11**: 617-623

LIPINSKI C., HOPKINS A. (2004) Navigating chemical space for biology and medicine.

Nature **432**: 855-861

LIPINSKI C.A., LOMBARDO F., DOMINY B.W., FEENEY P.J. (2001)

Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting. *Adv. Drug Deliv. Rev.* **46**: 3-26

MAYER T.U., KAPOOR T.M., HAGGARTY S.J., KING R.W., SCHREIBER S.L., MITCHISON T.J.

(1999) Small molecule inhibitor of mitotic spindle bipolarity identified in a phenotype-based screen. *Science* **286**: 971-974

MCCULLOUGH A. (2004) Phosphodiesterase-5 inhibitors:

clinical market and basic science comparative studies. *Curr. Urol. Rep.* **5**: 451-459

MITSOPOULOS G., WALSH D.P., CHANG Y.T. (2004) Tagged library approach to chemical genomics and proteomics. *Curr. Opin. Chem. Biol.* **1**: 26-32

PERLMAN Z.E., SLACK M.D., FENG Y., MITCHISON T.J., WU L.F., ALTSCHULER S.J. (2004)

Multidimensional drug profiling by automated microscopy. *Science* **306**: 1194-1198

PETERSON J., LOKEY R.S., MITCHISON T.J., KIRSCHNER M.W. (2001) A chemical inhibitor of N-WASP reveals a new mechanism for targeting protein interactions.

Proc. Natl Acad. Sci. U.S.A. **98**: 10624-10629

PETERSON J.R., MITCHISON T.J. (2002) Small molecules, big impact: a history of chemical inhibitors and the cytoskeleton. *Chem. Biol.* **9**: 1275-1285

RODI D.J., JANES R.W., SAGANEE H.J., HOLTON R.A., WALLACE B.A., MAKOWSKI L.

(1999) Screening of library of phage-displayed peptides identifies human Bcl-2 as a Taxol-binding protein. *J. Mol. Biol.* **285**: 197-203

ROOT D.E., KELLEY B.P., STOCKWELL B.R. (2002) Global analysis of large-scale chemical and biological experiments. *Curr. Opin. Drug. Discov. Dev.* **5**: 355-360

SCHREIBER S.L. (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**: 1964-1969

SHELANSKI M.L., TAYLOR E.W. (1967) Isolation of a protein subunit from microtubules.

J. Cell Biol. **34**: 549-554

SOLEILHAC E., NADON R., LAFANECHÈRE L. (2010) High-content screening for the discovery of pharmacological compounds: advantages, challenges and potential benefits of recent technological developments. *Expert Opin. Drug Discov.* **5**: 135-144

STOCKWELL B.R. (2000) Chemical genetics: ligand-based discovery of gene function.

Nat. Rev. Genet. **1**: 116-125

- STOCKWELL B.R. (2004) Exploring biology with small organic molecules. *Nature* **432**: 846-854
- STOCKWELL B.R., HAGGARTY S.J., SCHREIBER S.L. (1999) High-throughput screening of small molecules in miniaturized mammalian cell-based assays involving post-translational modifications. *Chem. Biol.* **6**: 71-83
- TARNAWSKI AS, CAVES TC. (2004) Aspirin in the XXI century: its major clinical impact, novel mechanisms of action, and new safer formulations. *Gastroenterology* **127**: 341-343
- VASSAL E., BARETTE C., FONROSE X., DUPONT R., SANS-SOLEILHAC E., LAFANECHÈRE L. (2006) Miniaturization and validation of a sensitive multi-parametric cell-based assay for the concomitant detection of microtubule-destabilizing and microtubule-stabilizing agents. *J. Biomol. Screen.* **11**: 377-389
- WALTERS WP, NAMCHUK M. (2003) Designing screens: how to make your hits a hit. *Nat. Rev. Drug Discov.* **2**: 259-266
- WARD G.E., CAREY K.L., WESTWOOD N.J. (2002) Using small molecules to study big questions in cellular microbiology. *Cell Microbiol.* **4**: 471-482
- WIGNALL S.M., GRAY N.S., CHANG Y-T., JUAREZ L., JACOB R., BURLINGAME A., SCHULTZ P., HEALD R. (2004) Identification of a novel protein regulating microtubule stability through a chemical approach. *Chem. Biol.* **11**: 135-146
- YARROW J.C., FENG Y., PERLMAN Z.E., KIRCHHAUSEN T., MITCHISON T.J. (2003) Phenotypic screening of small molecule libraries by high throughput cell imaging. *Comb. Chem. High Throughput Screen.* **6**: 279-286
- YARROW J.C., PERLMAN Z.E., WESTWOOD N.J., MITCHISON T.J. (2004) A high-throughput cell migration assay using scratch wound healing, a comparison of image-based readout methods. *BMC Biotechnol.* **4**: 21

Chapter 9

HIGH-CONTENT SCREENING IN FORWARD (PHENOTYPIC SCREENING WITH ORGANISMS) AND REVERSE (STRUCTURAL SCREENING BY NMR) CHEMICAL GENETICS

Benoît DÉPREZ

9.1. INTRODUCTION

High-content screening (HCS) aims not only to isolate molecules that are active towards a biological target but also to obtain the maximum amount of information *during* the screening about the effect of the molecule on this target. Here the notion of ‘biological target’ covers both molecularly defined biomolecules and complex biological systems. When the targets are identified molecularly, the effects of the targeted molecules can be studied on the isolated target *in vitro* (structural screening in parallel) or on the target in its cellular context, *in vivo*. These are procedures in reverse chemical genetics (chapter 8). It is also possible to search for molecules active on metabolic or signalling pathways without any molecular characterisation of the target. Phenotypic screening with cells or whole organisms are approaches in forward chemical genetics (chapter 8).

Traditionally, high-content screening was conducted secondarily to the primary high-throughput screening. More recently, thanks notably to the development of technologies permitting the parallel measurement of multiple parameters, complex cellular assays are used straight after the primary screening to annotate the effects of the molecules targeted. This chapter explores the colossal potential that is offered by the approaches both in **high-content screening with phenotypic screens** (forward chemical genetics) and in **structural screening in parallel** (reverse chemical genetics), at last achievable thanks to the progress in **nuclear magnetic resonance (NMR)**.

9.2. BENEFITS OF HIGH-CONTENT SCREENING

9.2.1. SUMMARISED COMPARISON OF HIGH-THROUGHPUT SCREENING AND HIGH-CONTENT SCREENING

High-content screening techniques occupy two critical steps in the drug discovery process:

- › the **validation of novel therapeutic targets** (functional genomics and chemogenomics),
- › the **discovery of novel lead molecules**, in particular against targets known to be difficult to screen in classical cell-free settings.

High-content screening can be distinguished from classic high-throughput screening by the nature of the question asked before the experiment and by the nature of the answer revealed by the experiment. **Classic high-throughput screening** poses a **univocal question**, the answer to which is a binary ‘negative/positive’ answer, with the aim of separating a population of biological entities (protein target, compound, materials) into two sub-populations. The very high number of molecules tested restricts the manner in which the result can be read, which must be simple and robust. This robustness is however sometimes achieved to the detriment of the quality and relevance of the information obtained. **High-content screening** aims to answer an **open question** of the type “*how does the molecule or the target exert its effect on the system under study?*”. The result permits the objects to be sorted into more than two categories, and to classify them by order of priority (table 9.1).

Table 9.1 - Comparison of the two screening approaches

Screening	High-throughput	High-content
Question asked	univocal	open
Number of molecules tested	high	low
Rate	rapid	moderate
Objective	binary sorting	multidimensional sorting and ranking
Type of signal measured	one-dimensional signal (fluorescence, absorbance, radioactivity etc.)	multidimensional signal (image, spectrum, integration signal)
Analysis models	structure-activity relationship	structure-property relationship
Quality factor	Z factor *	?

* The Z factor is explained in chapter 3.

9.2.2. ADVANTAGES OF HIGH-CONTENT SCREENING FOR THE DISCOVERY OF NOVEL THERAPEUTIC TARGETS

A major challenge in pharmaceutical research is the discovery of novel **therapeutic targets**. The first works elucidating animal and human genomes have now

finished. The analysis of mutations involved in genetic diseases and the study of gene expression in different physiological and pathological contexts has allowed the identification of **genes whose products are potential therapeutic targets**. The next step, probably more difficult, consists of interpreting these data and selecting the targets. To define some of these genes, animal models have been selected: the fruitfly (*Drosophila melanogaster*), the zebrafish (*Danio rerio*) and the roundworm (a **nematode**, *Caenorhabditis elegans*).

Until the early years of this century, the proteins characterised by classical functional genomics constituted a considerable scientific and market capital. The miniaturisation of techniques in proteomics, as well as the sequencing of genomes, has stimulated such a sharp acceleration in biomedical research that all the other steps in drug discovery were for several years considered to be trivial. After this phase of euphoria the realities of pharmaceutical research have once again set in. The discovery of innovative targets is today considered once more to be a difficult and essential step in the process of finding novel drugs, even if it is no longer the sole key to it. In the current economic model, the **exogenous ligand** (small molecule brought into proximity of the target) which can eventually constitute an **active principle of a drug**, becomes again also the **main object of industrial property**. Forward chemical genetic approaches (chapter 8), which validate a target by characterising the effects of small molecules on a biological model, are thus interesting in two respects, since in one step they can supply both a validated target and a lead compound. The two main prerequisites of this strategy are:

- › development of a relevant biological model (reflecting as well as possible the physiopathology) that is miniaturisable. For biological relevance it is often necessary to use relatively complex models such as differentiated cells (chapter 8) and in certain cases, tissues or whole organisms.
- › the availability of a structurally diverse chemical library containing compounds suitable for use in a complex medium. In fact, the systems used for this type of screening most of the time present physical barriers (e.g. intestinal epithelium, biological membranes) and chemical ones (e.g. cytochromes P450, esterases) between the small molecules and the protein targets. The compounds must therefore be metabolically stable and have good bioavailability (chapter 8).

9.2.3. THE NEMATODE *CAENORHABDITIS ELEGANS*:

A MODEL ORGANISM FOR HIGH-CONTENT SCREENING

Several biotechnology firms were founded on the exploitation of model organisms for the discovery of novel targets and innovative drug candidates. In this context, the **nematode *Caenorhabditis elegans*** has aroused a particularly huge interest (KALETTA *et al.*, 2006). Linking the complexity and richness of information from multicellular organisms to the ease of culturing and working with unicellular microorganisms, the ***C. elegans* model** offers a number of advantages for the parallel screening of whole organisms. *C. elegans* has been well known to geneticists

since the 1980s and was the first multicellular organism whose genome was completely sequenced (BIRD *et al.*, 1999). *C. elegans* is **easy to observe** (it is transparent in the visible spectrum) and possesses a genome that is simple and relatively homologous to those of mammals. Furthermore, this organism is **sensitive to RNA interference** (RNAi) which permits repression of the expression of any gene at the post-transcriptional level in a transient and specific manner. Consequently, by comparing the phenotypes of nematodes ‘during RNAi’ to those of wild-type nematodes, the functions of a large number of human genes can be revealed, provided of course that these genes have a homologue in *C. elegans* (example 9.1).

Example 9.1 - phenotypic screening in *Caenorhabditis elegans* in order to identify novel therapeutic targets against depression

Depression is linked to a drop in the concentration of serotonin, a transmitter (*neurotransmitter*) of the message propagated by nerves, released at the zones of contact with nerve cells (*synapses*). Medicines capable of increasing the synaptic concentration of serotonin, like inhibitors of its reuptake, have a beneficial effect in depressive patients. Nevertheless, for some types of depression which resist these treatments, research is underway to find new pharmacological means to **augment the tonus associated with serotonin synapses**.

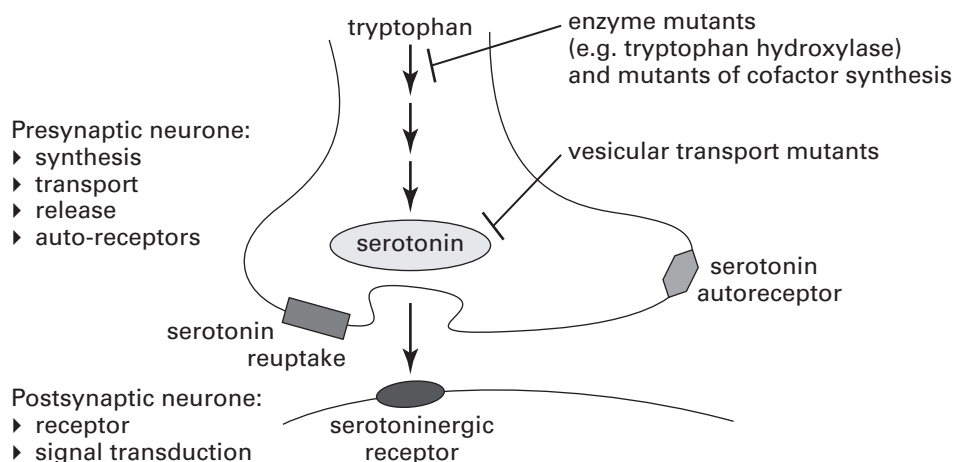


Fig. 9.1 - Synapse controlling the rhythm of pharynx contraction in *Caenorhabditis elegans*

The nematode possesses a very well-known serotoninergic synapse, which controls the contraction of its pharyngeal muscle.

Numerous mutants of enzymes involved in the biosynthesis of serotonin, its receptors and its transporters, both presynaptic and postsynaptic are known.

How can a very robust screening method be set up that is capable of specifically measuring the activity of this synapse, and consequently enable an evaluation of the activity of compounds or the impact on this synapse of repressing particular genes?

The critical **information can be** provided by a particular **phenotypic trait** of the nematode. An increase in the activity of the nematode's pharynx translates into an increase in the volume of liquid taken up by the worm per unit time. By measuring this volume the level of pharyngeal activity may be deduced, and consequently the activation level of the synapse due to serotonin.

In order to measure the volume imbibed by the nematode, it suffices to dissolve in the culture liquid the precursor to a fluorescent probe, activatable by the nematode's digestive enzymes. A quantity of the fluorophore proportional to the volume swallowed is activated by esterases in the worm's intestines. This gives a directly measurable proportional fluorescence, as the nematode is transparent in the visible and near-UV region of the spectrum. The probe used is calcein acetoxy-methyl ester.

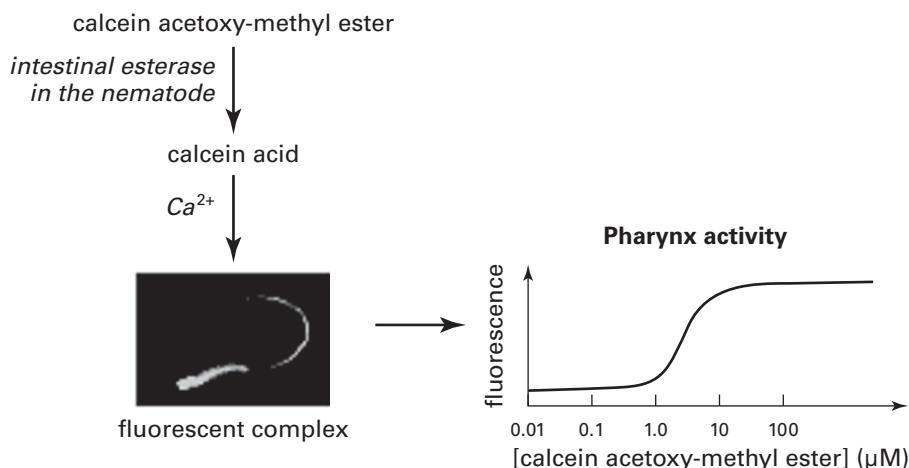


Fig. 9.2 - Measurement of pharyngeal activity in *Caenorhabditis elegans*

Thanks to a test based on this phenotype, screening could be undertaken. The majority of molecules selected proved to be ligands of known targets in the field of anti-depressant research like the 5HT1A autoreceptor or transporter, thus validating the approach. Some compounds, which were not ligands of known targets, were selected for entry into a target research programme, either by affinity chromatography or by genetic methods (searching for resistant mutants).

Another field in which *C. elegans* is of interest is **screening against targets whose functions are influenced by a complex environment**. Indeed, certain targets for which a pharmacological interest has already been demonstrated do not readily lend themselves to setting up classical functional screening assays that can be miniaturised and automated. Among these target types there are, for example, ion channels or enzymes with substrates that are difficult to handle (e.g. unstable or very lipophilic ones). High-throughput screening methods relate the activity of some of these targets to a directly measurable signal in the living nematode, in microtitre plates, as we saw in example 9.1 for pharyngeal activity. By using suitable promoters, it is actually possible to express an exogenous target in cells of the pharynx and to subject pharyngeal activity to the target's function. Compounds that modulate the activity of the target cause hyperactivation or inhibition of the pharynx muscles and a quantitative change in the fluorescence signal.

An interesting property of this type of functional screen is that it very rapidly provides quantitative information about the biological activity of the compound in the cellular environment. It is therefore possible to establish relationships between the molecular structure of the ligand and its bioactivity (chapters 11, 12 and 13) and to

orient very early on the development of families of molecules that modulate the activity of the target in the manner expected, something that classical tests do not permit. Lastly, the selected molecules, as well as being pharmacologically active, have proved their capacity to overcome the intestinal barrier and consequently show physicochemical profiles compatible with good bioavailability by oral administration.

9.2.4. ADVANTAGES OF HIGH-CONTENT SCREENING FOR REVERSE CHEMICAL GENETICS AND THE DISCOVERY OF NOVEL BIOACTIVE MOLECULES

In contrast to forward chemical genetic approaches, **structural screening** techniques are based on the structure determination of ligand-receptor complexes. The question asked here is: “*does the sample interact with the target and if yes, how, and with what affinity?*”. This type of structural screening is currently approached in two ways:

- › *in silico* virtual screening (chapter 16) for which the interaction between a protein structure and a small molecule is calculated computationally;
- › an experimental approach relying on parallel structural characterisation techniques. In the area of high-field NMR in particular, the appearance of new pulse sequences, which enable simplification of the spectra and selective or uniform labelling of the hydrogen and carbon atoms in proteins, has greatly facilitated what is termed NMR screening (HADJUK *et al.*, 1999; WIDMER *et al.*, 2004; SMET *et al.*, 2005)

NMR screening benefits from progress in the techniques of **high-field NMR** with spectrometers of greater than 800 Mhz which, as well as increasing the spectral resolution, have a sensitivity that allows a reduction in the sample concentrations and a shortening of the measurement times. The recent development of cryoprobes has further strengthened the potential of these spectrometers.

One of the most useful applications of **NMR** in pharmacology is the ability to detect **low-affinity protein-ligand interactions**. Indeed, one of the challenges of medicinal chemistry is to design compounds able to interfere with the intracellular protein-protein interaction networks that are involved in all signalling pathways. The major difficulty associated with the **search for protein-protein interaction inhibitors** is the **transient nature of these interactions** and above all the fact that they involve **large molecular surfaces**, typically with not very pronounced grooves. These interaction sites therefore often make poor binding sites for molecules of low molecular weight with medicinal potential. For these reasons, protein-protein interactions have been considered to be poor targets by the pharmaceutical industry. In classic binding assays based on fluorescence techniques, the signal is generally completely saturated at the concentrations necessary for significant binding to take place. It is precisely **in these weak-interaction conditions** that **NMR comes into its own** (example 9.2). Less sensitive than the other spectroscopic techniques, it permits the use of **more concentrated solutions**.

Example 9.2 - NMR measurement of changes in the chemical shift spectrum of proteins caused by the presence of a ligand

Nuclear magnetic resonance (**NMR**) is based on measuring the absorption of radiofrequency radiation by an atomic nucleus in a strong magnetic field. The absorption of the radiation pushes the nuclear spin to realign itself or to turn in the direction of the highest energy. After absorbing the energy, the atomic nuclei reemit radiation of radiofrequency and revert to their initial state at the lowest energy level. The energy of an NMR transition depends on the strength of the magnetic field as well as a proportionality factor applied to each nucleus. The local environment **around a given nucleus in a molecule** (such as a protein) **affects its transition energy**. This dependence on the transition energy with respect to the position of a particular atom in a molecule makes NMR extremely useful for **determining the structures of molecules**. Besides, the **presence of a near ligand can affect the transition energy of neighbouring atoms**.

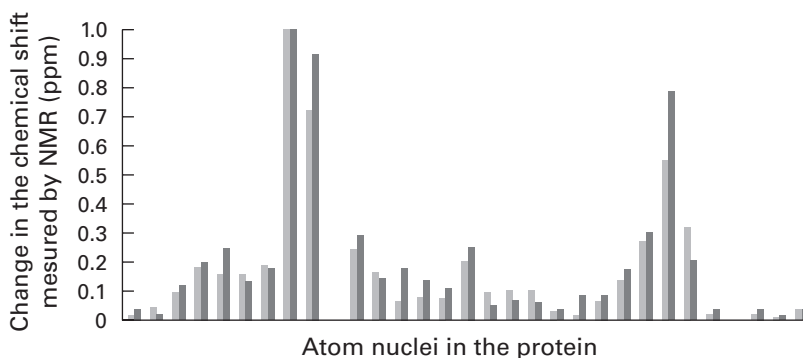


Fig. 9.3 - Result of an NMR experiment for a ligand binding to a protein receptor

Each numerical value corresponds to changes in the chemical shifts of certain nuclei in the protein (^{15}N and ^1H with respect to the α carbons of amino acids) near to the interaction site.

The technique consists of measuring the NMR spectrum of a protein (here labelled uniformly with ^{15}N) in the presence of different concentrations of small molecules. The presence of a ligand modifies the environment of the protein's atomic nuclei, which have altered chemical shifts. By measuring these changes at several ligand concentrations, the ligand's K_d (see chapter 5 for the definition of K_d) can be calculated. This K_d value constitutes the first information supplied by the experiment to chemists (fig. 9.3). If the protein spectrum has been assigned, the zones of interaction between the ligand and the protein can be determined, assuming that the amino acids in contact with the ligand are those having the most altered chemical shifts. This is only valid if ligand binding does not induce a long-range conformational change, which is not always true.

Such a histogram enables the chemist to narrow down the interactions between the ligand and the protein, which guides the design of novel molecules. Contrary to the multidimensional functional assay, it is possible to take advantage of this information for design purposes.

NMR is therefore a useful tool for initiating projects in which only hits with very weak affinity are expected, and for which knowledge of the mode of interaction aids the rational design of more refined analogues. Nevertheless, its low throughput and the large quantity of protein required for the experiments still limit the use of this technique.

9.3. CONSTRAINTS LINKED TO THROUGHPUT AND TO THE LARGE NUMBERS

9.3.1. KNOW-HOW

Handling complex biological objects (such as whole organisms) and sophisticated technologies (such as NMR, imaging etc.) requires advanced scientific know-how. The emergence of approaches with high information content in research and development (R & D) is today characterised by a very strong upstream research aspect (R). In the future, the development phase (D) will benefit from considerable effort in modelling both the complex processes implicated (chapter 6) and the target integrated in a biological context (chapters 1 and 14).

9.3.2. MINIATURISATION, RATE AND ROBUSTNESS OF THE ASSAYS

Conducting thousands of experiments every day is only possible if the unit cost is controlled. One measurement that can be performed for a few hundred experiments can become economically impossible upon increasing the rate. Besides, **an increase in the complexity of the system generally reduces the screening rate**. Due to this, the miniaturisation of tests with complex biological entities or those relying on heavy technologies (NMR) requires particular care in miniaturisation (chapters 3 and 8). The acquisition of multiple parameters is beneficial compared to acquiring single parameters, and consequently represents an economy, which must be taken into account.

In an industrial setting, a screening campaign with 50,000 to 300,000 compounds in general does not require more than one or two months with a screening team of three people plus their equipment. This campaign therefore represents between 50,000 and 150,000 € in personnel and depreciation of the screening equipment, which is comparable to the cost of reagents. The goal in terms of throughput therefore is between 5,000 to 10,000 samples per day. The impact of throughput on the cost of a screening campaign is high (about 50% of the cost).

To reach maximum efficacy, the screening assay must be as robust as possible. This robustness can be measured by the *Z* factor, which characterises, for a binary analysis of the signal, the separation of the populations of negatives and positives. The separation of these two populations governs the certainty of the conclusion about each sample. It is commonly agreed that if the *Z* factor is between 0.3 to 0.5, the experiments must be duplicated in order to reach a decision about a sample's activity. The duplication of experiments naturally doubles the cost of the screening campaign. It is therefore vital, while setting up a test, to spend sufficient time on improving its robustness and to bring the *Z* factor to a value higher than 0.5 (see chapter 3).

9.3.3. NUMBER, CONCENTRATION AND PHYSICOCHEMICAL PROPERTIES OF SMALL MOLECULES

This point is discussed in chapter 8 for high-content screening in the context of phenotypic screening.

With high-content screening in the context of **structural screening**, the **chemical libraries are in general small in size**. Structural screening aims rather to identify good-quality lead compounds, which will then be optimised in a classic way. For analyses using nuclear magnetic resonance (NMR) or X-ray diffraction (XD) a limited sample of compounds or compound mixtures are tested at **high concentration**. The high concentrations necessary for the experiment require the compounds to be very **soluble in water**, which implies *de facto*, when dealing with non-peptidic substances, a **low molecular weight** and **low complexity**. Besides the fact that the number of compounds required limits an exhaustive exploration of structural space, the **accessible diversity at this level of complexity is also reduced**.

9.4. TYPES OF MEASUREMENT FOR HIGH-CONTENT SCREENING

9.4.1. THE CRITICAL INFORMATION NEEDED FOR SCREENING

The question of what is the target of the screen is one of the most difficult points to define (see chapters 1 and 14). In high-content screening, the concept of **critical information** summarises what one wishes to measure, and what is the purpose of the screen. The critical information is therefore a major concept. We are evolving from **molecular target-specific screening** towards **phenotypic pathway-wide screening**. This latter strategy is carried out with complex biological systems, such as cells or whole model organisms (fly, nematode, zebrafish). Screening consists of using an external probe to disrupt a metabolic pathway or a signalling process and to identify a key gene or protein in this pathway. The test, with which it is possible to measure the signals that reflect the process targeted, thus takes into account the **critical (relevant) information permitting the selection of bioactive molecules**.

9.4.2. RAW, NUMERICAL RESULTS

The information gained can be presented in the form of a result from 1 to n numerical values. In **phenotypic screening** assays, this value may be the **result of integrating several biological phenomena** (crossing physical or metabolic barriers, activity towards several targets). In the case of the search for agonist ligands with the help of target-specific tests, the information is made up of at least two EC50/maximum-effect values (chapter 5).

9.4.3. RESULTS ARISING FROM EXPERT ANALYSES

Information is sometimes produced in the form of images, as is the case for phenotypic assays with cells. These assays can be handled by an expert experimenter for a few samples, but in order to be applied to thousands of samples it is necessary to program algorithms for form recognition.

With NMR screening, the primary acquisition is a spectrum (two-dimensional, 2D, or three-dimensional, 3D). A first analysis of the spectra can be automated and some quantitative parameters extracted from the spectra (example 9.2), such as by measurement of the chemical shift changes of chosen amino acids.

9.5. CONCLUSION

All of the screening examples described in this chapter yield a large quantity of information which has to be stored digitally and translated into a format usable by the biologist and chemist. In the last few years great progress has been made in informatics, with the creation of software (commercial or academic) that enables formatting, representation and simple use of the multidimensional data associated with chemical structures or with nucleic acid sequences. This software not only permits an analysis of the screening data, but also a comparison with archived data from other screening tests with the same compounds or genes. It is certain that a posterior analysis of historic data from screening laboratories constitutes the next generation of 'high-content screening' and which probably holds in store for us many interesting discoveries.

9.6. REFERENCES

- BIRD D.M., OPPERMAN C.H., JONES S.J., BAILLIE D.L. (1999) The *Caenorhabditis elegans* genome: a guide in the post genomics age. *Annu. Rev. Phytopathol.* **37**: 247-265
- HAJDUK P.J., GERFIN T., BOEHLEN J.M., HABERLI M., MAREK D., FESIK S.W. (1999) High-throughput nuclear magnetic resonance-based screening. *J. Med. Chem.* **42**: 2315-2317
- WIDMER H., JAHNKE, W. (2004) Protein NMR in biomedical research. *Cell Mol. Life Sci.* **61**: 580-599
- SMET C., DUCKERT J.-F., WIERUSZESKI J.-M., LANDRIEU I., BUEE L., LIPPENS G., DEPREZ B. (2005) Control of Protein-Protein Interactions: Structure-Based Discovery of Low Molecular Weight Inhibitors of the Interactions between Pin1 WW Domain and Phosphopeptides. *J. Med. Chem.* **48**(15): 4815-4823
- KALETTA K., HENGARTNER M.O. (2006) Finding function in novel targets: *C. elegans* as a model organism. *Nat. Rev. Drug Discov.* **5**: 347-398

SOME PRINCIPLES OF DIVERSITY-ORIENTED SYNTHESIS

Yung-Sing WONG

10.1. INTRODUCTION

Chemical genetics has recently established itself as a systematic and powerful tool for exploring the biological world by exploiting the structural diversity of small molecules, notably with the aim of probing proteins (see chapter 2). Inhibiting the activity of a protein with a small molecule (a ligand) corresponds to eliminating the gene with which it is associated (knock-out effect) and this allows the phenotype to be altered. This approach is in a way similar to classical genetics but with a far simpler technical setup.

In this context, obtaining novel small molecules in the form of a chemical library plays a central role. The boom in **combinatorial chemistry** at the end of the 20th century was the result of this high demand for small molecules. Nevertheless, with the exception of the French National Chemical Library (*Chimiothèque Nationale*; see chapter 2), the chemical libraries from commercial syntheses in general supply screening platforms with a low proportion of original structures. This is due notably to the design of current libraries, focussed principally on structures easily accessed by combinatorial chemistry and the diversity of which is provided only by the variation in the added residues. This approach massively produces molecules that harbour an identical central skeleton (scaffold) often with little structural complexity and thus result in little novel information.

In the quest for simplification we ought not to forget the essentials, that is, the search for quality information. This is why **Diversity-Oriented Synthesis, DOS** (fig. 10.1; SCHREIBER, 2000; BURKE *et al.*, 2004; SPRING, 2005) claims to go further and favours obtaining novel small molecules that combine quality (**diversity** and **complexity**) and accessibility (**simplicity** and **feasibility**) in order to fill in efficiently and rapidly the vast ‘chemical space’ left as yet unoccupied by the chemical libraries of today.

This introductory chapter aims to illustrate with a few examples the recent strategies in DOS. Why create complexity and diversity? How can they be defined?

How can the chemist quickly create them? Should products be synthesised pure or as a mixture? This chapter will attempt to provide some answers to these questions.

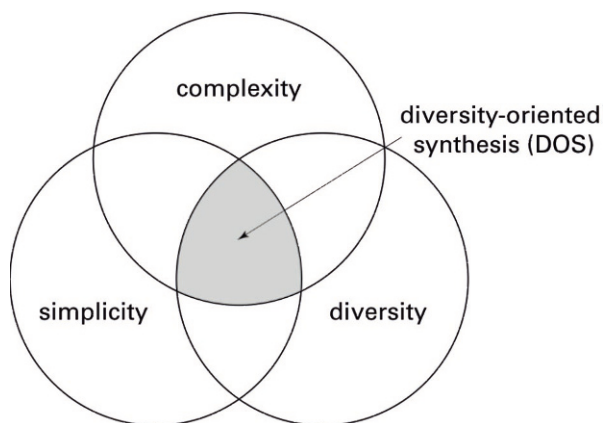


Fig. 10.1 Three principles on which the DOS approach relies

This introductory chapter aims to illustrate with a few examples the recent strategies in DOS. Why create complexity and diversity? How can they be defined? How can the chemist quickly create them? Should products be synthesised pure or as a mixture? This chapter will attempt to provide some answers to these questions.

On the other hand, we do not deal here with the technologies for simplification destined to facilitate and to accelerate the synthesis steps by automation, such as solid-phase synthesis for example (permitting notably ‘split/mix’ or ‘one bead/one compound’ strategies). To clarify, the reactions described will be considered to be carried out in solution inside reactors.

10.2. PORTRAIT OF THE SMALL MOLECULE IN DOS

If we attempt to establish a comparative profile between the protein target and the small molecules, we observe (fig. 10.2):

- » On the one hand, a protein, which is constituted of the linear combination of amino acids, comparable to **building blocks**, each related by the same type of bond (peptide bond). Proteins are **chiral** but exist as a single **enantiomer**. Chirality is present in each amino acid, which possesses an **asymmetric** centre (symbolised by *), essentially with the L configuration. The **degree of diversity** (see § 10.3) rests therefore on the length of the peptide chain and on the combination of about 20 chiral amino acids with various characteristics: hydrophobic, hydrophilic, protic, ionic, aromatic etc. This chain of amino acids is like a ‘pearl necklace’ and when folded is capable of giving rise to a **complex** and **chiral three-dimensional entity**, endowed with biological functions (e.g. recognition, specific catalyses).

- » On the other hand, the design using DOS of a small-molecule library, which must take into account its capacity to yield the richest information possible while considering production constraints, *i.e.* **simple** and quick access (achievable in 2, 3 or 4 steps for example) with great molecular **diversity** and producing maximum structural **complexity**. Considering that proteins cover a complex three-dimensional and asymmetric 'biological space', the probes used for exploring this space should, if possible, possess this same property of three-dimensional complexity in order to be potentially more refined. On the scale of the library, this must be as diverse as possible in order to explore the entire space. Even though the production of bioactive-peptide libraries remains relevant (FALCIANI *et al.*, 2005), the difficulty in obtaining them by synthesis increases proportionally with the desired length of peptide. DOS of small molecules seeks to express this diversity and complexity in an accessible way while still benefitting from the possibility of more broad variation.

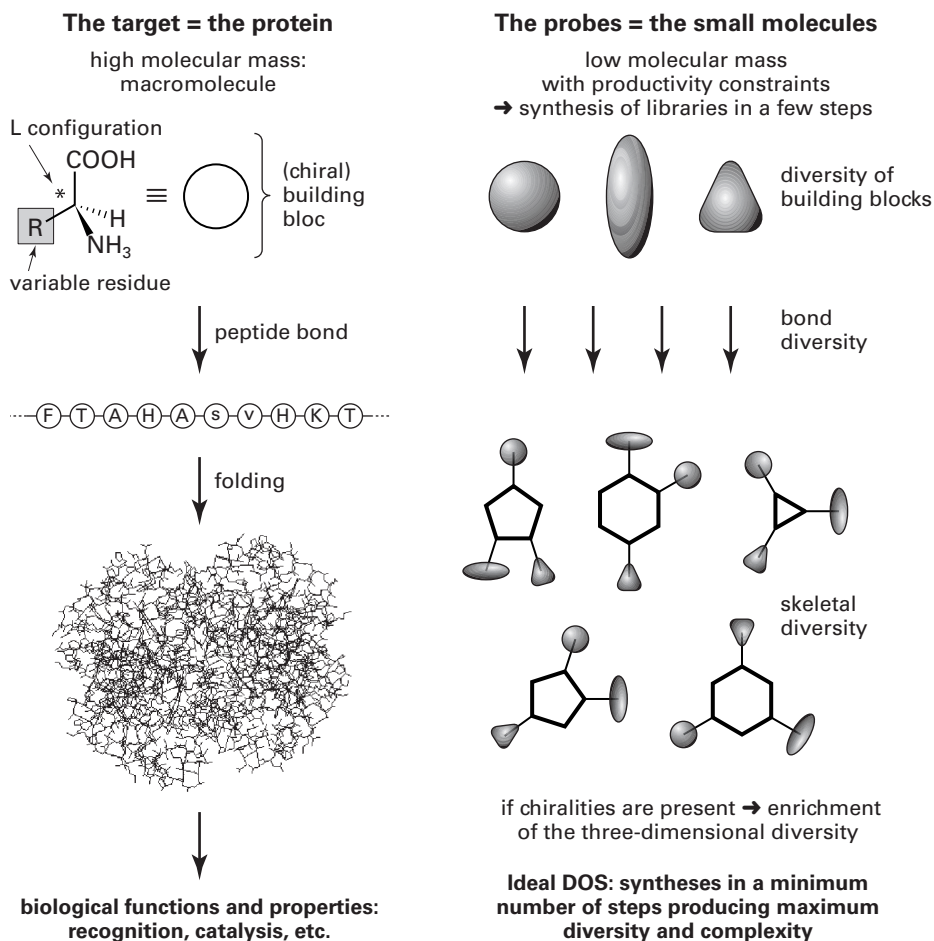


Fig. 10.2 - Comparative profile of the protein target and the probes

Thus, the chemist can experiment with (fig. 10.2):

- › the very great diversity of building-blocks,
- › the richness of chemical reactions available for creating diverse types of covalent bond,
- › the possibility of producing several types of skeleton,
- › the creation of chirality that easily generates complexity and diversity.

By combining judiciously these **elements of diversity**, it is possible to produce rapidly a complex and varied library of small molecules that is rich in exploratory potential.

10.3. DEFINITION OF THE DEGREE OF DIVERSITY (DD)

Whereas the degree of structural simplicity or complexity invoke more qualitative notions, the **degree of diversity** (or of ‘dissimilarity’) which we abbreviate here to **DD**, is more ‘quantifiable’. The elements of diversity can be determined on several levels during synthesis, specifically, **DD** of the **building blocks**, **DD** of **stereochemistry**, **DD** of **regiochemistry** and **DD** of the **skeleton**.

To give an outline, let us take as an example cycloaddition reactions (examples 10.1 and 10.2). These represent well the phenomenon of diversity and complexity generation, as they create in a single reaction step at least two covalent bonds with the appearance of a cyclic structure and new chiral centres.

10.3.1. DEGREE OF DIVERSITY OF THE BUILDING BLOCK

The **degree of diversity of the building block** (DD_{block}) is equal to the number of different and variable reactants involved in the same reaction. To yield quickly a wide diversity, it is preferable for this value to be **equal or greater than three** ($DD_{\text{block}} \geq 3$). These reactions, termed **multi-component reactions**, **MCR**, are often used in combinatorial chemistry (ZHU *et al.*, 2004). For example, a reaction with three components is a 3CR, with four components, a 4CR, and so on. This is a **convergent approach** where several reactants condense together to form a single type of product at the end. Leaving aside the **stereochemistry** and **regiochemistry** of the products formed (these aspects are dealt with in § 10.3.2 and 10.3.3), diversity will thus come about only through the potential of the combinations between these variable reactants. The number of possible products arising from a reaction and having the same central skeleton with distinct residues R^x is $N_A \times N_B \times N_C \times \dots \times N_Z$ where each letter A, B, C, ... Z, symbolises a different initial reactant and N the number of different species by reactant type.

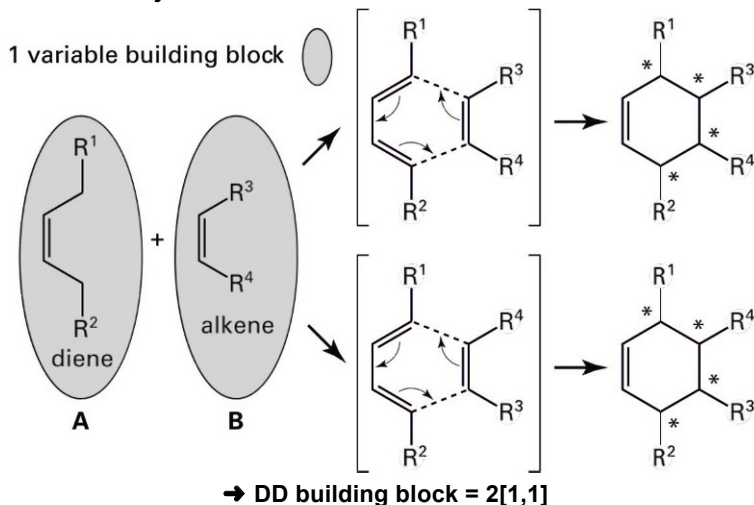
To simplify this, let us consider in the next examples that each reactant possesses the same number of different species ($N_A = N_B = N_C \dots = N_Z$), which brings the number of potential products with an identical central skeleton to $N^{DD_{\text{block}}}$.

In **case 1** from example 10.1, $DD_{\text{block}} = 2$ and the values in square brackets [1,1] indicate that only two distinct, variable building blocks exist and that each takes part only once in the reaction (the representation (AB) is also used).

Example 10.1 - evaluation of the degrees of diversity (DD)

Case of the cycloaddition reaction

Case 1 - DIELS-ALDER cycloaddition with 1 diene $R^1 R^2$ + 1 alkene $R^3 R^4$



In square brackets, each number symbolises a reactant:

- ▶ 1, the variable reactant only participates once in the reaction (case 1)
- ▶ 2, the variable reactant participates twice (see case 2) ...
- ▶ 0, for a reactant that does not vary (see case 3).

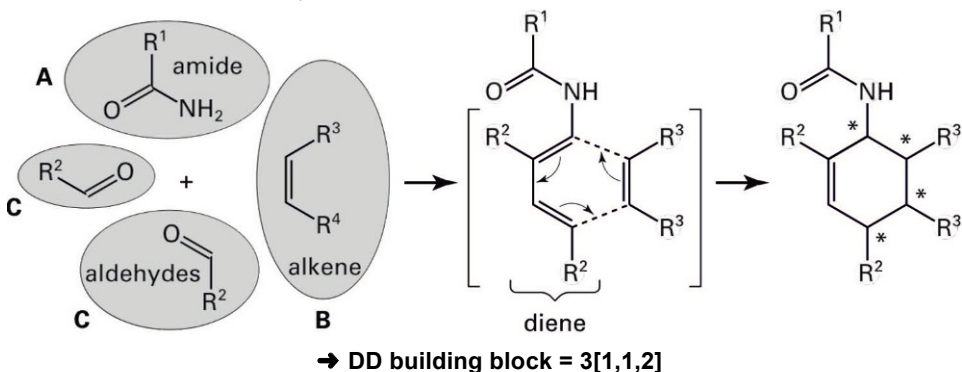
If $R^1 \neq R^2 \neq R^3 \neq R^4 \neq \text{hydrogen}$, 4 asymmetric centres are created (symbolised by * which is a binary-type variable, i.e. *R* or *S* configuration).

→ DD stereochemistry = 4

→ DD regioisomer = 2

Case 2 - DIELS-ALDER cycloaddition

with 1 amide R^1 + 2 aldehydes R^2 + 1 alkene $R^3 R^4$ (MCR = 4CR)

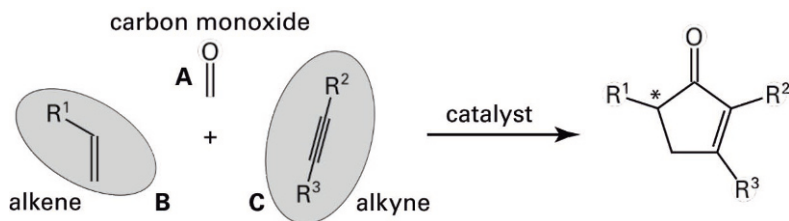


An intermediate diene is generated *in situ* by the condensation of 1 amide + 2 identical aldehydes.

→ DD stereochemistry = 4

→ DD regioisomer = 1

**Case 3 - PAUSEN-KHAND cycloaddition
with 1 carbon monoxide + 1 alkene R^1 + 1 alkyne $R^2 R^3$ (MCR = 3CR)**



→ DD building block = 2[0,0,1]

Although the reaction has three reactants, only two of them (the alkene and the alkyne) can vary:

→ DD stereochemistry = 1

In an MCR, it is important to consider the numbers that vary and not the number of fragments taking part in the reaction. In **case 2**, although the reaction involves four reactants (MCR of type 4CR), only the contribution of three variable reactants can create diversity (amide R^1 , aldehyde R^2 and alkene R^3R^3). This is why here $DD_{\text{block}} = 3$, and in square brackets [1,1,2], where the value 2 indicates that a reactant (the aldehyde R^2 in our example) takes part twice in the reaction (the representation (AB2C) is also used). If for example $N = 5$, there are 125 possible products (5^3); if $N = 10$, 1,000 different products are accessible (10^3). We can see that the potential of possible cycloadducts attainable in **case 2**, while starting for example with only 30 reactants ($dd_{\text{block}} = 3$, $N = 10$) can reach 1,000 products, whereas in **case 1** with 30 reactants ($dd_{\text{block}} = 2$, $N = 15$), only 225 products are accessible. Lastly, **case 3** is a three-component MCR (3CR), but only two reactants can vary. This is why $DD_{\text{block}} = 2$, with [0,1,1]. It is notable that in this case the representation (ABC) which we come across in the literature is not suitable for expressing the true degree of diversity of this reaction (it would be more correct to write (0ABC)).

Thus, with the same number of initial reactants, it is more advantageous to carry out a reaction with the highest DD_{block} possible.

10.3.2. DEGREE OF STEREOCHEMICAL DIVERSITY

The **degree of stereochemical diversity** (DD_{stereo}) corresponds to the number of asymmetric centres created during the course of the synthesis step. The stereochemical diversity increases the number of potential orientations of chemical residues based on the same skeleton in three-dimensional space. In example 10.1 **case 1**, if the **regioisomers** are excluded (this aspect will be dealt with in □10.3.3) and the residues R^1 , R^2 , R^3 and R^4 are all different from each other, during the cycloaddition reaction a number of **stereoisomers** equal to $2^{DD_{\text{stereo}}}$ may potentially be formed. This constant 2 can be explained by the fact that a chiral centre either has the *R* configuration or the *S* configuration (variable of binary type symbolised by *). In example 10.1 for **case 1** and **case 2**, DD_{stereo} is equal to 4. Consequently, 16 different stereoisomers with the same structural formula can theoretically be synthesised.

In reality, it is not always possible to access all stereoisomers. During the creation of asymmetric centres, interactions and tendencies exist that prevent or hinder the formation of certain stereoisomers. It is therefore difficult to estimate beforehand the number of stereoisomers and their respective proportions. If some stereoisomers predominate in quantity with respect to others at the start of the reaction, the latter will be called **stereoselective** (**stereospecific** for the formation of a single stereoisomer). It is necessary to distinguish **diastereoselective** reactions, which discriminate between **diastereoisomers**, and **enantioselective** reactions, which discriminate between **enantiomers** (see example 10.2). Although this stereoselectivity acts against diversification in the synthesis of mixtures, it is important for the parallel synthesis of libraries of pure products. It is noteworthy that a mixture of diastereoisomers is often not very homogenous in terms of the quantitative ratio between the species.

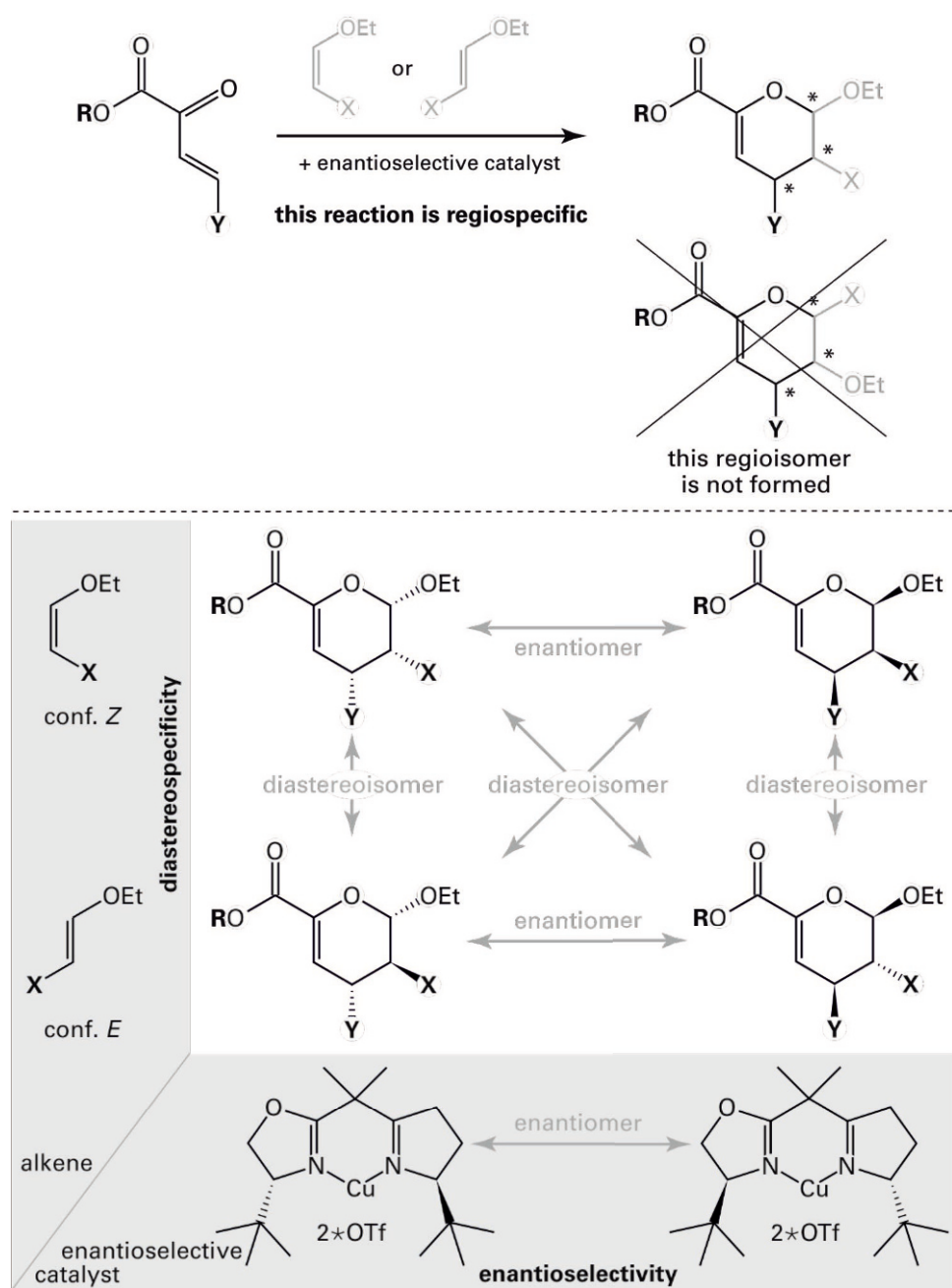
Example 10.2 illustrates the possibility of combining enantioselectivity and diastereospecificity in a reaction to produce selectively three-dimensional diversity and thus to obtain a matrix of pure products (STAVENGER *et al.*, 2001). On the horizontal axis, two chiral catalysers are enantiomers with respect to each other. They are capable of inducing a highly enantioselective cycloaddition, giving rise respectively to products that are themselves enantiomers relative to one another. On the vertical axis, two initial products (alkenes) are distinguished by the arrangement of their **-OEt** and **-X** groups: the *Z* configuration for residues **-OEt** and **-X** are positioned on the same side and the *E* configuration where **-OEt** and **-X** are in opposition. The reaction is diastereospecific and from *Z* alkenes cycloadducts are formed with the **-OEt** and **-X** groups pointing into the same region of space, and from *E* alkenes, the **-OEt** and **-X** groups are directed in opposite directions. Thus, by combining the two stereoselective factors, it is possible selectively to make a matrix of four distinct products presenting the same structural formula, but having a different spatial layout of their residues.

10.3.3. DEGREE OF REGIOCHEMICAL DIVERSITY

The **degree of regiochemical diversity** (DD_{regio}) is the number of **regioisomers** potentially accessible. The regioisomers possess the same scaffold, but their residues are localised on the scaffold at different positions. In example 10.1 **case 1**, during the cycloaddition, the alkene can present its R^3 group either from the side of the R^1 group, or from the side of R^2 of the diene.

It is thus possible to form two different products. In **case 1**, in the absence of elements that control this selectivity (**regioselective** element), DD_{regio} will be equal to 2. In example 10.2, the dienes and alkenes present complementary polarities, which induce **regiospecificity** during the cycloaddition by only forming a single regioisomer ($DD_{\text{regio}} = 1$).

Example 10.2 - combinations of the enantioselective and diastereospecific properties of the reaction to generate a matrix of four distinct stereoisomers



The double arrows indicate that the products are related to each other by being either enantiomers or diastereoisomers.

10.3.4. DEGREE OF SKELETAL DIVERSITY

The **degree of skeletal diversity** (DD_{skel}) is the number of new, different skeletons to which the initial skeleton is added. Each new skeleton must arise in a single step from an initial common skeleton. A new skeleton is defined either by the conversion of an existing ring structure (addition or removal of links in the ring), by the gain of a ring (e.g. from monocyclic to bicyclic), or by ring disappearance (e.g. from a monocyclic to a linear form). Two approaches can be envisaged and are schematised in figure 10.3.

For Approaches 1 and 2 in figure 10.3, four new skeletons arise from a common skeleton. Their DD_{skel} is equal to 5 as the library will in the end consist of 5 types of different skeletons (including the initial skeleton).

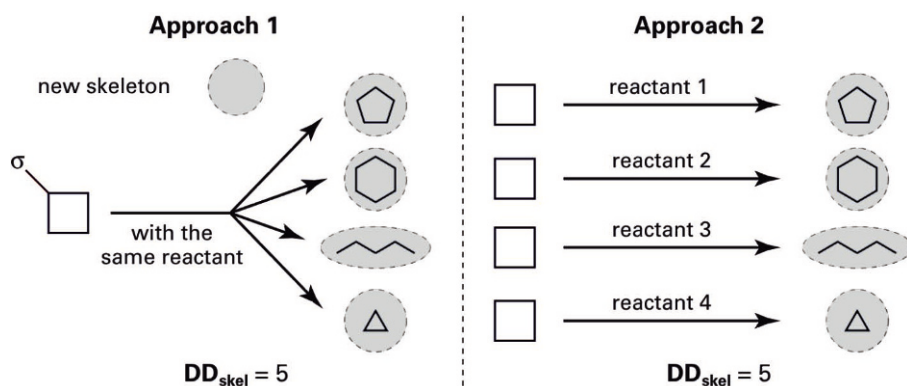


Fig. 10.3 = Two general approaches for expanding the diversity towards new skeletons from one common initial skeleton

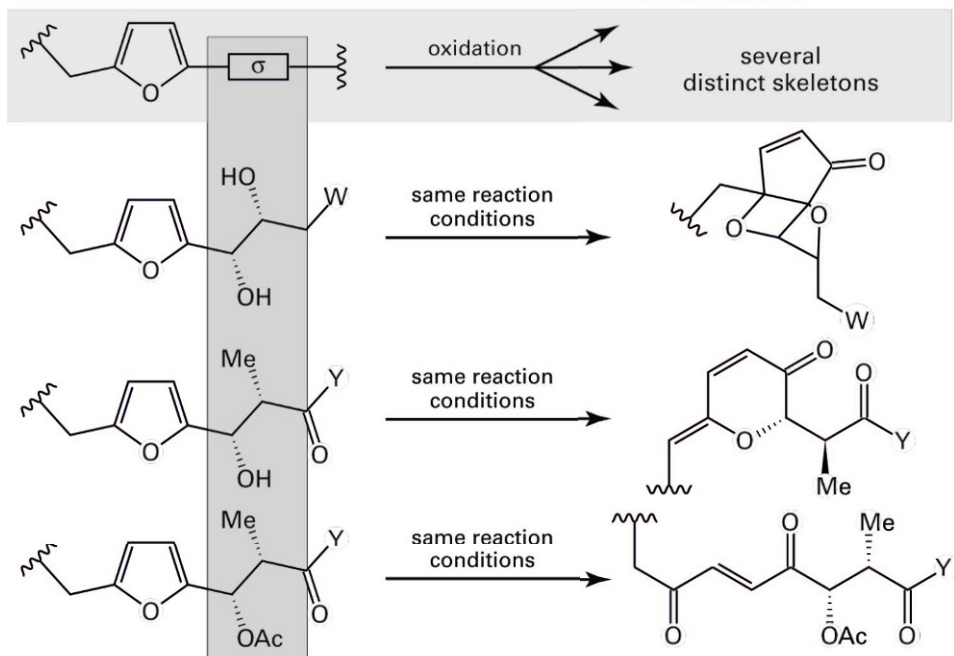
Approach 1 relies on the substrate's properties: a single reactant allows the generation of different skeletons depending on the nature of the **element σ** attached to the initial skeleton; **Approach 2 relies on the reactant's properties:** the initial compound is placed in different reactors and a different reactant is applied to each one, selectively giving a new skeleton.

Example 10.3 gives an illustration of Approach 1. The variable element σ (depicted by a grey background) of the molecule can, depending on its nature, lead to the formation of different structures under the same reaction conditions. The element σ may act in one dimension (a single element σ on the molecule) or by combining two elements σ_1 and σ_2 at the same time (BURKE *et al.*, 2003, BURKE *et al.*, 2004).

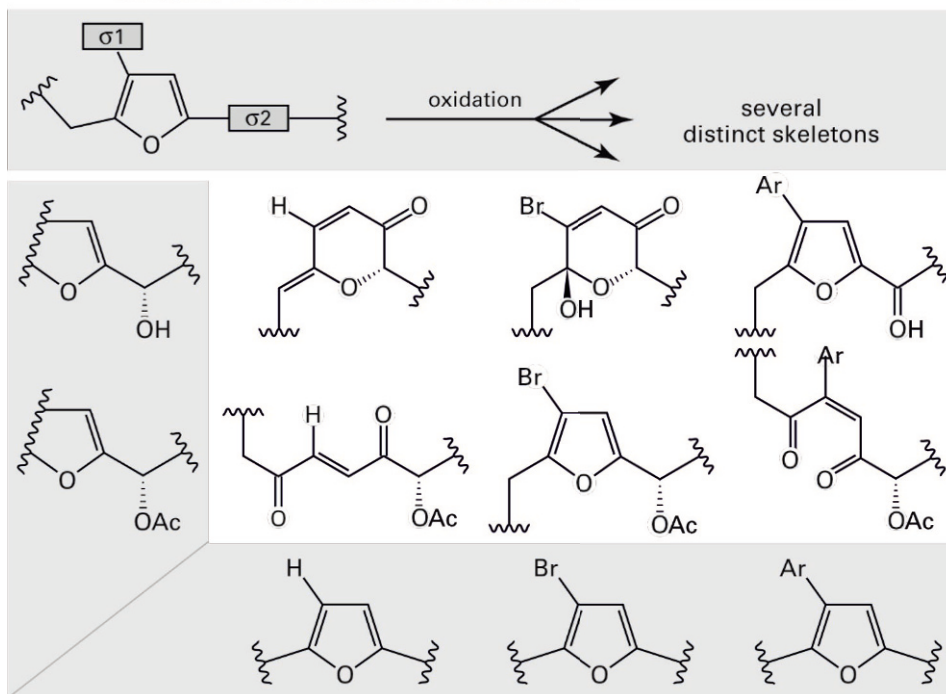
In the present example, the effect of varying the element σ is linked principally, on the one hand, to an alcohol functionality, which is either protected or not, and on the other hand to the substitution of the furan functionality by different stabilising groups. So that DOS remains a simple approach it is important that the products bearing the σ elements are derived in only a few steps from one common synthetic intermediate.

Example 10.3 - generation of skeletal diversity from hydroxy-furan functionality by single or double induction from the element σ

Influence of the element σ on a one-dimensional matrix

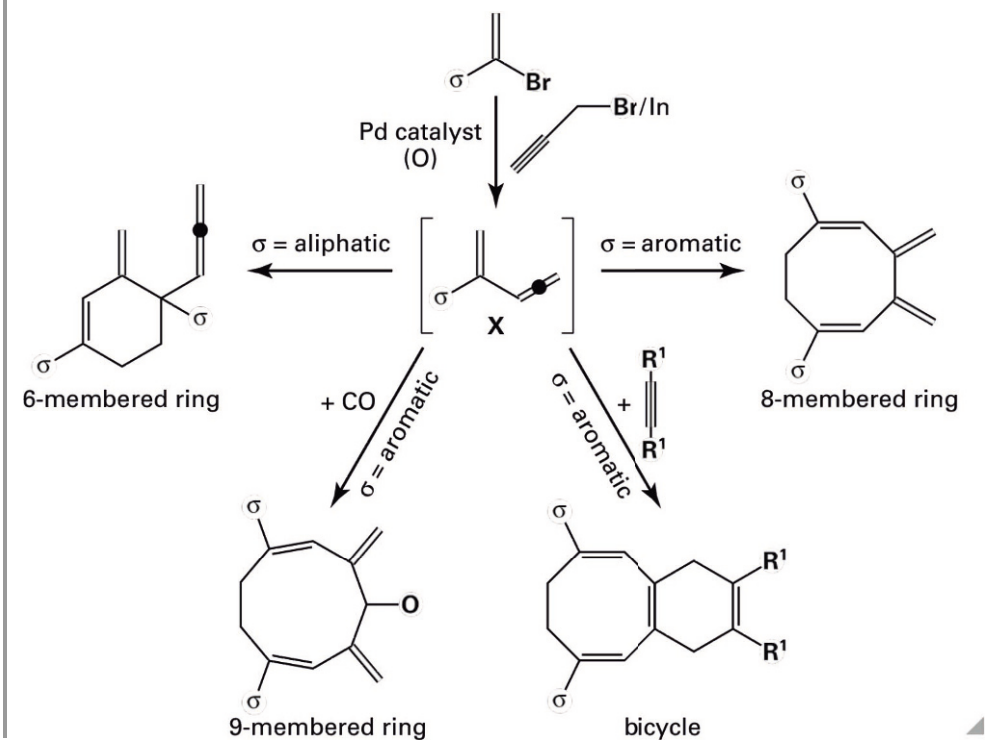


Influence of the elements σ_1 et σ_2 on a two-dimensional matrix



It is also possible to combine Approaches 1 and 2 simultaneously. Example 10.4 represents a dimerisation reaction of one non-isolatable reaction intermediate **X** (LEE *et al.*, 2006). Depending on the aliphatic (*i.e.* having saturated carbon chains) or aromatic nature of the element σ , a 6- or 8-membered ring will be formed, respectively. If carbon monoxide (CO) is present in the mix, a 9-membered ring will be produced. In the presence of a triple bond, a DIELS-ALDER reaction will lead to the formation of a fused bicycle.

Example 10.4 - dimerisation reaction leading in one step, depending on the element σ and the reaction conditions, to the selective formation of 6-, 8- and 9-membered rings



In conclusion to this section dedicated to the degrees of diversity, we have seen how rational exploitation of each element of diversity can lead to the rapid creation of a multitude of new products, while generating structural complexity. The simple fact of shrewdly combining elements of diversity together during the same synthetic step amplifies the potential for diversification.

An estimate of the number of potential products, n , accessible *in one step* is defined by the formula: $n = (N_A \times N_B \times N_C \times \dots \times N_Z) \times 2^{\text{DD}_{\text{stereo}}} \times \text{DD}_{\text{regio}} \times \text{DD}_{\text{skel}}$, and where $n = N^{\text{DD}_{\text{block}}} \times 2^{\text{DD}_{\text{stereo}}} \times \text{DD}_{\text{regio}} \times \text{DD}_{\text{skel}}$, if $(N_A = N_B = N_C = \dots = N_Z)$. Leaving aside the problems associated with product purification, it is more useful to have a value of **DD** that is as high as possible in order to make available more novel and diverse molecules. Although more constraining than with mixture-

synthesis strategies, it remains advantageous to design strategies that favour obtaining libraries of pure products (chimio-, stereo- and regioselective syntheses). It is firstly necessary to consider the qualitative aspect. Indeed, in a mixture, the presence of all of the potential products cannot be guaranteed and their proportion is not homogenous, which can lead to false negatives in biological tests (an analysis by HPLC coupled to mass spectrometry of each mixture is therefore required to establish the presence and the proportion of each species). It is then necessary to take into account the possible reuse of pure-product libraries. They can continue to supply the creation of other pure-product libraries by offering a wider range of possible new combinations (see § 10.4 and 10.5).

10.4. DIVERGENT MULTI-STEP DOS

BY COMBINING ELEMENTS OF DIVERSITY

When planning a library synthesis based on *more than one step*, one of the simple strategies in DOS is to adopt a **divergent approach**. The effect of this is to increase quickly the number of new molecules per synthesis step. A divergent approach implies that each compound synthesised is capable of passing on to the next generation *at least* two types of novel product, and so on (fig. 10.4).

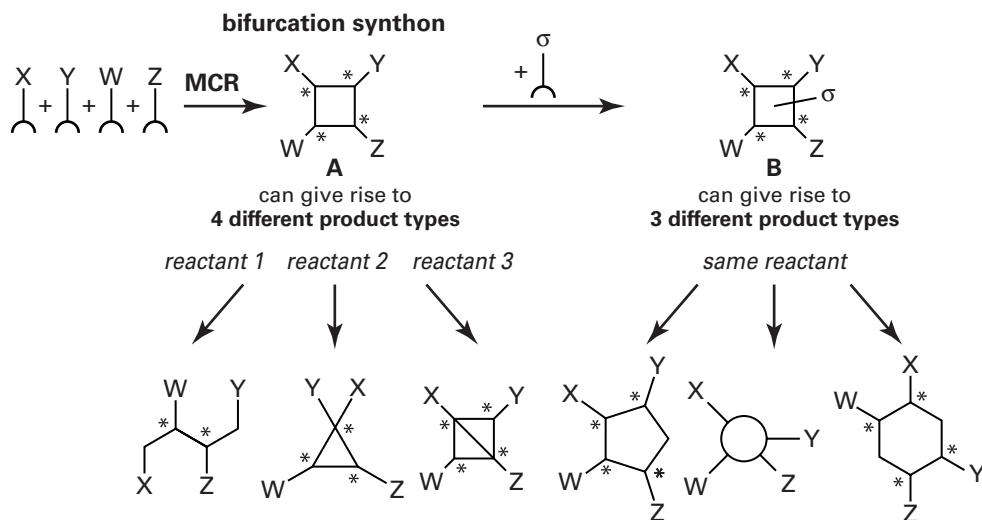


Fig. 10.4 - Using a divergent approach, the combination of different diversity elements provides access in a few steps to a succession of very diverse libraries containing varied, complex structures. In this schematic example, a multi-component reaction (MCR) permits the assembly of four building blocks to give a compound library **A** (stereoselectively or not, depending on whether we want to work with products that are pure or in a mixture). By using different reaction conditions (with 1, 2 or 3 etc. reactants), it is possible to access new diverse skeletons. An element σ can also be attached to the molecules in **A** to access a new library of compounds **B**. The latter can then be transformed into yet another new library possessing other original scaffolds.

We can compare this approach with the structure of a tree, composed of a trunk dividing successively into more and more branches and culminating at the tips in thousands of leaves.

Thus, another important point about this divergent multi-step approach is its intrinsic branching structure. It is important to plan ahead the design of efficient and flexible **bifurcated synthons**. These can be viewed as ‘mother’ molecules localised to different strategic nodes of a network and on which depends, as a function of its ‘fertility’, the existence of numerous ‘daughter’ molecules.

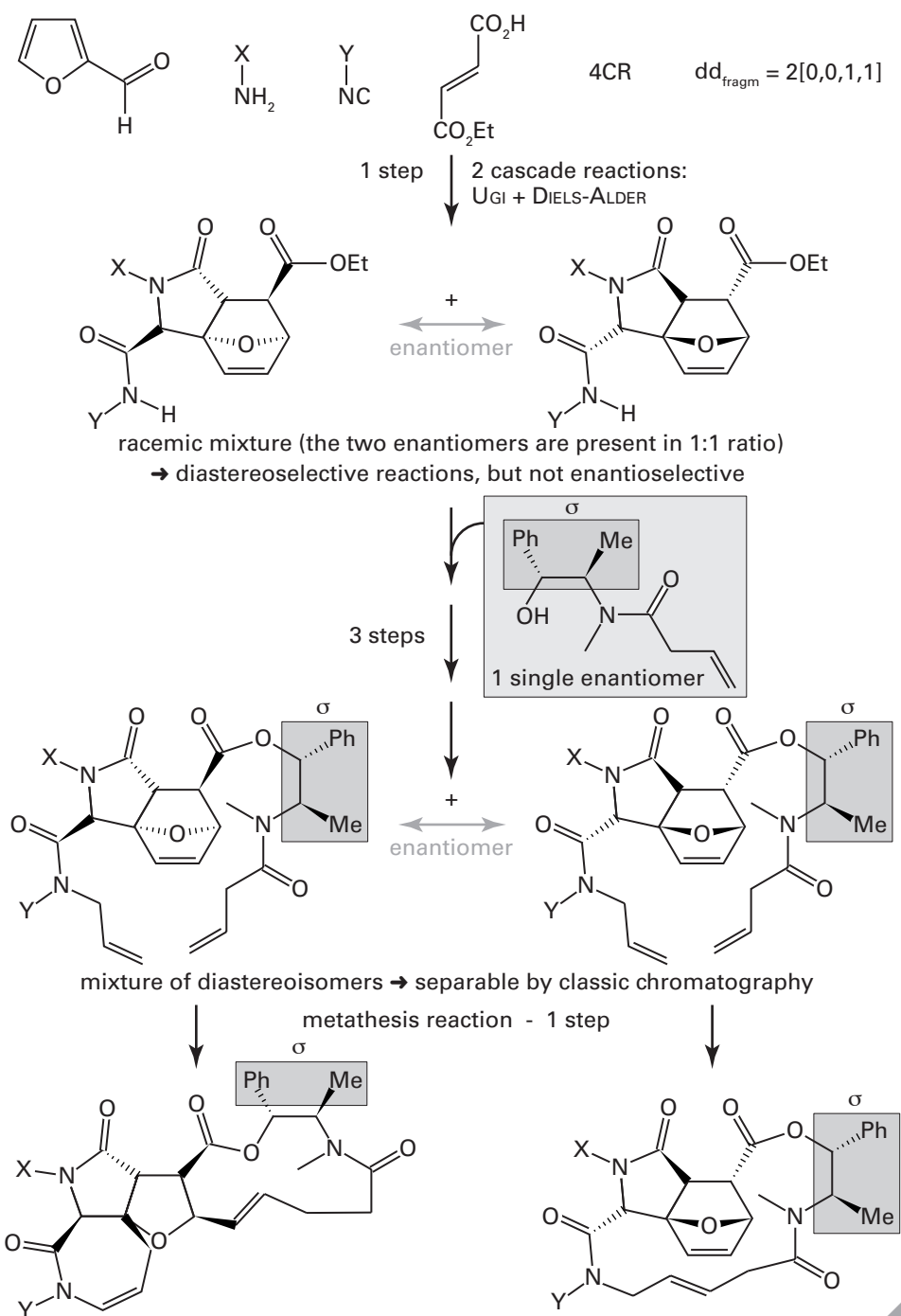
This branching approach also confers on DOS an enhanced ‘reactivity’ with respect to the biological information sought. Not only this, DOS is designed to achieve a greater incidence of affinity with biological targets, but this branching approach allows it to ‘readapt’ quickly and easily its syntheses in view of the outcome of biological tests. All of the molecules are connected to one another by bifurcated synthons available for reorienting the synthesis of new molecules towards novel branches of the network, according to already established protocols.

It is also possible to combine different diversity elements together in several steps in order to create synergistic effects that simply and quickly amplify the potential for diversification.

Thus in example 10.5 (SELLO *et al.*, 2003), a multi-component reaction and a DIELS-ALDER cascade reaction allows in one step, and in a diastereoselective way, a racemic mix of polycycles to be obtained (mixture of two enantiomers in a 1 : 1 ratio). It is of note that at present there are still very few examples of general and efficacious enantioselective multi-component reactions (RAMÓN *et al.*, 2005). To avoid having to work with a mixture of enantiomers, a chiral fragment in the form of a single enantiomer (example 10.5, on a light grey background) was attached to these two enantiomers. At the end of the reaction we therefore obtain this time a mixture of two diastereoisomers, separable by simple classic chromatography. Thus, two libraries of pure complex products having a diastereoisomeric relationship have been synthesised in only a few steps.

The metathesis reaction enables double bonds to combine together generating in one step new and more stable skeletons. Thus, for a diastereoisomer introduced with a metathesis catalyst, a quadracyclic structure is selectively obtained whereas the other diastereoisomer forms an entirely different polycyclic structure under identical reaction conditions. We notice therefore that the same chiral group (example 10.5, on a dark grey background), responsible initially for the derivatisation of the racemic mixture, is also in a position to control efficiently the skeletal diversity with the element σ .

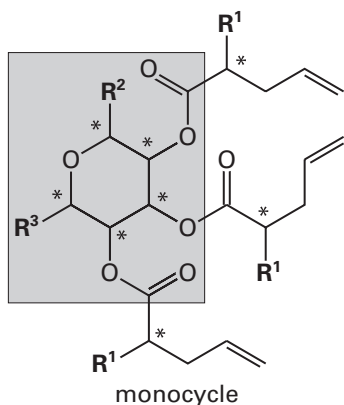
Example 10.5 - derivation of a mixture of enantiomers and control of skeletal diversity by diastereochemical effects



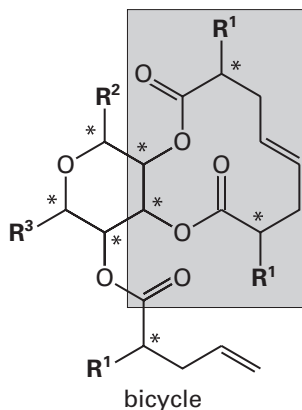
On a different note (example 10.6), the power to combine together double bonds by metathesis was exploited in an initial study seeking to establish a relationship between the stereochemical and skeletal diversity of small molecules and the biological space of the cell (KIM *et al.*, 2005). On different sugars three identical aliphatic residues possessing chirality and a terminal double bond were attached to three neighbouring alcohol functional groups, forming a library of 122 monocyclic products. These then underwent a metathesis reaction yielding the corresponding 122 bicyclic compounds. The 244 molecules were subjected to biological screening with 40 different cell lines. The biological results show that a relationship exists between the biological space of the cell and the topology of the small molecules induced by the stereochemistry and the types of scaffold formed. When the product has a monocyclic form, the stereochemistry of the 6-membered sugar ring dictates the biological activity. On the other hand, regarding the results of the bicycles the stereochemistry of their macrocycle predominates in structure-activity correlations.

Example 10.6 - relationship between the biological space of the cell and the stereochemical and skeletal diversity of small molecules

The stereochemistry of the **sugar** is a dominant factor of characterised 'hits'



The stereochemistry of the **macrocycle** is a dominant factor of characterised 'hits'



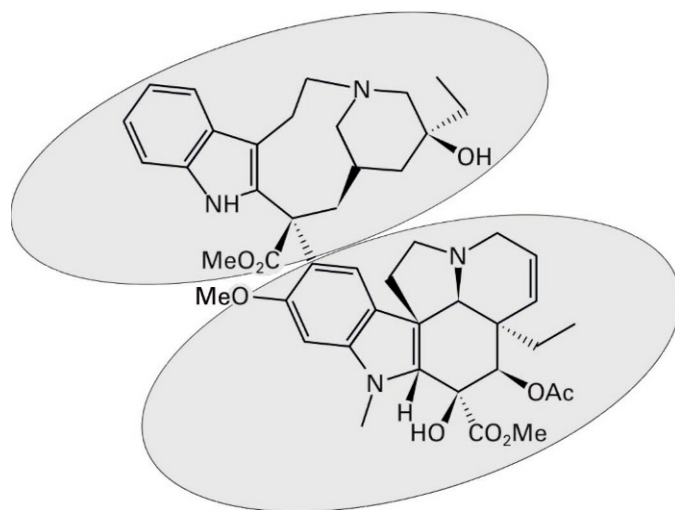
metathesis
reaction
→

122 monocycles and 122 bicycles were tested on 40 different cell lines. ▲

10.5. CONVERGENT DOS:

CONDENSATION BETWEEN DISTINCT SMALL MOLECULES

Having synthesised by divergent DOS a multitude of small molecules displaying a substantial degree of complexity and diversity, it is still possible to enrich the collection with novel molecules, this time in a **convergent** manner (BEELER *et al.*, 2005). Two different types of small molecule can condense together to form a novel hybrid entity (MEHTA *et al.*, 2002, TIETZE *et al.*, 2003). We find in Nature biologically active compounds that are created using this concept (example 10.7).

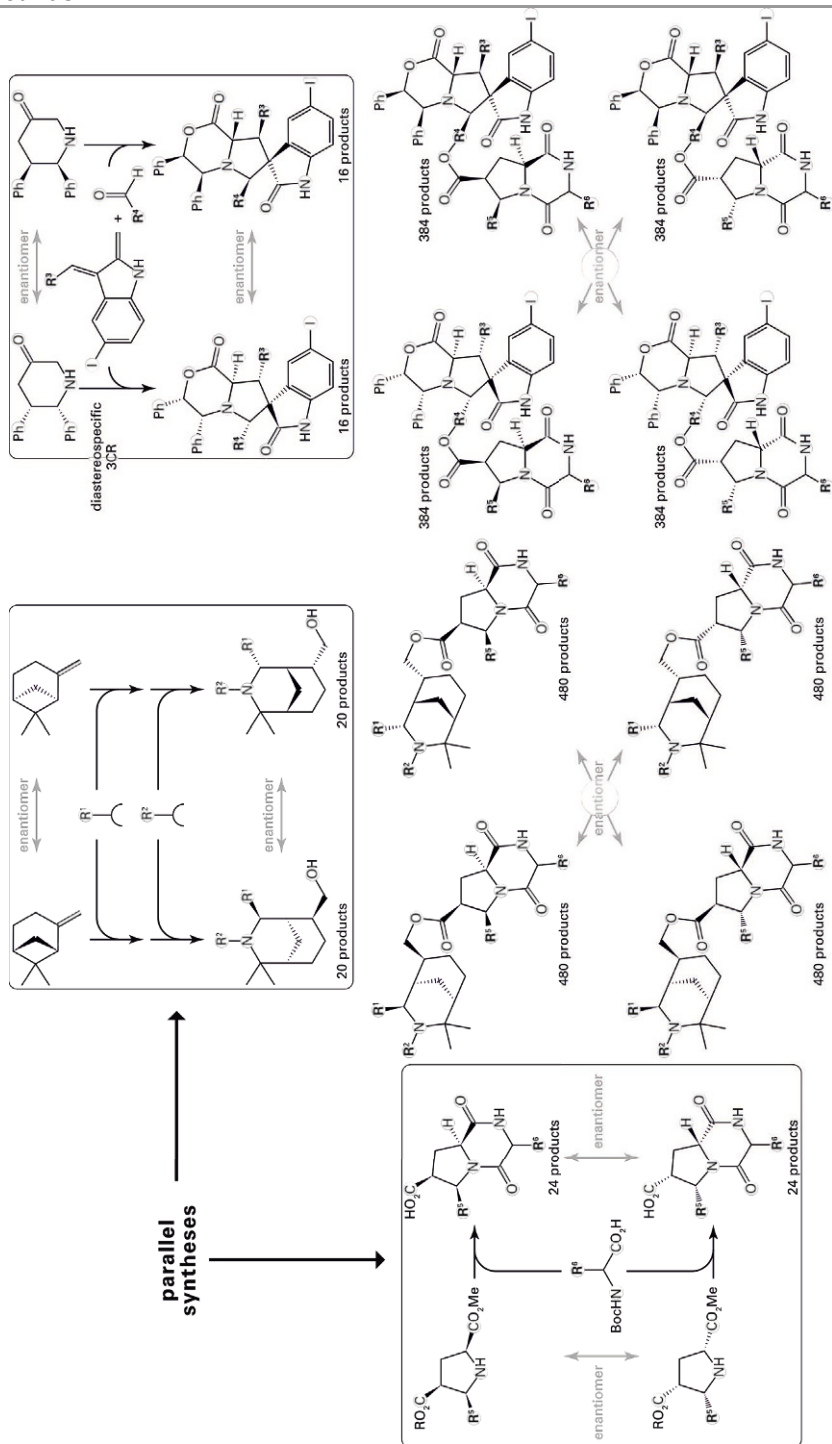
Example 10.7 - vinblastine, a hybrid and biologically active natural molecule

Vinblastine, a natural anticancer product, is made up of two distinct parts linked by a covalent bond. This stratagem has permitted Nature to create easily novel, complex (or hybrid) molecules from different pre-existing molecules. ▲

Example 10.8 illustrates well the exploitation of this strategy to create easily novel, complex molecules (CHEN *et al.*, 2005). Three types of chiral skeleton are synthesised in parallel from easily accessible initial products and in an enantiomerically pure way (*i.e.* not as a racemic mixture). These initial products come either from natural products or from highly enantioselective reactions (not represented in example 10.8). Thus, on the horizontal axis we have for the first type of skeleton two libraries: one containing 20 different molecules varying in their R^1 and R^2 groups, and the other containing the other 20 molecules, which are mirror images of the former. In the same way, for the second type of skeleton two libraries of 16 products were obtained each varying in their R^3 and R^4 residues and for the third type of skeleton, two libraries of 24 products each with varying R^5 and R^6 groups.

We have therefore four libraries containing two types of skeleton on the horizontal axis (72 products), which are compatible with the formation of a hybrid with two libraries containing one type of skeleton, on the vertical axis (48 products).

The ability to assemble them through a single covalent bond has generated a matrix of 3456 (72×48) different and novel products, obtained pure, with the creation of two types of novel skeleton. It is notable that this condensation between different skeletons has permitted the generation of structures that increase the possibilities for the spatial orientation and combination of R^x groups.

Example 10.8 - convergent DOS in the development of a matrix of 3456 hybrid compounds

10.6. CONCLUSION

If we must compare so-called ‘classical’ combinatorial chemistry with what innovations DOS has already brought to this field, we can define DOS in chemical terms as being a combinatorial chemistry capable of combining in several dimensions, in a simple and well thought-out manner, different elements of diversity in one or several steps. DOS amplifies and exploits efficiently the creation of topological diversity through stereochemistry, the formation of novel skeletons etc., with the goal of accessing quickly a multitude of novel, structurally complex and very varied molecules.

In terms of chemical genetics, DOS defines a ‘chemical space’ that is populated by small molecules accessible to the chemist. These hold an inherent exploratory potential for their complex and diverse structures. The impact of these small molecules on the ‘biological space’ to be explored (the protein) must allow useful information to be acquired. These data then become relevant for the development of novel small-molecule libraries high in information content.

DOS thus seeks to make the chemical space created by the probe molecules coincide with the known ‘biological space’ of proteins (the reverse chemical genetic approach) or those as yet undiscovered (by forward chemical genetics).

10.7. REFERENCES

- BEELER A.B., SCHAUS S.E., PORCO J.A. (2005) Chemical library synthesis using convergent approaches. *Curr. Opin. Chem. Biol.* **9**: 277-284
- BURKE M.D., BERGER E.M., SCHREIBER S.L. (2003) Generating diverse skeletons of small molecules combinatorially. *Science* **302**: 613-618
- BURKE M.D., SCHREIBER S.L. (2004) A planning strategy for diversity-oriented synthesis. *Angew. Chem., Int. Ed.* **43**: 46-58
- BURKE M.D., BERGER E.M., SCHREIBER S.L. (2004) A synthesis strategy yielding skeletally diverse small molecules combinatorially. *J. Am. Chem. Soc.* **126**: 14095-14104
- CHEN C., LI X., NEUMANN C.S., LO M.M.C., SCHREIBER S.L. (2005) Convergent diversity-oriented synthesis of small-molecule hybrids. *Angew. Chem., Int. Ed.* **44**: 2249-2252
- FALCIANI C., LOZZI L., PINI A., BRACCI L. (2005) Bioactive peptides from libraries. *Chem. Biol.* **12**: 417-426
- KIM Y.-K., ARAI M.A., ARAI T., LAMENZO J.O., DEAN E.F., III, PATTERSON N., CLEMONS P.A., SCHREIBER S.L. (2004) Relationship of stereochemical and skeletal diversity of small molecules to cellular measurement space. *J. Am. Chem. Soc.* **126**: 14740-14745
- LEE P.H., LEE K., KANG Y. (2006) *In situ* generation of vinyl allenes and its applications to one-pot assembly of cyclohexene, cyclooctadiene, 3,7-nonadienone, and

bicycle[6.4.0]dodecene derivatives with palladium-catalyzed multicomponent reactions. *J. Am. Chem. Soc.* **128**: 1139-1146

MEHTA G., SINGH V. (2002) Hybrid systems through natural product leads: an approach towards new molecular entities. *Chem. Soc. Rev.* **31**: 324-334

RAMÓN D.J., YUS M. (2005) Asymmetric multicomponent reactions (AMCRs): the new frontier. *Angew. Chem., Int. Ed.* **44**: 1602-1634

SCHREIBER S.L. (2000) Target-oriented and diversity-oriented organic synthesis in drug discovery. *Science* **287**: 1964-1969

SELLO J.K., ANDREANA P.R., LEE D., SCHREIBER S.L. (2003) Stereochemical Control of Skeletal Diversity. *Org. Lett.* **5**: 4125-4127

SPRING, D.R. (2005) Chemical genetics to chemical genomics: small molecules offer big insights. *Chem. Soc. Rev.* **34**: 472-48

STAVENGER R.A., SCHREIBER S.L. (2001) Asymmetric catalysis in diversity-oriented organic synthesis: Enantioselective synthesis of 4320 encoded and spatially segregated dihydropyrancarboxamides. *Angew. Chem., Int. Ed.* **40**: 3417-3421

TIETZE L.F., DELL H.P., CHANDRASEKHAR S. (2003) Natural product hybrids as new leads for drug discovery. *Angew. Chem., Int. Ed.* **42**: 3996-4028

ZHU J., BIENAYMÉ H. (2005) *Multicomponent Reactions*. Wiley-VCH, Weinheim

THIRD PART

TOWARDS AN IN SILICO EXPLORATION OF CHEMICAL AND BIOLOGICAL SPACES

MOLECULAR DESCRIPTORS AND SIMILARITY INDICES

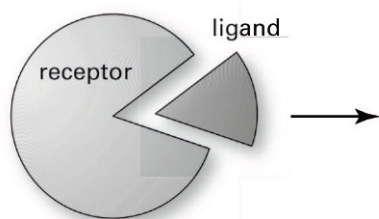
Samia Aci

11.1. INTRODUCTION

High-throughput and high-content pharmacological screening methods generate such a flood of chemical and biological data that their analysis would barely be feasible without the use of computational tools (chapters 6 and 15). As is the case with forward chemical genetics (chapter 8), for which the target itself is unknown and the biological information incomplete, the small molecule represents the best defined piece of data. Chemists, biologists and informaticians have to pull together as best they can in order to create from their joint observations ingenious hypotheses about the reasons for a molecule's bioactivity, or lack thereof. An apparently simple question to ask is: what is a molecule? And what sensible representation to make of it? Example 11.1 illustrates the range of possible answers.

Example 11.1 - different ways to 'look' at a molecule

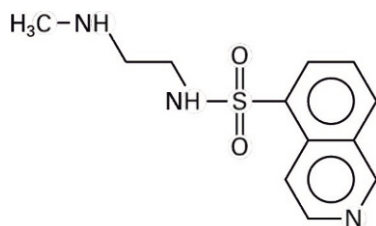
For the biologist



Global properties:

- ▶ IC₅₀
- ▶ log *P*
- ▶ molecular mass
- ▶ etc.

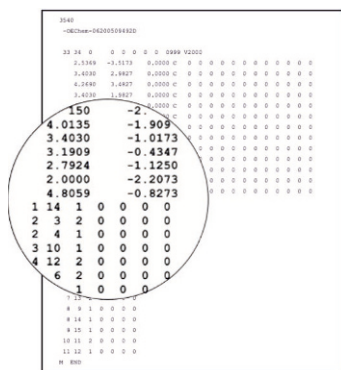
For the chemist



Functional properties:

- ▶ pK_a (acid/base character)
- ▶ H-bond donor/acceptor character
- ▶ nucleophilic/electrophilic character
- ▶ etc.

For the informatician



Graph properties:

- ▶ number of nodes
- ▶ number of edges
- ▶ number of node neighbours
- ▶ etc.

How do we go about bringing together these different views, each of which represents only one facet of the chemical entity, to enable the extraction of innovative concepts as to the reasons for bioactivity? More particularly, with regards to **virtual screening** (chapter 16), which **representation language** should be adopted to code for the chemical and structural properties of molecules in a format exploitable by the informatician and thus enabling him or her to extract chemically and biologically relevant information? This chapter presents a certain number of descriptors utilised in informatics for the characterisation of molecules. The aim of such a description is to introduce the concept of **similarity** between objects evaluating their degree of coverage in the space of the descriptors used to characterise them. How should this evaluation be carried out? Which measurement index should be adopted depending on the properties of the molecules to be evaluated for similarity or dissimilarity? Several of these indices will be detailed here. This chapter is not an exhaustive inventory of the whole set of molecular descriptors – which number in their hundreds or indeed thousands – nor of similarity measures (barely less numerous) but seeks to give to the reader an overview of the general categories that are available (see also chapter 12 for the particular point about hydrophobicity and chapter 13 for the recent developments in the annotation and classification of chemical space). The reader desiring more complete information is invited to consult the works dedicated to these subjects by TODESCHINI and CONSONNI (2000), and WILLET *et al.* (1998).

11.2. CHEMICAL FORMULAE AND COMPUTATIONAL REPRESENTATION

The calculation of, as well as the type of information given by, molecular descriptors depend on the chemical formula. They also depend on the computational representation of this chemical formula. Here, therefore, follows firstly a brief reminder of the concept of chemical formula. We shall see thereafter a way in which this information may be represented in calculations.

11.2.1. THE CHEMICAL FORMULA: A REPRESENTATION IN SEVERAL DIMENSIONS

A chemical formula can be depicted in three dimensions roughly corresponding to 1D to 3D geometrical spaces.

» 1D representation

The simple **molecular** formula gives information about the nature and the number of atoms that constitute a molecule (example 11.2).

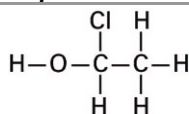
Example 11.2 - example of a molecular formula



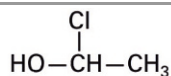
» 2D representation

The **KEKULÉ** (or **line**) structure and **condensed** formula provide information about the linkage of atoms in a molecule taking into account the bond types (example 11.3).

Example 11.3 - example of a KEKULÉ structure and a condensed formula



KEKULÉ structure

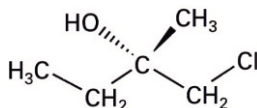


Condensed formula

» 3D representation

The **stereochemical** or **spatial** formula is a representation of the three-dimensional structure in a plane (example 11.4).

Example 11.4 - example of a stereochemical formula



11.2.2. MOLECULAR INFORMATION CONTENT

Each of the three chemical formulae mentioned above describes the assembly of different levels of molecular information, which can be broken down as follows:

[a]: the **atoms** of a molecule

[b]: the **inter-atomic bonds**

[c]: the **bond types**

[d]: the **elements of local geometry** (bond length, valence angle)

[e]: **chirality**, if it exists

[f]: dihedral **angles** (conformations)

Some of this information is deducible by grouping together several other levels. Thus level [d], comprising elements of local geometry, can be determined if the information in levels [a], [b] and [c] is already known. On the other hand, the chirality (level [e]) must be explicitly given as it cannot be deduced from the preceding levels.

To redescribe the chemical formulae as a function of these levels of molecular information, the **molecular formula** corresponds to information level $[a]$, the **condensed formula** contains levels $[a]$, $[b]$ and $[c]$ and, in addition to these, the **stereochemical formula** includes level $[e]$. The gaps in inherent molecular information with each of these formulae mean that going from one dimension to another leads to great ambiguity. Indeed, several condensed formulae can correspond to a single molecular formula (they are termed **structural** or **constitutional isomers**); similarly several spatial formulae can correspond to a single condensed formula (generating **stereoisomers**). The ambiguity is not totally resolved at the 3D-representation level since, in the case of the stereochemical formula, the lack of awareness of level $[f]$ implies that several conformational isomers or conformers may correspond to a single stereochemical formula. Example 11.5 illustrates these cases.

Example 11.5 - examples of the ambiguity brought about by the different chemical formulae: cases of isomerism

Structural isomers

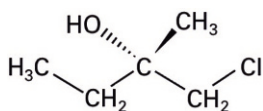
Molecular formula	2 corresponding condensed formulae		Type of isomerism
C_4H_{10}	$H_3C-CH_2-CH_2-CH_3$ <i>butane</i>	$H_3C-\underset{\substack{ \\ CH_3}}{CH}-CH_3$ <i>methylpropane</i>	Chain isomerism: isomers differing in their carbon chains
C_3H_8O	$HO-CH_2-CH_2-CH_3$ <i>propan-1-ol</i>	$H_3C-\underset{\substack{ \\ OH}}{CH}-CH_3$ <i>propan-2-ol</i>	
C_3H_8O	$HO-CH_2-CH_2-CH_3$ <i>propan-1-ol</i>	$H_3C-O-CH_2-CH_3$ <i>methoxyethane</i>	Functional group isomerism: isomers differing in their type of functional group

Stereoisomer configuration

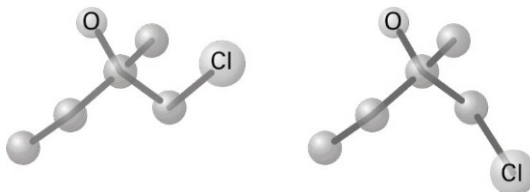
Condensed formula	2 corresponding stereochemical formulae		Type of isomerism
$\begin{array}{c} Cl \\ \\ HO-CH-CH_3 \end{array}$ <i>1-chloroethanol</i>	$\begin{array}{c} Cl \\ \\ HO-C \begin{array}{l} \nearrow H \\ \searrow CH_3 \end{array} \end{array}$ <i>(1S) 1-chloroethanol</i>	$\begin{array}{c} Cl \\ \\ H-C \begin{array}{l} \nearrow H \\ \searrow CH_3 \end{array} \end{array}$ <i>(1R) 1-chloroethanol</i>	Enantiomers: compounds that are non-superimposable mirror images of one another (also called optical isomers)
$H_3C-CH=CH-CH_3$ <i>2-butene</i>	$\begin{array}{c} H_3C & & CH_3 \\ & \diagdown & / \\ & C=C \\ & / & \diagdown \\ H & & H \end{array}$ <i>Z configuration</i> (from <i>zusammen</i> , meaning together in German)	$\begin{array}{c} H_3C & & H \\ & \diagdown & / \\ & C=C \\ & / & \diagdown \\ H & & CH_3 \end{array}$ <i>E configuration</i> (from <i>entgegen</i> , meaning opposite in German)	
			Diastereoisomers: geometrical stereoisomers

Conformational isomers (or conformers)

Stereochemical formula



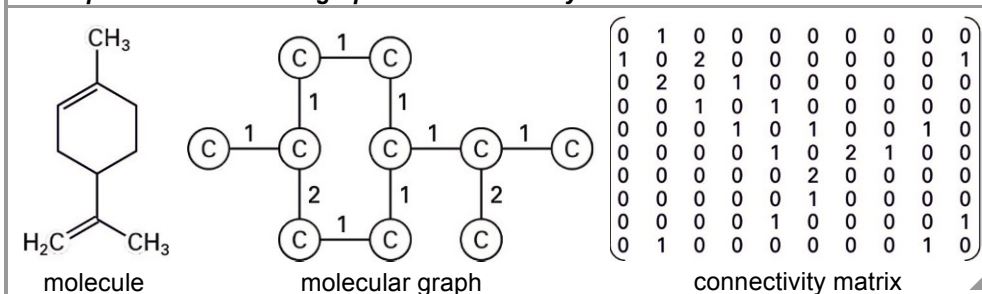
2 corresponding three-dimensional conformations



The term 3D representation appears to be relatively imprecise with respect to the stereochemical formula. Indeed, it does seem surprising to keep a formula, termed 3D, that does not actually permit the coordinates of the molecule's atoms in three-dimensional space to be known. The designation of '3D representation' for the stereochemical formula comes from the fact that it contains some three-dimensional information such as chirality, thus placing it above the 2D level. However this information is not sufficient to obtain a precise geometry of the molecule since the dihedral angles are unknown; it just allows the ambiguity about the possible stereoisomers to be lifted (see example 11.5). The determination of conformers arising from a single stereochemical formula is more or less difficult to achieve depending on the molecule studied. The number of possible conformations of the same molecule is very dependent on the number of **rotatable bonds** that it possesses, and the more linear and flexible a molecule, the more arduous it becomes to understand all of its conformations. A better 3D representation of a molecule therefore corresponds to knowing its three-dimensional conformations. Later in this chapter we shall consider this definition of a molecule's 3D formula instead of its stereochemical formula.

11.2.3. MOLECULAR GRAPH AND CONNECTIVITY MATRIX

One efficient representation of a molecule's 2D structure from the point of view of calculation is the **molecular graph**. This graph is a simple structure composed of a set of points called **vertices**, related to each other by links termed **edges** or arcs. In a molecular graph, the vertices are associated to the atoms of the molecule and the edges correspond to existing bonds between atoms. A graph can be weighted, *i.e.* a 'weight' is given to each edge of the graph; in the case of molecules, this feature of these graphs frequently serves to code for the bond type (single bond = 1, double bond = 2 etc.). Generally, hydrogen atoms are omitted from this kind of representation. An alternative way to describe this graph computationally is to use a connectivity matrix, in which the rows and columns correspond to the graph's vertices. The value of the edge (or bond) between vertices i and j is contained in the element C_{ij} of the matrix. These representations of the molecule are illustrated in example 11.6.

Example 11.6 - molecular graph and connectivity matrix for limonene

11.3. MOLECULAR DESCRIPTORS

There are different ways **to classify molecular descriptors**. Each method has its advantages and none allows an absolute classification of molecular descriptors.

Let us take for example two methods that are commonly employed to classify descriptors:

- » dependent on the nature of the descriptor: *i.e.* its capacity to describe the molecule as a function either of its physicochemical properties, or of its structural or constitutive or even electronic properties.
- » dependent on its dimension: *i.e.* the dimension of the chemical formula (1D, 2D or 3D, as shown above) from which the descriptor can be extracted.

For some authors, the dimension of the molecular descriptor does not correspond to the dimension of the chemical formula from which it was extracted but to the spatial dimension corresponding to this descriptor. For example, a molecular descriptor composed of *one* unique index (and therefore *one* number associated with the molecule) will correspond to a *1D* index whereas a molecular descriptor composed of *n* indices will describe a space in *n* dimensions.

For other authors, a classification can also be made between ‘first order’ and ‘second order’ descriptors: every predictive empirical model returning the value calculated from a macroscopic property as a function of ‘basic’ (or ‘first order’, *i.e.* arising directly from the structural analysis) molecular descriptors becomes automatically a molecular descriptor of the ‘second order’. For example, the lipophilicity index (chapter 12) can be considered to be a ‘second order’ descriptor; calculated from ‘basic’ descriptors (either 2D – constitutive or fragment-based – or 3D; see chapter 12), it can thereafter be used as a descriptor in a model of intestinal permeability.

However in the rest of this chapter, for reasons of simplicity, the **molecular descriptor’s dimension** will correspond to the **dimension of the chemical formula from which it was extracted** and so the molecular descriptors mentioned will be classified with respect to this dimension.

11.3.1. 1D DESCRIPTORS

1D descriptors are those extracted from the simple molecular formula of a molecule. The number of descriptors arising from such a representation is limited however, and they only capture very global properties of a molecule since the molecular formula does not even offer a distinction between structural isomers. Descriptors such as the **molecular weight** or the number and types of atoms constituting the molecule provide examples of the most frequently used 1D molecular descriptors found in the literature.

11.3.2. 2D DESCRIPTORS

The descriptors arising from a molecule's structural formula (KEKULÉ or condensed formula) permit an association to be made between the properties of the molecule studied and its topology. Several hundred of these descriptors can be found in the literature; they can be distinguished as much by the nature of the molecule's properties they describe as by their form.

Indices

Indices are coded descriptors in the form of one or more numbers. Several hundred exist in the literature and it would be tedious to describe here the complete list. The interested reader can find already sorted a substantial list of references on the following website:

http://www.taletе.mi.it/products/dragon_molecular_descriptors.htm

- » **Topological indices** characterise the size, the degree of branching and the global form of the molecular structure but ignore the nature of the atoms involved. They serve rather in the context of a single class of compounds; for example MERCADER *et al.* (2001) used topological indices to predict the enthalpies of formation of about sixty hydrocarbons.

Example 11.7 - WIENER index

$$W = \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N D_{ij}$$

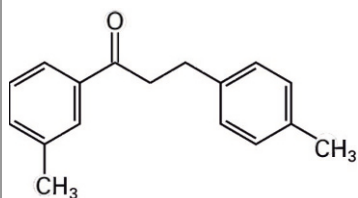
where D_{ij} is the number of bonds (= topological distance) between atoms i and j of the molecule.

- » **Constitutional indices** reveal the nature of a molecule's different components. They not only describe the nature of the atoms involved, but also their degree of hybridisation, the nature of the bonds and the presence of aliphatic chains (for example, $\#arings$ = number of aromatic rings in a molecule). Some indices permit the electronic state and the valence of the atoms to be related to their topological arrangement (KIER and HALL, 1990; HALL and KIER, 1995).
- » **Property indices** allow a description of the overall physicochemical properties of the molecule, e.g. the lipophilicity index (chapter 12).

Descriptors based on a decomposition into fragments

- » **Dictionary-based fingerprints** (or **structural keys**) are composed of a predefined set of molecular fragments, stored in **bits** and linked in a **bit string**. Depending on the fingerprint, the bit strings are either ‘binary’ or ‘continuous’ (example 11.8). In a binary fingerprint, only the presence or absence in the described molecule of predefined fragments can be detected. If a given fragment exists in the molecule, the corresponding bit in the string is assigned the value 1; else if it does not, the value 0 is assigned. For a ‘continuous’ type of fingerprint, not only is it specified whether a fragment does or does not appear in a molecule, but the number of occurrences of this fragment in the molecule is also quantified.

Example 11.8 - ‘Binary’ and ‘continuous’ fingerprints



Let us imagine a molecule (left) with the following dictionary of predefined fragments:

1	2	3	4	5	6
$\square\text{CH}_3$	$\square\text{CH}_2\square$	$\square\text{CH}<$	C_6H_5	$>\text{C}=\text{O}$	$\square\text{NH}_2$

A **binary** $\square$ fingerprint for this molecule would correspond to:

1	2	3	4	5	6
1	1	0	1	1	0

A **continuous** $\square$ fingerprint corresponds to:

1	2	3	4	5	6
2	2	0	2	1	0

- » **Hashed fingerprints** are very similar to the aforementioned fingerprints with the difference that the stored fragments are not predefined by the user but constructed from the molecule; a fingerprint is thus built specifically for each molecule.
- » The **‘atom pair’ descriptor** (CARHART *et al.*, 1985) also relies on a bit string: however, the bits no longer correspond to particular substructures or chemical functions sought in the molecule, but to portions of its topological graph. An ‘atom pair’ has the form:

$$\text{Type } X_n\text{-(}d_T\text{)-Type } X_n$$

where *Type* is the atom type, *n* the number of its non-hydrogen atoms and *d_T* the topological distance between the two atoms of the pair.

- » Based on the same principle, *i.e.* the description of the molecule’s topological graph, NILAKATAN *et al.* (1987) described a **‘topological torsion’ descriptor** as follows:

$$(n\text{Pi-Type-nBr})\text{-(nPi-Type-nBr)}\text{-(nPi-Type-nBr)}\text{-(nPi-Type-nBr)}$$

where *nPi* is the number of *Pi* electrons in the atom, *Type* the type of atom and *nBr* the number of bonds the atom in question makes with its non-hydrogen neighbours.

Descriptors based on pharmacophore properties

These descriptors link the physicochemical properties of the atoms to the molecule's topology. Here, the atoms are no longer described by their type nor by their properties directly accessible in the Periodic Table, but by the properties acquired by their layout in the molecular scaffold. The pharmacophore properties mainly utilised in this type of representation are:

- › hydrogen bond donor,
- › hydrogen bond acceptor,
- › cation (or acid),
- › anion (or base),
- › hydrophobic centre,
- › aromatic ring.

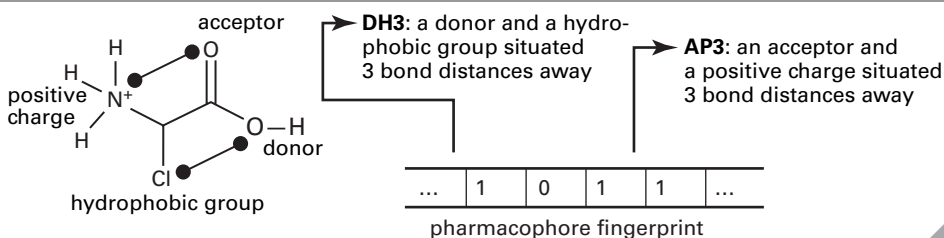
Even if the properties described by these descriptors are the same for all, the information is however conveyed differently depending on the descriptor used.

» 2-point pharmacophores

Inspired by the 'atom pair' and 'topological torsion' descriptors KEARSLEY *et al.* (1996) conceived the first version of these descriptors by replacing the atom types with their pharmacophore functions. The principle for building these descriptors is as follows: first of all, the set of all possible pairs combining the types of pharmacophore points considered (for example, the six listed above) with all of the topological distances considered for the molecule (e.g. the distances of 1 to 12 bonds) is calculated. These pairs are next stored in a bit string analogous to that described for structural keys. In the same way, for each molecule information is stored about either the presence or absence of a given pair, or the number of occurrences of this pair in the molecule. The string thus obtained is called the **pharmacophore fingerprint** and is used in similarity calculations (example 11.9).

This type of descriptor possesses, among other advantages, the capacity to offer the chemist readily understandable and usable reasons for the activity and/or inactivity. Indeed, an activity rule of the type: "*for a molecule to be active, a hydrogen bond donor function and an aromatic ring are needed situated a topological distance away from each other of between 5 and 7 inclusive*", 'speaks' much more to a chemist seeking a new active compound than a rule of the type: "*the molecule's WIENER index must be between 15 and 20*".

Example 11.9 - 2-point pharmacophore



» **BCUT descriptor**

This method, developed by PEARLMAN *et al.* (1998) derives from a technique described by BURDEN *et al.* (1989) and aims to identify the organic molecules by an index calculated from connectivity matrices arising from a molecule's graph. In the matrix calculated by BURDEN, the atom numbers of non-hydrogen atoms are placed in the diagonal elements and the type of bond between an atom i and an atom j is contained in the element $[M_{ij}]$. The index, strictly speaking, then consists of a combination of the N (typically two) greatest eigenvalues extracted from the matrix. PEARLMAN adopted this same matrix but in the place of atom numbers, the diagonal contains either their charge, their polarisability, or their capacity to form hydrogen bonds. PEARLMAN thus had 3 different matrices and used as a molecular descriptor 6 numerical values from these matrices by extracting from each its largest and smallest eigenvalues.

11.3.3. 3D DESCRIPTORS

The main disadvantage of 2D descriptors is their inability to take into account the real form of molecules (including chirality!) and the steric effects liable to interfere with the interactions between the receptor and its substrate. A new generation of descriptors based on the three-dimensional conformation of a molecule has thus been developed with the aim of alleviating this problem. More often employed (as the habit of working with 3D molecular conformations develops), other types exist that are even more diversified than 2D descriptors.

Indices

The majority of indices originally used for 2D molecular structures have been adapted in order to be calculatable for the three-dimensional conformations of these same molecules.

- » **Geometric indices** provide information about the spatial organisation of the atoms.
- » **Property indices** supply mostly global information either about the physicochemical properties of the molecule, or about the form of the molecular conformation, for example the VAN DER WAALS volume, the solvent-accessible surface, the principal moment of inertia, and also again the lipophilicity index.

Descriptors based on pharmacophore properties

These are the same descriptors as described in 2D but adapted for use with 3D structures.

» **2-, 3- and 4-point pharmacophores**

The same reasoning as that explained previously for the 2-point pharmacophores in 2D is applied but intervals of geometric distance are employed here to construct pharmacophore fingerprints. For example, PICKETT *et al.* (1996) constructed 3-point pharmacophores (see chapter 13) and used in their study the

following six geometric distance intervals: {2-4.5, 4.5-7, 7-10, 10-14, 14-19, 19-24}, which allowed them to generate a set of 5916 possible triangles. These descriptors were also extended for use with 4-point pharmacophores (MASON *et al.*, 1999). This generalisation with 2- to 4-point pharmacophores can also be applied to 2D.

» **BCUT descriptor**

The generation of descriptors of the BCUT type from 3D structures is distinguished from 2D descriptors since, for a molecule's i and j atoms, the element $[M_{ij}]_{i \neq j}$ of the corresponding matrix no longer contains the type of possible bond but the geometric distance between the two atoms.

Descriptors based on interaction fields

These descriptors relate the biological activity to the interaction properties and the three-dimensional form of the molecule. They offer the advantage of taking into account the flexible character of molecules. Calculation of a molecule's interaction field in space and the use, in order to achieve this, of three-dimensional calculation grids, bring together these descriptors for **molecular modelling** and drug-design methods. They are particularly suited to the search for explanatory reasons for a molecule's activity, however, **they require knowledge of the reliable three-dimensional conformations** of the molecule, as well as having a suitable force field. The best known of these descriptors are the CoMFA (for Comparative Molecular Field Analysis, CRAMER *et al.*, 1988) and CoMSIA (Comparative Molecular Similarity Index Analysis, KLEBE *et al.*, 1999) approaches. Besides their considerable requirements from the point of view of calculation, the main difficulty with these descriptors resides in the need to align molecules together. The GRIND descriptor (*GRid-INdependent Descriptors*), published by PASTOR (PASTOR *et al.*, 2000), offers the advantage of allowing the calculation of a molecular interaction field while avoiding the problem of molecular alignment.

Descriptors derived from spectroscopic methods

The use of three-dimensional structures for molecules, with a view to represent their real nature as closely as possible, led researchers quite naturally to consider that an efficacious means to associate an *in silico* measurement with a molecule would be to use the same methods as those routinely used in the laboratory, *i.e.* spectroscopic methods. This idea led to the emergence of descriptors using the same methods as for chemical analysis, e.g. the EVA descriptor (standing for *Eigen-Value*; GINN *et al.*, 1997) calculated from the vibrational spectrum of molecules. This path was also chosen to calculate the descriptors such as 3D-MoRSE (Molecule Representation of Structures based on Electron diffraction; SCHUUR *et al.*, 1996) or RDF (Radial Distribution Function; HEMMER *et al.*, 1999).

Other descriptors

The growing use of three-dimensional conformations of molecules has led to the design of numerous other descriptors that are difficult to classify into a specific category. Let us take for example the WHIM descriptor (TODESCHINI *et al.*, 1994) the objective of which is to avoid molecular alignment problems by the use of projection along axes, or the GETAWAY descriptor (CONSONNI and TODESCHINI, 2001), which aims to establish a correspondence between the molecular geometry and topology, and the physicochemical properties of atoms.

11.3.4. 3D VERSUS 2D DESCRIPTORS?

The advent of 3D descriptors, thanks to the increasing use of spatial conformations of molecules, permitted exploitation not only of the atoms' organisation relative to one another but also the general form of a molecule and the interaction properties thus conferred upon it. Some of these 3D descriptors, not merely possessing a predictive capacity – indeed able to explain the bioactivity – are, besides, capable of constructing a 3D 'model' of the motif responsible for the activity. However, although the advances achieved in the last few years in the design of powerful 3D descriptors are considerable, the use of 2D descriptors and hence the corresponding structures remains highly developed in the scientific community. This behaviour holds in part from the conviction of some scientists that judicious use of 2D descriptors would allow prediction of the 3D parameters and would thus be equally efficient in predicting the bioactivity of molecules as with 3D descriptors and for a lower cost of calculation (ESTRADA *et al.*, 2001). Another reason for the mistrust incurred by the use of 3D descriptors is the necessity, in order to obtain them, of generating three-dimensional conformations of the molecules to be targeted, engendering the difficulty of increasing considerably the number of molecules for processing without the assurance even that the active conformation(s) are really present among those obtained. In the absence of information about the potential bioactive geometries (which is normally the case in searching for structure-activity relationships) it is necessary to ensure that the descriptor is tolerant of geometric fluctuations, such that the region occupied by a molecule as a 'point' with coordinates D_i in the 'structural' space defined by these descriptors (see chapter 13) does not change radically if the conformations first employed to calculate the coordinates D_i are altered. Otherwise, there is a risk of finding a molecule to be highly dissimilar to itself, and simply by having used conformers that were too different!

The real problem lies, in fact, in evaluating judiciously the cost of inclusion in the modelling of 3D conformations as a function of the information gathered and thus of determining whether the calculation of conformations is necessary to obtain a piece of crucial information, or whether or not this information can be estimated from the 2D representation.

11.4. MOLECULAR SIMILARITY

11.4.1. A BRIEF HISTORY

The concept of similarity has gained its importance in the field of biology and not only in the search for molecules of biological interest. This is why many researchers set about the arduous task of designing indices likely to convey the similarity between objects that are similar in nature (SNEATH and SOKAL, 1973). The first research into similarity applied to molecules, instead of searching for common substructures as was common practice, dates from the middle of the 1980s and was carried out quite close together in time by CAHART *et al.* (1985) and WILLET *et al.* (1986). In these two studies the authors used fragments of molecules to make their comparisons. It is of note that this practice then became generalised, as it is principally by using 2D descriptors of the dictionary-based, molecular or pharmacophore fingerprint type that similarity research has seen its great boom in the world of virtual screening.

11.4.2. PROPERTIES OF SIMILARITY COEFFICIENTS AND DISTANCE INDICES

Four main types of similarity coefficient, from all domains, can be distinguished:

» **Coefficients based on distance**

These coefficients are analogous to distances in geometric space; the best known of these coefficients is quite simply Euclidian distance.

» **Coefficients based on association**

These coefficients compare the attributes that two objects have in common relative to the set of attributes; these indices are popular in part because their construction even confers a sort of self-normalisation, the most frequently used is the TANIMOTO index (example 11.10).

Example 11.10 - calculation of the TANIMOTO index for binary fingerprints

Let *A* and *B* be two binary fingerprints, and *a* the number of bits with a value of 1 in *A*, *b* this number in *B* and *c* the number of bits having a value of 1 common to both *A* and *B*.

A:	1	0	0	1	1	0	0	1	0	1	1	1	<i>a</i> = 7	<i>c</i> = 3
B:	0	0	0	1	0	1	0	1	1	0	1	0	<i>b</i> = 5	

The TANIMOTO index *S* is given by:
$$S = \frac{c}{a+b-c} = \frac{3}{7+5-3} = 0,33.$$

» **Correlation coefficients**

These methods measure the degree of correlation between a pair of objects as, for example, in linear regression.

» **Probabilistic coefficients**

These indices are based on the principle that it is less probable for two objects to have two rare fragments in common than two more common fragments. It is therefore necessary to give more weight to the fact of having two rare fragments

in common. An example of a probabilistic coefficient is the GOODALL index (GOODALL, 1964). It is notable that, according to ADAMSON and BUSH (1975), this type of similarity method is less efficient than the others in chemistry.

11.4.3. A FEW SIMILARITY COEFFICIENTS

With regards to their advantages and disadvantages, it is not easy to define the best similarity coefficient and this is not the purpose of this chapter. Depending on the needs and the type of similarity, and on the types of descriptors used and the size of molecules, the reader will favour one or the other. Nevertheless, it can be said that the TANIMOTO coefficient remains a preferential criterion for similarity searching involving binary-type fingerprints, while Euclidian distance remains the most used similarity coefficient when the data are of a continuous type (see also chapter 13). [Table 11.1](#) opposite presents the formulae of a few commonly used indices in chemical similarity searching.

11.5. CONCLUSION

The choice of descriptors as well as the similarity index are fundamental criteria in virtual screening and for the prediction of potentially bioactive molecules. Currently, there exists no unique and universal representation of molecules for this purpose, and it is up to each researcher to decide the most representative and suitable descriptors to use depending on the type of search and the answers desired.

The search for the best possible description of a molecule leads us incidentally to ask ourselves about the level of representation of the molecule to keep: should we only keep the properties acquired by the atoms due to their own arrangement within a molecule, or do the nature of these atoms risk intervening in some other manner? What level of information is the most sensible to keep? Do they not in fact all have a role to play? How should we also represent molecules that can adopt different chemical forms (tautomers, for example)? In this case, perhaps it is necessary to ask the question: how can we integrate multiple representations of a molecule in order to harvest all of the necessary information for our research, whatever the level of representation it belongs to?

Table 11.1 - Formulae for common indices in chemical similarity searches

Let $A = \{x_{iA}\}_{i=1}^N$ et $B = \{x_{iB}\}_{i=1}^N$ be two molecular fingerprints.

In the case of binary molecules, let a be the number of 'bits' having a value of 1 in A , b this number in B and c the number of 'bits' in common for A and B .

	Formula for 'continuous' data	Formula for 'binary' data
HAMMING distance	$D_{AB} = \sum_{i=1}^N x_{iA} - x_{iB} $	$D_{AB} = a + b - 2c$
Euclidian distance	$D_{AB} = \left[\sum_{i=1}^N (x_{iA} - x_{iB})^2 \right]^{1/2}$	$D_{AB} = \sqrt{a + b - 2c}$
SOERGEL distance	$D_{AB} = \frac{\sum_{i=1}^N x_{iA} - x_{iB} }{\sum_{i=1}^N \max(x_{iA}, x_{iB})}$	$D_{AB} = \frac{a + b - 2c}{a + b - c}$
TANIMOTO coefficient	$S_{AB} = \frac{\sum_{i=1}^N x_{iA} x_{iB}}{\sum_{i=1}^N (x_{iA})^2 + \sum_{i=1}^N (x_{iB})^2 - \sum_{i=1}^N x_{iA} x_{iB}}$	$S_{AB} = \frac{c}{a + b - c}$
DICE coefficient	$S_{AB} = \frac{2 \sum_{i=1}^N x_{iA} x_{iB}}{\sum_{i=1}^N (x_{iA})^2 + \sum_{i=1}^N (x_{iB})^2}$	$S_{AB} = \frac{2c}{a + b}$
COSINE similarity	$S_{AB} = \frac{\sum_{i=1}^N x_{iA} x_{iB}}{\left[\sum_{i=1}^N (x_{iA})^2 \sum_{i=1}^N (x_{iB})^2 \right]^{1/2}}$	$S_{AB} = \frac{c}{\sqrt{a \cdot b}}$
TVERSKY index	—	$S_{Tversky} = \frac{c}{\alpha(a - c) + \beta(b - c) + c}$ where α and β are chosen by the user

11.6. REFERENCES

- ADAMSON G.W., BUSH J.A. (1975) A comparison of the performance of some similarity and dissimilarity measure in the automatic classification of chemical structures. *J. Chem. Inf. Comput. Sci.* **15**: 55-58
- BURDEN F.R. (1989) Molecular identification number for substructure searches. *J. Chem. Inf. Comput. Sci.* **29**: 225-227
- CARHART R.E., SMITH D.H., VENKATAGHAVAN R. (1985) Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **25**: 64-73
- CONSONNI V., TODESCHINI R. (2001) GETAWAY descriptors: new molecular descriptors combining geometrical, topological and chemical information for physico-chemical properties modelling and drug design. In *Rational approaches to drug design* (HÖLTJE H.D., SIPPL W. Eds) Prous Science, Barcelona, Spain: 235-240
- CRAMER R.D., PATTERSON D.E., BUNCE J.D. (1988) Comparative molecular field analysis (CoMFA). Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**: 5959-5967
- ESTRADA E., MOLINA E., PERDORNO-LOPEZ I. (2001) Can 3D structural parameters be predicted from 2D (topological) molecular descriptors? *J. Chem. Inf. Comput. Sci.* **41**: 1015-1021
- GINN C.M.R., TURNER D.B., WILLETT P. (1997) Similarity searching in files of three-dimensional chemical structures: evaluation of the EVA descriptor and combination of rankings using data fusion. *J. Chem. Inf. Comput. Sci.* **37**: 23-37
- GOODALL D.W. (1964) A probabilistic similarity index. *Nature* **203**: 1098
- HALL L.H., KIER L.B. (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. *J. Chem. Inf. Comput. Sci.* **35**: 1039-1045
- HEMMER M.C., STEINHAEUER V., GASTEIGER J. (1999) Deriving the 3D structure of organic molecules from their infrared spectra. *Vibrat. Spect.* **19**: 151-164
- KIER L.B., HALL L.H. (1990) An electrotopological state index for atoms in molecules. *Pharm. Res.* **7**: 801-807
- KEARSLEY S.K., SALLAMACK S., FLUDER E.M., ANDOSE J.D., MOSLEY R.T., SHERIDAN R.P. (1996) Chemical Similarity Using Physicochemical Property Descriptors. *J. Chem. Inf. Comput. Sci.* **36**: 118-127
- KLEBE G., ABRAHAM U. (1998) Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J. Comput. Aided Mol. Des.* **13**(1): 1-10
- MASON J.S., CHENEY D.L. (1999) Ligand-receptor 3-D similarity studies using multiple 4-point pharmacophores. *Pac. Symp. Biocomput.* **4**: 456-467

- MERCADER A., CASTRO E.A., TOROPOV A.A. (2001) Maximum topological distances based indices as molecular descriptors for QSPR. 4. Modeling the enthalpy of formation of hydrocarbons from elements. *Int. J. Mol. Sci.* **2**: 121-132
- NILAKANTAN R., BAUMAN N., DIXON J.S., VENKATARAGHAVAN R. (1987) Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **27**: 82-85
- PASTOR M., CRUCIANI G., MCLAY I., PICKETT S., CLEMENTI S. (2000) Grid independent descriptors (GRIND). A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **43**: 3233-3243
- PEARLMAN R.S., SMITH K.M. (1998) Novel software tools for chemical diversity. *Perspect. Drug Discovery Des.* **9-10-11**: 339-353
- PICKETT S.D., MASON J.S., MCLAY I.M. (1996) Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries (PDQ). *J. Chem. Inf. Comput. Sci.* **36**: 1214-1223
- SCHUUR J.H., SELZER P., GASTEIGER J. (1996) The coding of the three-dimensional structure of molecules by molecular transforms and its application to structure-spectra correlations and studies of biological activity. *J. Chem. Inf. Comput. Sci.* **36**: 334-344
- SNEATH P.H.A., SOKAL R.R. (1973) In *Numerical taxonomy: the principle and practice of numerical classification*, Freeman and Co, San Fransisco, USA
- TODESCHINI R., CONSONNI V. (2000) In *Handbook of molecular descriptors* (MANNHOLD R.H., KUBINYI H., TIMMERMAN H. Eds) Wiley, New York, USA
- TODESCHINI R., LASAGNI R., MARENGO E. (1994) New molecular descriptors for 2D and 3D structures. *Theory. J. Chemometric* **8**: 263-272
- WILLET P., BARNARD J.M., DOWNS G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**: 983-996
- WILLET P., WINTERMAN V. (1986) Implementation of nearest-neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **26**: 36-41

Chapter 12

MOLECULAR LIPOPHILICITY: A PREDOMINANT DESCRIPTOR FOR QSAR

G rard GRASSY - Alain CHAVANIEU

12.1. INTRODUCTION

The discovery of a new medicine is a complex and costly process, in which the role of chance and fortuitous observation is often predominant. Substantial effort has been afforded to rationalise the discovery process as much as possible. The methodologies relevant to **rational drug design** (according to international terminology) are today omnipresent as much in academic research centres as in industry.

Rational drug design is undertaken in a very different way depending on the level of knowledge that one has about the biological system concerned. In an ideal case, studies in structural biology allied to good knowledge of the physiological mechanism allow observation of the protein target in the presence of natural agonists and to understand completely the nature of all the key interactions. In other cases, one possesses no element permitting the development of a rational strategy. This situation, unfortunately, occurs frequently and one can only attempt to develop a strategy for establishing a **structure-activity relationship (SAR)** if a group of products is available that is likely to constitute a reliable basis for statistical studies. In each case the process of establishing a SAR begins with a molecule displaying a biological activity with therapeutic potential and consists of jointly optimising both its chemical structure and its activity. The majority of SAR studies were initiated by chemists having sufficient potential to explore the reactivity of a compound and to construct a series of analogous compounds with which to experiment. The setup of SAR studies or rather **QSAR** when they are quantitative, assumes the establishment of wide experimental support.

12.2. HISTORY

The relationship between chemical structure and bioactivity was accepted as axiom by chemists from the middle of the XIX century, but the first **quantitative relation-**

ship between activity and property was established in 1899 by MEYER and OVERTON, two English researchers. They studied separately the anaesthetic power of certain gases (xenon, chloroform, cyclopropane, fluorinated gases), and perceived that their activity was correlated with their solubility in lipids (MEYER, 1899; MEYER, 1901; OVERTON, 1901); the more a gaseous molecule was liposoluble, the greater its anaesthetic power. This seemed logical since neuronal membranes are composed of 90% lipids (phospholipids, cholesterol etc.). However, experiment often contradicts this theory! Gases do exist which, although possessing a very high solubility coefficient, have no anaesthetic power. It has since been accepted that the target of anaesthetics is a protein and does not correspond to simple solubility in the membranes of neurons.

The formalisation of structure-activity relationships goes back to the 1950s. Corwin HANSCH founded the conceptual bases for **quantitative SAR (QSAR)** and the first methodological elements in detail (HANSCH and MUIR, 1951; HANSCH and MUIR, 1961; HANSCH *et al.*, 1962). Since this period, QSAR has seen a significant progress, integrating notably statistical methodologies, the benefits of structural biology, three-dimensional information about biomolecules (chapters 14 and 16) and methods arising from artificial intelligence like neuronal networks and genetic algorithms (chapter 15). If the evolution of this practice over the last thirty years were to be depicted symbolically, then one could write that **Qsar** has evolved into **qSAR**. The notion ‘quantitative’, so important at the beginning in optimisation studies, is no longer an absolute priority today.

12.3. THEORETICAL FOUNDATIONS AND PRINCIPLES OF THE RELATIONSHIP BETWEEN THE STRUCTURE OF A SMALL MOLECULE AND ITS BIOACTIVITY

12.3.1. QSAR, QPAR AND QSPR

The basic principle resides in the axiom stating that there exists a causality dependence between a molecule’s chemical structure and its activity on a given biological entity. In an analogous way to ANFENSEN’s principle which stipulates that “*all the information necessary to obtain the native conformation of a protein in a given environment is contained in the amino-acid sequence*” (ANFENSEN and REDFIELD, 1956) it is accepted that a molecule’s chemical notation, as long as it determines its structure without ambiguity, contains all of the information necessary to predict its activities, including biological ones. But exploitation of this founding axiom necessitates the relationship to be established in a wider sense enabling the activity to be predicted from structural information. It is therefore necessary to establish all of the elements of the following relationship:

$$\text{Activity} = f(\text{structure}) \quad (\text{eq. 12.1})$$

Establishing a reliable study requires the determination and perfect knowledge of the three parts of the triptych: **chemical structure**, **bioactivity** and **relationship function**. From a strictly epistemological point of view, true relationships between structure and activity are rare since ‘quantifying the notion of structure’ necessitates recourse to techniques for form recognition. It is preferable **to extend the notion of structure to the properties and physicochemical descriptors stemming from a chemical structure**. The majority of QSAR (**Quantitative Structure-Activity Relationship**) studies are in reality QPAR (**Quantitative Property-Activity Relationship**) studies where the **S** of structure is replaced by the **P** of properties. This supposes as a corollary the establishment of a QSPR (**Quantitative Structure-Property Relationship**) which refers to the relationship between the structure and molecular properties.

$$\text{Activity} = f_1(\text{property}) \text{ and } \text{Property} = f_2(\text{structure}) \quad (\text{eq. 12.2})$$

In this book, we shall look at all of the aspects implicated in these relationships step by step and in detail.

12.3.2. BASIC EQUATION OF A QSAR STUDY

First, we define the following: BR_A is the biological response associated with a compound A , P_{nA} , the properties associated with the structure of compound A , and a_n the regression coefficients associated to the property n . These descriptors can refer either to a complete molecule (e.g. $\log P$, see below) or, in a homogenous series possessing a common skeleton, only represent the contribution of different substituents present on the basic skeleton (e.g. π). The basic equation of a QSAR study is given by:

$$BR_A = a_1 P_{1A} + a_2 P_{2A} + a_3 P_{3A} + \dots + a_n P_{nA} + C \quad (\text{eq. 12.3})$$

where C is a constant. In the best case, the set of properties taken into account in the equation must suitably represent the three main property types involved in the biological activity:

- › the descriptors of molecular lipophilicity supposed to be related to the distribution in biological microenvironments,
- › the steric descriptors representative of molecular crowding,
- › the electronic descriptors assumed to represent the interactions between the compound and its molecular biological target.

12.4. GENERALITIES ABOUT LIPOPHILICITY DESCRIPTORS

12.4.1. SOLUBILITY IN WATER AND IN LIPID PHASES: CONDITIONS FOR BIOAVAILABILITY

Water is the most important solvent and communication means adopted in living systems, as much by biomolecules as by xenobiotics (exogenous compounds). The

essential properties involved in the interaction between a substance and water are both its solubility and the characteristics of its partition equilibrium between water and another phase. Modelling the pharmacological action of a substance in an organism involves these two primordial notions which govern both its **bioavailability** and its fate. Paradoxically, whereas theoretical advances permit quite an easy explanation of a compound's **hydrophilicity** and yet its lipophilicity and hydrophobicity are much less accessible theoretically, it is the generic term **lipophilicity** that has been chosen to refer to the behaviour of organic molecules in solution within an organism. To estimate experimentally the distribution properties of a substance within an organism, we rely on its **partition coefficient between an aqueous phase and a more or less miscible organic phase**. This type of parameter is today considered by medicinal chemists as being representative of molecular lipophilicity.

12.4.2. PARTITION COEFFICIENTS

The first partition coefficient was used a century ago by MEYER and OVERTON to describe the 'non-specific' pharmacological activities of narcotic substances. **Initially measured between oil and water**, the partition coefficient is nowadays measured for other systems of partially miscible solvents. Many compounds are indeed hard to dissolve in oils. In current practice the partition coefficient measured for an **n-octanol/water system** ($P_{O/W}$) is the most often used for describing these partition properties. Octanol is considered rightly or wrongly as a near universal model for membranes (above all, those of erythrocytes), organic fats, and nearly all non-aqueous biophases, in so far as it allows satisfactory empirical correlations with the observed biological activity. Besides, its chemical stability, low toxicity, its almost non-existent volatility coupled with an extremely low absorbance in the ultraviolet region have favoured its use. Although many works have demonstrated the partial or total inadequacy of this system, it remains standard. A substance's solubility in water is strongly correlated to its partition coefficient $P_{O/W}$. Due to this fact, the majority of data in the literature concern the octanol-water system, and for this reason we shall take this as our study model.

12.4.3. THE PARTITION COEFFICIENT IS LINKED TO THE CHEMICAL POTENTIAL

The chemical potential μ of a non-electrolytic substance in a given solvent is related to the concentration expressed as the mole fraction x according to the equation:

$$\mu = \mu^\circ + RT \log x \quad (\text{eq. 12.4})$$

where μ° is the chemical potential of unitary mole fraction. For dilute solutions, this relation leads to:

$$\mu = \mu'^\circ + RT \log C \quad (\text{eq. 12.5})$$

where C is the solute's molarity in the solvent and μ'° the chemical potential of a 1 Molar solution under standard conditions of pressure and temperature.

When a solute is in equilibrium between two totally or partially miscible solvents, the chemical potential of this solute becomes equal in the two phases. With the octanol-water system and for a given compound:

$$\mu_{\text{oct}} = \mu_{\text{wat}} \quad (\text{eq. 12.6})$$

$$\mu_{\text{oct}}^{\circ} + R T \log x_{\text{oct}} = \mu_{\text{wat}}^{\circ} + R T \log x_{\text{wat}} \quad (\text{eq. 12.7})$$

$$\mu_{\text{wat}}^{\circ} - \mu_{\text{oct}}^{\circ} = R T \log x_{\text{oct}} - R T \log x_{\text{wat}} = R T \log (x_{\text{oct}}/x_{\text{wat}}) = R T \log K_{\text{O/W}} \quad (\text{eq. 12.8})$$

with $K_{\text{O/W}}$ being the partition coefficient in octanol/water expressed as the ratio of the mole fractions.

Classically, the partition coefficient is evaluated from the molarities of the solute in each of the solvents, in moles per litre: $P_{\text{O/W}} = C_{\text{oct}}/C_{\text{wat}}$. The value obtained is numerically different from that arising from the expression in mole fractions: $K_{\text{O/W}} = x_{\text{oct}}/x_{\text{wat}}$. Taking logarithms of the partition coefficients:

$$\log P_{\text{O/W}} = \log K_{\text{O/W}} - 0.94 \quad (\text{eq. 12.9})$$

12.4.4. THERMODYNAMIC ASPECTS OF LIPOPHILICITY

The quantity $(\mu_{\text{wat}}^{\circ} - \mu_{\text{oct}}^{\circ})$ corresponds to the standard free-energy change accompanying the transfer of one mole of solute from one solvent to another. The logarithm of the partition coefficient K is related linearly to the change in free energy of the transfer, ΔG_t .

$$\mu_{\text{wat}}^{\circ} - \mu_{\text{oct}}^{\circ} = R T \log K_{\text{O/W}} = \Delta G_t \quad (\text{eq. 12.10})$$

ΔG_t can be deconstructed into an enthalpic part and an entropic part:

$$\Delta G_t = \Delta H_t - T \Delta S_t \quad (\text{eq. 12.11})$$

at constant pressure and temperature. A negative ΔG_t value promotes the transfer towards water. The partition coefficient therefore represents the result of these two effects (ΔH_t and ΔS_t), knowledge of which can help to clarify certain factors intervening at the molecular level, such as **solvation**, **conformational changes** etc. The separate study of changes in ΔH_t and ΔS_t has not undergone very significant developments. However, a theoretical understanding of the phenomena associated with solubility requires knowledge of the factors affecting the changes in ΔG_t . The very strong correlation that has already been mentioned between the partition coefficient and solubility in water allows reasoning by analogy about the effects of water solubilisation.

For ideal solutions that can be simplified to solvent and solute molecules having the same size and form (spherical), it is possible to explain the factors influencing the different thermodynamic parameters of dissolution.

» Factors influencing ΔH_t

When a substance dissolves in water, the forces maintaining the cohesion of water molecules to each other are ruptured and these are replaced by interaction

forces between water and solute. Since the cohesive forces between water and organic solute are generally weak compared to those existing between water molecules, the **aqueous medium has a tendency to expel the organic molecules towards the exterior**. This phenomenon manifests as a **change in ΔH_t of water > 0** . However negative ΔH_t values are observed in many examples of solubilisation (benzene, toluene, xylene) due to the **hydrophobic effect**. There are actually two opposing phenomena occurring during a substance's solubilisation in water. The creation of a cavity in water by the solute (which leads to a rise in transfer enthalpy) is accompanied by a *hyperstructuring of water molecules* (hydrophobic effect) around the cavity (which diminishes the transfer enthalpy). The global result is that the change in ΔH_t is small, indeed negative, especially at low temperatures.

» **Factors influencing ΔS_t**

Entropy measures the tendency of a system to reach maximum disorder. Accounting for the fact that the molecular distribution of the solute is necessarily more disordered in solution than in the pure liquid state, **ΔS_t is necessarily positive and favours dissolution**. A precise study of the changes in ΔS_t requires knowledge of:

- › the form of the molecules involved,
- › the statistical nature of their mutual arrangements,
- › co-association and solvation effects.

In particular if there is association between solvent and solute inducing local, ordered structures, the n term will be smaller.

In general, study of the partition coefficient of a substance between solvents requires a comparison in each medium of the solvent-solvent and solvent-solute association forces (which influences the enthalpic term) as well as the respective molecular arrangements (entropic term). In studying the octanol-water system the insight obtained is limited most often to a qualitative understanding. Generally, solute and solvent molecules have neither the same size nor form, although each case should be individually examined. From an entropic point of view, the molecular dissimilarity between solvent and solute can be considered to correspond to an increase in disorder in the solution, which will increase the ΔS_t term.

12.5. MEASUREMENT AND ESTIMATION OF THE OCTANOL/WATER PARTITION COEFFICIENT

12.5.1. MEASUREMENT METHODS

Shake-flask method

To determine experimentally the partition coefficient $P_{O/W}$ of a compound, the simplest method consists of **distributing a known quantity of the substance between octanol and water** (mutually saturated beforehand) by shaking in a flask,

referred to as the **shake-flask method**. In order for the results obtained to be reproducible, the duration and intensity of shaking as well as the separation of phases (centrifugation at 2,000 rpm for 1 hour) are standardised. The concentrations used are of the order of 10^{-3} M and a quantitative analysis is generally carried out by spectrophotometry. In the case of weak acids and bases, it is the partition of the non-ionised form that is measured:

$$P = \frac{C_{\text{oct}}}{C_{\text{H}_2\text{O}}(1-\alpha)} \quad (\text{eq. 12.12})$$

with α being the degree of ionisation.

The partition is achieved between octanol and an aqueous buffer of fixed pH, which allows the deduction of α if the pK of the substance is known.

For acids:
$$\alpha = \frac{1}{1+10^{(\text{pK}_a - \text{pH})}} \quad (\text{eq. 12.13})$$

For bases:
$$\alpha = \frac{1}{1+10^{(\text{pH} - \text{pK}_a)}} \quad (\text{eq. 12.14})$$

In reality this method presents many practical disadvantages (long handling times, difficulties linked to the stability and/or the purity of the products). In order to alleviate these difficulties, chromatographic methods have been developed to determine the partition coefficient.

Method based on Reverse Phase Thin-Layer Chromatography (RPTLC)

This method is a simple and quick alternative to the shake-flask method. The partition coefficient is obtained by applying the following equation, tested for numerous chemical series:

$$\log P = \log a + Rm \quad (\text{eq. 12.15})$$

where a is a constant and Rm a function of the coefficient of migration in thin-layer chromatography (Rf).

$$Rm = \log \frac{1}{Rf - 1} \quad (\text{eq. 12.16})$$

Method based on High Pressure Liquid Chromatography (HPLC)

This more recent method has been used with success. The equation giving the partition coefficient is written as:

$$\log P = \log a + \log k' \quad (\text{eq. 12.17})$$

with

$$k' = \frac{t_R - t_0}{t_0} \quad (\text{eq. 12.18})$$

where t_R is the retention time of the substance, t_0 the elution time of the solvent and a a constant.

For most compounds, the chromatographic HPLC parameters can be obtained more easily and more precisely than the partition coefficients. In particular, chromatographic determinations are the only ones applicable in the case of highly lipophilic molecules.

12.5.2. PREDICTION METHODS

The interest in methods for the prediction of physicochemical parameters has already been underlined as they provide a significant gain in time, as syntheses, purification of the products and experimental measurement of $\log P$ can thus be avoided. Thanks to these methods the lipophilicity can be evaluated before the molecule is actually synthesised.

HANSCH method

FUJITA and HANSCH (1967) developed a method of constitutive and additive calculation of the partition coefficient, for which **the overall partition coefficient of a molecule is equal to the sum of the elementary contributions from each constitutive element of the structure** (example 12.1).

Example 12.1 - prediction of $\log P$ by the HANSCH method

Let π_x be the contribution of a substituent x fixed on a 'carrier' structure to the overall molecular $\log P$:

$$\pi_x = \log(P[R-X]) - \log(P[R-H])$$

Taking the example of substituents on a C_6H_6 core:

$$\pi_x = \log(P[C_6H_5-X]) - \log(P[C_6H_6])$$

$$\pi_x = \log \frac{P[C_6H_5-X]}{P[C_6H_6]}$$

Evaluation of the overall $\log P$ is therefore given by:

$$\log(P[C_6H_5-X]) = \pi_x + \log(P[C_6H_6])$$

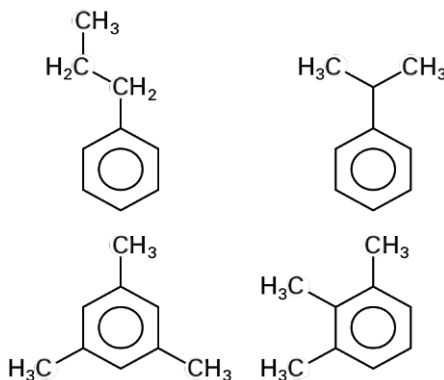


Fig. 12.1 - Example of linking substituents to a C_6H_6 core

The value of π_x is obtained after experimentally determining the respective partition coefficients of the compound possessing the relevant substituent and of the non-substituted homologous structure. However, when a first substituent is already present on an aromatic core (e.g. NH_2 , OH) the value of π_x varies appreciably depending on the nature of the first functional group. These differences reflect the model's sensitivity to the influence of mutual interactions between the two substituents. Several sets of π values are therefore necessary to take into account the different structural types encountered in an aromatic series (FUJITA and HANSCH, 1967).

The determination of these values has required considerable experimental work as well as an elaborate statistical analysis. Nevertheless it is of note that the π values arising from the aforementioned work must always be added to the $\log P$ value that is representative of a non-substituted structure. While this method is satisfactory for an homogenous chemical series, when the first terms have been prepared and their $\log P$ measured, it does not avoid recourse to experimentation (determination of the $\log P$ of the non-substituted base structure) in the case of original chemical series.

Stemming from the formalism of HANSCH's method the contribution of a hydrogen atom to the molecular lipophilicity is considered to be nil. The contributions to the overall lipophilicity of a substrate by the CH_3 , CH_2 , CH and C groups will be identical. For example, molecules of isopropyl-benzene and n-propyl benzene on the one hand and molecules of 1,3,5-trimethyl benzene and 1,2,3-trimethyl benzene, on the other, have the same $\log P$ when estimated by the HANSCH method. This leads to the use of numerous corrective terms, which often complicates the use of the π parameters.

Method of hydrophobic fragmental constants

An attractive objective in the domain of prediction is the evaluation *ex nihilo* of $\log P$ only from knowledge of the molecular structure. Such an approach involves the summation of the contribution to the partition coefficient of all the different component 'fragments' of a molecule. The procedure put forward by NYS and REKKER (REKKER, 1977) is described by the following equation:

$$\log P = \sum_{i=1}^n a_i f_i + \sum_{j=1}^m c_j \quad (\text{eq. 12.19})$$

where a_i is the number of identical fragments of rank i , f_i the contribution of the fragment i to the lipophilicity and c_j a correction factor. The different values for fragments are obtained from a very large number of molecules whose partition coefficients $P_{\text{O/W}}$ (statistical mean) have been measured. The c_j correction factor takes into account principally the proximity effects of polar groups. It is remarkable that these corrective terms are made up of a constant, estimated to be close to 0.29, and multiplied by a whole number. This constant has been referred to as the 'magic constant' after the first description of REKKER's methodology. Now with a revised value of 0.219, this constant must be applied a whole number of times (positive or negative) so as to take into account the listed interactions in the method.

Calculation *in silico* of the partition coefficient is thus made possible (VAN DE WATERBEEMD *et al.*, 1989), but a breakdown of the structure into fragments and the choice of the number k to be multiplied by the **magic constant** C_m are left to the experimenter's initiative (example 12.2).

Example 12.2 - prediction of log P by the method of NYS and REKKER

Let π_x be the contribution of a substituent x attached to a 'carrier' structure to the overall molecular log P :

N (aliphatic)	– 2.074
CONH ₂	– 2.011
Pyridinyl	+ 0.534
C ₁₅ H ₂₃	+ 6.342
�Cm	+ 0.219
TOTAL	+ 2.570

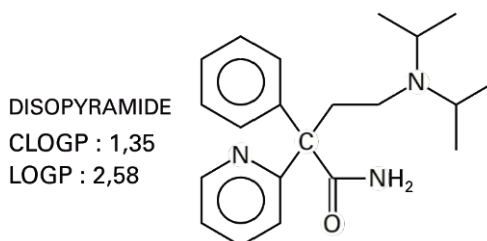


Fig. 12.2 - Example of disopyramide

Method of LEO and HANSCH

A method derived from that of HANSCH is based on a system analogous to that proposed by REKKER by elementary fragmentation. The principle of fragmentation is the following: first of all the **isolated carbons, IC**, that are not linked to heteroatoms by multiple bonds (CH₃–CH₃, CH₃OH, CH₂=CH₂) are defined. In the calculation an IC will always be considered to be hydrophobic. A hydrogen atom linked to an IC also constitutes a hydrophobic fragment (*ICH*). All of the covalently bonded atoms, which remain after selection of the IC and *ICH*, are polar atoms or groups: CN, OH, Cl etc. These polar fragments can be assigned a different numerical value depending on whether they are bonded to an sp² or sp³ carbon. Lastly, these polar fragments are split into different classes as a function of their capacity to form hydrogen bonds, depending on whether or not they comprise a hydroxyl. The correction factors take into account the structural characteristics, flexibility, branching and the electronic interactions between polar fragments.

Methods based on molecular connectivity

A molecule can be represented in the form of a molecular graph, in which only the atoms (nodes of the graph) and the bonds (edges or lines of the graph) feature. The use of **mathematical graph theory** in the context of molecular topology has led to the **construction of mono- or multidimensional indices** (numerical values or vectors) in order to attempt to represent the dependent molecular properties of the structure.

In the specific case of molecular lipophilicity, the first index developed was by RANDIC. This index (χ) is obtained by the summation over all the graph edges linking heavy atoms of the inverse square root of the product of the connectivities (δ) of the two atoms i and j involved.

$$\chi = \sum \frac{1}{\sqrt{\delta_i \delta_j}} \quad (\text{eq. 12.20})$$

The indices obtained are simple, easy to determine but they do not take into account unsaturation. For example, they are identical for benzene and cyclohexane.

Consequently, KIER *et al.* (1975) introduced the notion of valence index (χ^v) calculated in a similar way but for which the atomic connectivity (δ) is replaced by χ^v , the valence index between heavy atoms. For example, for a bond in benzene, $\delta^v = 2$; for C, $\delta^v = 3$. These indices show a strong correlation with the partition coefficient P and are often used to represent the properties of lipophilicity.

Based on molecular graphs, the **technique of autocorrelation vectors** put forward by BROTO *et al.* (1984) following the introduction of the concept of molecular autocorrelation, which we shall explain later, enables calculation of the partition coefficient from the atomic contributions.

12.5.3. RELATIONSHIP BETWEEN LIPOPHILICITY AND SOLVATION ENERGY: LSER

LSER (Linear Solvation Energy Relationships) is a fundamental approach attempting to rationalise solvent-solute interactions. Proposed by KAMLET *et al.* (1986), LSER is expressed by an equation having the general form:

$$\log(\text{properties}) = \Sigma \text{terms [steric, polarisability, hydrogen bonds]} \quad (\text{eq. 12.21})$$

For the properties of solubility and partition, the **steric term** is the molar volume. This molar volume represents the energy necessary to disrupt the cohesive forces between solvent molecules and to create a 'cavity' for the solute. The **polarisability term** is a measure of the dipole-dipole interactions induced between solvent and solute. The **hydrogen bonds term** represents the proton donor or acceptor power. These three groups of parameters must be orthogonal so as to minimise the covariance.

Only the first term corresponding to the molar volume can be calculated; the two others are measured, and this method does not have predictive power. WILSON and FAMINI (1991) suggested calculating the value of the terms relating to polarisability and to the hydrogen bonds. This approach is called **TLSER (Theoretical LSER)**. The polarisability term defines the ease with which the electron cloud can be distorted. The third term is obtained from quantum methods. Very significant correlations are obtained in several series of active principles between the molecular toxicity and the lipophilicity evaluated by these procedures.

12.5.4. INDIRECT ESTIMATION OF PARTITION COEFFICIENTS FROM VALUES CORRELATED WITH MOLECULAR LIPOPHILICITY

The solubilisation of a substance in a solvent depends on a very large number of parameters which govern the physics of a compound's solubilisation. Many physicochemical properties are in this respect highly correlated with lipophilicity, and hence they constitute an indirect means to estimate the lipophilicity.

» Aqueous solubility

The solubility of a substance in water S_w (solubility in mole/litre) is highly correlated with its partition coefficient $P_{O/W}$. One of the first relationships between these two values was established by HANSCH *et al.* (1968); based on 156 compounds it demonstrates the linearity of the equation relating $\log P$ (estimated by the HANSCH method) to S_w determined experimentally.

$$\log S_w = -1.34 \log P + 0.978 \quad (\text{eq. 12.22})$$

This relationship, which only applies to liquid solutes at room temperature, has subsequently been generalised and extended by YALKOWSKY *et al.* (1980).

» Temperatures of change of state

The boiling point of a liquid and the temperature of solidification (freezing) reflects the characteristics of auto-association of the molecules of which they are composed. Considering the predominance of polar interaction forces, a strong correlation has been observed between aqueous solubility and the temperature of the change of state.

» Parachor

Parachor is defined for liquid substances as follows:

$$\text{Parachor} = \frac{PM \cdot \gamma^{1/4}}{\rho_{\text{liquide}} - \rho_{\text{vapeur}}} \quad (\text{eq. 12.23})$$

with MW , the molecular weight, γ , the surface tension and ρ_{liquid} and ρ_{vapour} , the respective densities of the liquid and the vapour.

When the liquid density is much higher than the vapour density we obtain a simplified expression for parachor as a function of the molar volume V :

$$\text{Parachor} = V \cdot \gamma^{1/4} \quad (\text{eq. 12.24})$$

Parachor can be estimated by the group-contribution method. It is very highly correlated with the partition coefficient $P_{O/W}$ for a number of chemical series.

» Molar and molecular volume

We have previously mentioned the need for a solute during dissolution to create a cavity in the solvent. The work involved in this process influences the enthalpic term ΔH_i for the free energy of dissolution. The size of this cavity is related to the molar or molecular volume. Many authors have demonstrated the relationship between these values and solubility in diverse solvents. The **molar volume** can be readily calculated with the equation:

$$V = M / \rho \quad (\text{eq. 12.25})$$

where M is the molar mass of the solute and ρ its density.

The **molecular volume** can be deduced from the molecular geometry by common algorithms (CONOLLY, 1985; SHERAGA, 1987) based on a calculation of the total VAN DER WAALS volume and estimation of the intersections with the

atomic spheres. This calculation is currently available in almost all molecular modelling programs. It is however necessary to note that the best relationships between molecular volume and solubility are obtained by using an additional term featuring the molecular polarity.

» **Molecular surface**

The fact that the molecular volume alone does not represent the nature of the interactions between solute and solvent molecules and that these interactions are localised to the molecular surface has led researchers to **estimate molecular surface properties in order to understand better the effects involved in lipophilicity**. The validity of these relationships between molecular surface values and lipophilicity was demonstrated from the calculation of polar and apolar surfaces.

Calculation of molecular surfaces is only straightforward in the case of helium. For other cases it can be achieved from the external VAN DER WAALS surface or even from solvent-accessible surfaces. The relationships between molecular surfaces and lipophilicity are simple only when the molecules are small and rigid. For flexible molecules interacting with a solvent, conformational changes can intervene in such a way as to enable the best **stabilisation of the solvent-solute super-molecule**. These changes of conformation manifest in an overexposure of mobile polar groups to a polar solvent and the burial of these same groups in an apolar environment. These phenomena, which have mainly been studied and simulated in relation to proteins, are general and complex.

12.5.5. THREE-DIMENSIONAL APPROACH TO LIPOPHILICITY

The involvement of molecular conformation in solvation effects requires in many cases a **three-dimensional view of lipophilicity**. In parallel to the development of molecular potential quantities such as the **molecular electrostatic potential (MEP)** to explain reactivity phenomena, the concept of lipophilic potential was introduced by AUDRY *et al.* (1986). The **molecular lipophilic potential (MLP)** does not correspond directly to anything observable molecularly, it is determined empirically based on atomic lipophilic contributions and has the following analytic form:

$$\text{PLM}_x = K \sum \frac{f_i}{1+d_x} \quad (\text{eq. 12.26})$$

where MLP_x is the lipophilic potential at the point x ; K , a constant for converting to kcalorie/mole; f_i is the atomic contribution to lipophilicity of atom i , and d_x is the distance between the point x and the atom i .

Other analytic methods for calculating this potential, such as that of decreasing exponentials, were subsequently proposed.

The **use of MLP values** in the context of establishing **structure-activity relationships** is analogous to that of MEP. They allow in particular a **semi-quantitative evaluation of the influence of conformational modifications on lipophilicity**. They can be represented either as an isopotential surface (relative to a given energy) or serve to colour a molecular surface (for example, the solvent-accessible surface). It is important to note that although the molecular lipophilicity potential cannot be directly related to thermodynamic quantities, it remains **extremely useful during guided molecular fitting**.

12.6. SOLVENT SYSTEMS OTHER THAN OCTANOL/WATER

A complete estimation of molecular lipophilicity requires knowledge of the molecular behaviour in different solvents, as they are likely to cause the nature of the interactions to vary, and an exploration of molecular adaptability to the environment. The *octanol/water* system is widely used to measure the lipophilicity of the active principles. HANSCH considered this to be a representative model of biological systems: it represents essentially solvation effects and hydrophobic interactions. The validity of the model is generally demonstrated by a significant correlation observed between the biological properties of a substance and its partition coefficient P measured in this system. For example, the binding to blood plasma proteins by numerous xenobiotics is expressed by the following equation:

$$\log \frac{1}{C} = 0,75 (\pm 0,07) \cdot \log P_{O/E} + 2,30 (\pm 0,15) \quad (\text{eq. 12.27})$$

This equation was determined from 42 molecules (phenols, aromatic amines, alcohols), with C , the concentration necessary to form an equimolecular complex between the substance and bovine serum albumin. The slope of the straight line is less than 1, which means that desolvation of the solute is incomplete during binding to the protein.

The crossing of biological membranes is generally well represented by the octanol/water system. In certain instances, such as crossing the hematoencephalic (blood-brain) barrier, the octanol/water system is not representative and other solvent systems are required.

These systems can be broadly classified into four groups according to the nature of the organic phase:

- **inerts** - alkane/water (mainly cyclohexane, heptane) and aromatic/water (benzene, xylene, toluene),
- **amphiprotic** - octanol/water, pentanol/water, butanol/water, oleic alcohol/water,
- **proton donors** - CHCl_3 /water,
- **proton acceptors** - propylene glycol dipelargonate, butanone-2/water.

These systems or even combinations of these systems can give very good results (example 12.3).

Example 12.3 - lipophilicity of antagonists of histamine H_2 receptors situated in the central nervous system

The classical antagonists against H_2 receptors are very polar compounds that do not cross the blood-brain barrier (the barrier between the blood circulation and the central nervous system).

Very surprisingly, molecules having optimal lipophilicity according to HANSCH ($\log P_{OW} = 2$), prepared by YOUNG *et al.* (1988), also do not cross the blood-brain barrier.

A model to evaluate adequate lipophilic properties was put forward with the help of $\Delta \log P$, the difference between $\log P$ (octanol/water) and $\log P$ (cyclohexane/water). For six molecules (clonidine, mepyramine, imipramine and three H_2 receptor antagonists) penetration into the central nervous system and the partition between the two systems of measured solvents led to the development of the following equation:

$$\log \frac{C_{\text{cerveau}}}{C_{\text{sang}}} = -0,64 (\pm 0,169) \Delta \log P + 1,23 (\pm 0,56) \quad (\text{eq. 12.28})$$

This model suggests that penetration into the nervous system can be improved by reducing the capacity of the active principle to form hydrogen bonds. The first active molecule, zolantidine, was obtained using these lipophilicity properties. ▽

Of course, solvent systems other than octanol/water have not been hugely worked upon. In particular, it is difficult to find parametric databases allowing them to be estimated theoretically.

12.7. ELECTRONIC PARAMETERS

The **electronic distribution** around a molecular structure is responsible for establishing interactions between a receptor and its ligand and underlies all of the chemical properties. We all know that **the electron distribution forms within a volume englobing all of the nuclei** and that to discretise (meaning: to reduce continuous data to discrete, *i.e.* discontinuous, intervals,) a charge by relating it to a particular atom has no physical sense since charge is not a molecular observable. However many studies utilise **local charges** as a **parameter in QSAR equations** since they are very easily predicted theoretically using software for elementary quantum mechanics or even programs for the empirical attribution of charges like those often present in molecular modelling software.

As in the case of lipophilicity, we find global parameters and substituent parameters, the latter only being usable for the study of homogenous series.

12.7.1. THE HAMMETT PARAMETER, σ

It is well known that the strength of carboxylic acids varies as a function of the substituents connected to the carboxylate group. A large number of relationships have been established between the substituent groups of an aromatic series, and their reactivity. In several cases these relationships are expressed quantitatively and

are useful in the interpretation of mechanisms, for predicting rates and equilibria. The best known is the **HAMMET-BUKHARDT equation**, which relates rates to the equilibria of numerous reactions involving substituted phenyls. This parameter was initially determined in an aromatic series by studying the Ka values of differentially substituted benzoic acids. If Ka is the acidity constant of benzoic acid and Kax that of a benzoic acid substituted with the group X , the HAMMET equation is given by:

$$\log \frac{Kax}{Ka} = \sigma \quad (\text{eq. 12.29})$$

This equation is generalisable to other aromatic-acid series. Only the constant ρ changes, which is either higher or lower than 1 depending on the sensitivity of the series to the effects of the substituents. The σ constants are positive for electron donor substituents and negative for electron attractors.

$$\log \frac{Kx}{K} = \rho \quad (\text{eq. 12.30})$$

By extension, the **HAMMETT-BURKHARDT equation** takes into account electronic effects; k and k_0 are rate constants associated with a given reaction for a substituted or unsubstituted benzene ring. ρ thus takes into account the sensitivity of the reaction to electronic effects.

$$\log \frac{k}{k_0} = \rho \sigma \quad (\text{eq. 12.31})$$

Let us note the modifications made for cases where there is a delocalisation of charge between the reaction site and the substituent. For delocalisation of a negative charge, σ becomes σ^- , and of a positive charge, σ becomes σ^+ .

$$\log \frac{k}{k_0} = \rho [\sigma + r(\sigma^+ - \sigma^-)] \quad (\text{eq. 12.32})$$

12.7.2. SWAIN AND LUPTON PARAMETERS

Another approach to the treatment of field and resonance effects was proposed by **SWAIN and LUPTON**. It consists of separating the effects. The substitution constant expresses the sum of these two effects:

$$\sigma = fF + rR \quad (\text{eq. 12.33})$$

The method for determining the four parameters in place of a single one requires determination of a reaction series by regression. Beyond the historical parameters, there is not really a rule for selecting a representative value of the electronic molecular properties. The HOMO (Highest Occupied Molecular Orbitals) and LUMO (Lowest Unoccupied Molecular Orbitals) energies, the electronic densities around some atoms, the molecular electrostatic potential values in a point in space, or any other parameter can be used.

12.8. STERIC DESCRIPTORS

Historically, the first free-energy parameters being derived from the steric hindrance of substituents were those of TAFT, complementing those of HAMMET, and describing the influence of the steric hindrance of substituents in the basic hydrolysis reaction of aliphatic esters. The description of steric hindrance characteristics and a molecule's form has greatly evolved over the course of the years 1990-2000 with the advent of 3D drug design and the descriptors that are used today also take into account the molecular form (see chapters 11 and 13).

12.9. CONCLUSION

Faced with an avalanche of results arising from automated pharmacological screens, the relationship between the chemical structure of a ligand and its bioactivity is a challenge. We emphasise the fact that rather than the structure (the *S* in *QSAR*), it is really the physicochemical properties (the *P* in *QPAR*) that permit a description of the molecules of interest. Chapter 13 deals with this precise point. By adopting a variety of methods for the measurement and prediction of molecular properties that show a correlation with pharmacological activity, much light has been shed on **lipophilicity**, estimated with the partition coefficient, the popular **log P**. In parallel to the developments in molecular modelling, we are witnessing today a veritable upheaval in the concepts associated with molecular lipophilicity. Lipophilicity is no longer considered to be a statistical characteristic but a dynamic property. If its flexibility allows, a molecule can adopt different conformations in different solvents, exposing to the solvent a maximum number of polar groups in an aqueous environment, yet conversely burying these groups when interacting with a lipid solvent. The perspectives in the field of *QSAR* thus veer towards ever more dynamic descriptions of molecules and take into account ever more diverse molecular properties (GRASSY *et al.*, 1998). Looking for correlations between the complex properties of small molecules and their bioactivity now demands sophisticated computational methods and moves towards the techniques in artificial intelligence (chapter 15).

12.10. REFERENCES

- ANFINSEN C., REDFIELD R. (1956) Protein structure in relation to function and biosynthesis. *Adv. Protein Chem.* **48**: 1-100
- AUDRY E., DUBOST J.P., COLLETER J.C., DALLET P. (1986) Le potentiel de lipophilie moléculaire, nouvelle méthode d'approche des relations structure-activité. *Eur. J. Med. Chem.* **21**: 71-72
- BROTO P., MOREAU G., VANDYCKE C. (1984) Molecular structures: perception, autocorrelation descriptor and SAR studies. *Eur. J. Med. Chem.* **19**: 71-78

- CONNOLLY M.L. (1985) Computation of molecular volume. *J. Am. Chem. Soc.* **107**: 1118-1124
- FUJITA T., HANSCH C. (1967) Analysis of the structure-activity relationship of the sulfonamide drugs using substituent constants. *J. Med. Chem.* **10**: 991-1000
- GRASSY G., CALAS B., YASRI A., LAHANA R., WOO J., IYER S., KACZOREK M., FLOCH R., BUELOW R. (1998) Computer-assisted rational design of immunosuppressive compounds. *Nat. Biotechnol.* **16**: 748-752
- HANSCH C., LIEN E.J., HELMER F. (1968) Structure-activity correlations in the metabolism of drugs. *Arch. Biochem. Biophys.* **128**: 319-330
- HANSCH C., MALONEY P.P., FUJITA T., MUIR R.M. (1962) Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature* **194**: 178-180
- HANSCH C., MUIR R.M. (1961) Electronic effect of substituents on the activity of phenoxyacetic acids. In *Plant growth regulation*, Iowa State University Press: 431
- HANSCH C., MUIR R.M. (1951) Relationship between structure and activity in the substituted benzoic and phenoxyacetic acids. *Plant Physiol.* **26**: 369-374
- KAMLET M., DOHERTY R., ABBOUD, J.L., ABRAHAM M., TAFT R. (1986) Linear solvation energy relationships: 36. molecular properties governing solubilities of organic nonelectrolytes in water. *J. Pharm. Sci.* **75**: 338-349
- KIER L.B., HALL L.H., MURRAY W.J., RANDIC M. (1975) Molecular connectivity. I: Relationship to nonspecific local anesthesia. *J. Pharm. Sci.* **64**: 1971-1974
- MEYER H.H. (1899) Theorie der Alkoholnarkose. *Arch. Exp. Pathol. Pharmacol.* **42**: 109-118
- MEYER H.H. (1901) Zur Theorie der Alkoholnarkose. III. Der Einfluss wechselnder Temperatur auf Wirkungs-stärke and Teilungs Koeffizient der Narkotika. *Arch. Exp. Pathol. Pharmacol.* **154**: 338-346
- OVERTON C.E. (1901) *Studien  ber Narkose, zugleich ein Beitrag zur allgemeinen Pharmakologie*. Fisher, Jena, Allemagne
- REKKER R.F. (1977) *The hydrophobic fragment constant*. Elsevier, New York, USA
- VAN DE WATERBEEMD H., TESTA B., CARRUPT P.A., TAYAR N. (1989) Multivariate data analyses of QSAR parameters. *Prog. Clin. Biol. Res.* **291**: 123-126
- WILSON L.Y., FAMINI G.R. (1991) Using theoretical descriptors in quantitative structure-activity relationships: some toxicological indices. *J. Med. Chem.* **34**: 1668-1674
- YALKOWSKY S.H., VALVANI S.C. (1980) Solubility and partitioning. I: solubility of nonelectrolytes in water. *J. Pharm. Sci.* **69**: 912-922
- YOUNG R.C., MITCHELL R.C., BROWN T.H., GANELLIN C.R., GRIFFITHS R., JONES M., RANA K.K., SAUNDERS D., SMITH I.R., SORE N.E. (1988) Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H₂ receptor histamine antagonists. *J. Med. Chem.* **31**: 656-671

Chapter 13

ANNOTATION AND CLASSIFICATION OF CHEMICAL SPACE IN CHEMOGENOMICS

Dragos HORVATH

13.1. INTRODUCTION

How do we recognise a drug? Can one, by looking at any chemical formula, declare: *“it is out of the question that this molecule could act as a drug – it is simply not drug-like!”*? Is this a question of intuition or does it lend itself to a mathematical analysis? Using which tools? Lastly, what can we expect from modelling the biological activity of molecules? The complexity of the living world for the moment evades all attempts of a reductionist analysis from the underlying physicochemical processes. However, the ‘blind’ search for drugs, expecting to come across by chance a molecule that illicitly the ‘right’ effect in vivo – too expensive, too slow and ethically questionable as it involves many animal tests – is nowadays no longer an option. Chemoinformatics, a recent discipline developed with the aim of rationalising the drug discovery process, proposes a ‘middle way’ between the impossible modelling from first principles and the blind screening of compound libraries (XU and HAGLER, 2002). It uses the maximum amount of information obtained experimentally in order to find possible correlations between the structures of tested molecules and their success in biological tests. Such empirical correlations can then be used successfully to guide the choice of novel compounds to synthesise and test, while ensuring a better success rate compared to a random choice.

13.2. FROM THE MEDICINAL CHEMIST’S INTUITION TO A FORMAL TREATMENT OF STRUCTURAL INFORMATION

With the development of medicinal chemistry as an entirely separate discipline, it has been recognised that drugs are organic molecules, which, despite their extreme diversity, show a series of common traits differentiating themselves from other categories of compounds. This is not surprising in regard of the fact that, aside from the specificity of each molecule for its own biological target, its success as a

drug depends also on its capacity to reach the target within a living being (see chapters 8, 9 and 12). Now, the tests that a drug candidate must pass in order to penetrate into an organism's cell are essentially the same: the passage in solution into the intestine, intestinal absorption, resistance to enzymes responsible for metabolising xenobiotics (compounds foreign to the organism, toxic in low doses), and finally maintenance in the blood while avoiding excretion. These are therefore the constraints of good **pharmacokinetics** which, in spite of their vast structural diversity, force drugs to share a significant number of common physicochemical traits. Better still, the physicochemical reasons for a ligand's affinity for its receptor are universal: the bulk of free energy of ligand binding within an active site is contributed by hydrophobic contacts leading to a minimisation of the total hydrophobic surface exposed to solvent. A drug candidate supposed to interact reversibly with its target must necessarily be capable of establishing these contacts. We can easily understand why the flexibility of compounds should also play a role in defining a drug's characteristics, as the adoption of the bioactive conformation fixed in the active site will be more unfavourable (entropically) for more flexible molecules. More prosaically, economic criteria also feature in terms of limiting the structural complexity of viable drugs, since the idea of developing industrial synthesis procedures for molecules with around ten asymmetric centres is only ever met with little enthusiasm – here is a major source of differences between a synthetic drug and a natural bioactive product extracted from a living organism.

Nevertheless, it is not easy to identify a consistent series of rigorous *structural* rules listing the common traits hidden in the great diversity of drugs. A summarised **drug description** might define it as an **organic molecule based on a rigid skeleton, adorned with well-separated hydrophobic groups** (to avoid the intramolecular 'hydrophobic collapse') **and a polar head ensuring solubility** (POULAIN *et al.*, 2001). Estimating the pertinence of a structure as a drug candidate was for a long time a matter of *savoir-faire* and of the medicinal chemist's 'flair', whose occasional lapses of intuition greatly inflated the cost of pharmaceutical research, due to the failures of very expensive clinical tests with insoluble molecules that did not cross into the blood or were metabolised immediately afterwards, without reaching their target. The rational analysis of these problems was therefore a favourite subject in modern medicinal chemistry, notably with the emergence of the **concept of privileged structures** (HORTON *et al.*, 2003) based on the observation that certain organic fragments appear statistically more often in drugs than in other organic molecules.

However, the inventory of fragments has its limits as an analysis tool for the universe of potential drugs. Firstly, how do we choose the fragments to include? What size must they be in order for the analysis to be meaningful? Furthermore, different functional groups can very well interact similarly with the active site, and thus be interchangeable without a loss of activity. Would it therefore be justified to count them in the statistics as independent entities?

These questions led to the introduction of a fundamental concept in medicinal chemistry, that of **pharmacophore**, which defines the elements necessary for a molecule to show a particular bioactivity in terms of the **spatial distribution of pharmacophore points**. These points specify not the exact chemical identity of the fragment expected at this position, but its physicochemical nature: *hydrophobicity*, *aromaticity*, hydrogen bond *acceptor* or *donor*, *positive* or *negative* charge. The most simple example (and the most famous) of an analysis of the nature of drugs with reference to the pharmacophore nature of chemical groups is that of LIPINSKI (LIPINSKI *et al.*, 1997). This analysis demonstrated that 90% of the drugs known to date respect certain limits in terms of size, lipophilicity and the number of hydrogen bond acceptors (< 10) and donors (< 5) (see chapter 8). Of course, pharmacophore models also have their disadvantages, because the contrived classification of all organic functional groups according to the few pharmacophore types listed ends up sooner or later with inconsistencies.

The general idea arising from the previous paragraph is therefore the necessity of finding a means to extract information quantitatively from a molecular structure, and to convert this to a form that lends itself readily to statistical analysis. Any quantitative aspect in relation to a molecule's structure, be it the number of phenyl groups, the number of positive charges or the dipole moment obtained by quantum calculation, as well as every experimentally measured value (the refraction index, for example), can in principle be used as a **molecular descriptor**. A molecule M will thus be encoded as a series of such descriptors, which we shall generically refer to as $D_i(M)$, with $i = 1, N$ (the total number of descriptors chosen from the near-infinite available options). The design of relevant descriptors and the choice of the optimal descriptors for resolving a given problem are central questions in the science (or art?) of chemoinformatics. The 'translation' of a chemical structure into a series of descriptors amounts to the introduction of an N -dimensional **structural space**, where each axis i corresponds to a descriptor D_i and in which each molecule M corresponds to a point having the coordinates $D_1(M)$, $D_2(M)$... $D_N(M)$. The problem of dependence on the biological properties of a compound's structure can thus be reformulated in a way more suited to a mathematical treatment: "*expressing these properties as a function of the location of a molecule in structural space*".

The most commonly used molecular descriptors are presented in chapter 11.

Typically the following types of descriptor are distinguished:

- 1D, calculatable from the molecular formula, without needing to know connectivity tables,
- 2D or topological, utilising solely the information contained in molecular graphs (atom types and connectivity, plus bond order), and finally,
- 3D, including in addition structural information (interatomic distances, solvent-accessible surfaces, intensity of fields generated by atoms etc.).

Another classification of descriptors is based on the manner in which the atom type is encoded, in terms of which one may distinguish between:

- › graph-theoretical indices that completely ignore the nature of the atoms,
- › fragment-based, counts of predefined organic fragments in the molecule,
- › pharmacophore pattern counts, which note only the pharmacophore nature of the groups, while ignoring their chemical identity (see example 13.1 below),
- › implicit terms, in which the atoms are represented by means of their implicit calculatable properties: charge, VAN DER WAALS radius, electronegativity.

13.3. MAPPING STRUCTURAL SPACE: PREDICTIVE MODELS

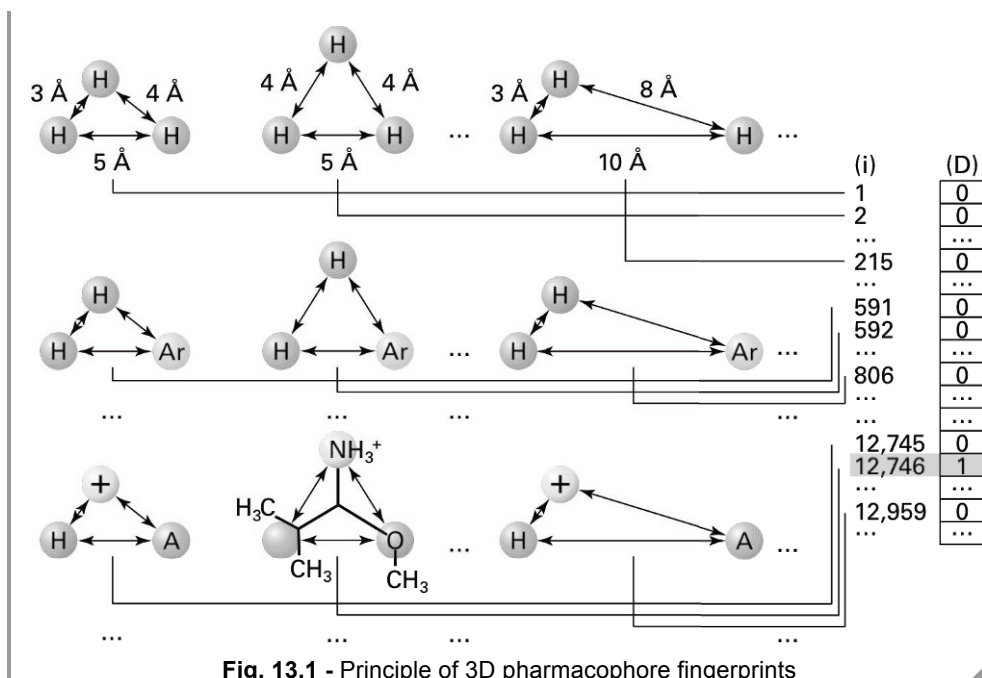
13.3.1. MAPPING STRUCTURAL SPACE

Once the **structural space** has been chosen, an analysis of the biological properties of molecules becomes a problem of space **mapping**, in a quest to find zones rich in compounds having the desired properties, while assuming, quite obviously, that such desired zones exist (if this is not the case, *i.e.* if the active compounds are found to be scattered uniformly throughout space, then the choice of descriptors used must be revised!). Undertaking this mapping demands, of course, a preliminary effort to explore this space: firstly, the property studied must be visited and then measured for a minimal number of points (molecules) in this space before being able to suggest a reasonable hypothesis about what might be discovered in the points that have not yet been reached. Mathematically speaking, a **map** is nothing other than a **function** (e.g. altitude in metres, a function of the latitude and longitude of a point on the globe). In this case, $P^{calc} = f[D_1(M), D_2(M), \dots, D_N(M)]$ will be the **predictive model** (QSAR, Quantitative Structure-Activity Relationship; see chapter 12) yielding an **estimation of a molecule's property occupying the point M in structural space**.

A predictive model can in principle be a non-linear function of arbitrary complexity, to be calibrated using previously visited points in space (molecules of known property, $P(M)$), such that $P^{calc}(M)$ is as close as possible to the true value $P(M)$ for each molecule M in this **calibration set**.

Example 13.1 - the 3D pharmacophore fingerprint

3D pharmacophore fingerprints (PICKET *et al.*, 1996) are binary 3D descriptors: each element D_i corresponds to a possible location of three pharmacophore centres, and this for each conceivable combination (H, H, H) , (H, H, Ar) , ..., (H, Ar, A) , ..., (H, A, D) , ..., $(H, A, -)$, ..., $(+, +, +)$ where H = hydrophobic, Ar = aromatic, A = hydrogen acceptor, D = hydrogen donor, $-$ = anion and $+$ = cation. For each triplet of properties a series of triangles of a size compatible with a drug molecule are indexed, and each of these several tens of thousands (!) of triangles will have its 'bit' i (0 or 1) in the fingerprint. All of the triangles i which are represented in a stable conformation of a molecule are thus indicated by $D_i = 1$, whereas for all the others $D_i = 0$. The binary vector $D(M)$ therefore characterises the pharmacophore profile of a molecule M , by listing the pharmacophore triangles that they contain.



In reality, the model's complexity will be initially limited by the size and quality of the available data for the calibration, but also by the impossibility of exploring *all* of the functional forms imaginable in order to find an absolutely perfect dependence between descriptors and properties. Certain categories of model have been developed as a priority (linear and non-linear approaches: neuronal networks, partition trees, and neighbourhood models) and will be discussed in more detail.

13.3.2. NEIGHBOURHOOD (SIMILARITY) MODELS

This is the most intuitive approach as it represents the mathematical reformulation of the classic **similarity principle** in medicinal chemistry, which states that similar structures will have similar properties (see chapter 11). The idea of considering the (Euclidian) **distance** between two molecules M and m in structural space as a **measure (metric) of similarity** feels natural:

$$d(M, m) = \sqrt{\sum_{i=1}^n [D_i(m) - D_i(M)]^2} \quad (\text{eq. 13.1})$$

However, there is no particular reason to presume that structural space is Euclidian rather than Riemannian or obeying other metrics routinely used despite their apparent exoticism (WILLETT *et al.*, 1998). Indeed, the only true criterion for the validity of a dissimilarity metric is good **neighbourhood behaviour**; that is, statistically speaking a majority of pairs of *neighbouring* molecules (m , M), according to a valid metric of structural space, will also be neighbouring with respect to their

experimental properties (PATTERSON *et al.*, 1996). The dissimilarity scores can include *adjustable parameters* for calibration in order to maximise the correct neighbourhood behaviour with respect to a specific property or a complete profile of properties (HORVATH and JEANDENANS, 2003).

Example 13.2 - the importance of the normalisation of axes (descriptors) of structural space for calculating dissimilarity

Sometimes, the different D_i defining structural space may include variables of very different magnitude. Taking the example of a two-dimensional structural space with the molecular mass (MW) as a descriptor D_1 and the partition coefficient of octanol/water, $\log P$, as D_2 , and three molecules:

$$M(MW = 250, \log P = -2)$$

$$N(MW = 300, \log P = -2)$$

$$P(MW = 250, \log P = +4)$$

Calculating the distances (dissimilarities) according to EUCLID:

$$d^2(M, N) = (250 - 300)^2 + (-2 + 2)^2 = 2500$$

$$d^2(M, P) = (250 - 250)^2 + (-2 - 4)^2 = 36$$

Therefore, M seems to be much closer to P than to N ! However, P is a very hydrophobic species, whereas M and N are hydrophilic. Furthermore, the difference in mass between M and N is after all not very large, knowing that the variance of this parameter among 'drug-like' molecules is of the order of 100 DALTONS! The artefact comes from the fact of having 'mixed apples and pears' in the similarity score, as the two descriptors are not directly comparable.

In order to bring all of the descriptors onto a common scale, each D_j must undergo a re-centering with respect to its mean $\langle D_j \rangle$ and **normalisation** with respect to its variance $\text{Var}(D_j)$:

$$D_j^{\text{norm}} = (D_j - \langle D_j \rangle) / \text{Var}(D_j)$$

Note - The variance $\text{Var}(D) = [\langle D^2 \rangle - \langle D \rangle^2]^{1/2}$ cannot be nil except if D is constant for all molecules and, due to this, useless. The calculation of means and variances used for normalisation should be performed from as wide and diverse a set of compounds possible having properties compatible with bioactivity.

The distances calculated in normalised space integrate well the fact that the difference observed for $\log P$ is more significant (with respect to the variance of $\log P$ of 'drug-like' molecules) than the mass difference in view of the typical fluctuation of molecular weights! ▲

If M is a **reference compound** of known activity and $d(m, M)$ a valid metric of structural space, the subset of neighbours m near to M , having $d(M, m)$ less than a **dissimilarity threshold** d^{lim} , is supposed to be richer in bioactive compounds than any other subset of the same size, including randomly chosen molecules m' . This is the principle of **virtual screening** by similarity of a database of molecular structures. Now, the medicinal chemist knows very well how to recognise similar molecules – of similar connectivity, belonging to the same **chemotype**. Therefore, the main purpose of an algorithm for similarity screening is (aside from the high speed of automation) to prove '**scaffold hopping**': **hidden similarity relationships, less well spotted at first glance by a chemist, but leading to a definite (and unexpected!) similarity in the properties**. This is the case with metrics for pharmacophore similarity (HORVATH, 2001a), illustrating the analogy to pharmacophore motifs hidden within chemical structures – an analogy that can be illustrated

intuitively by means of a molecular fitting procedure (HORVATH, 2001b), which can reveal functional complementarity of two apparently different molecules (fig. 13.2). The discovery of these alternative **hits** is much valued in medicinal chemistry since, while acting on the same target, they may be complementary in other respects (pharmacokinetics, synthetic feasibility, patentability etc.), which offers an alternative way out to the research programme if the development of one compound then fails in one of these parameters.

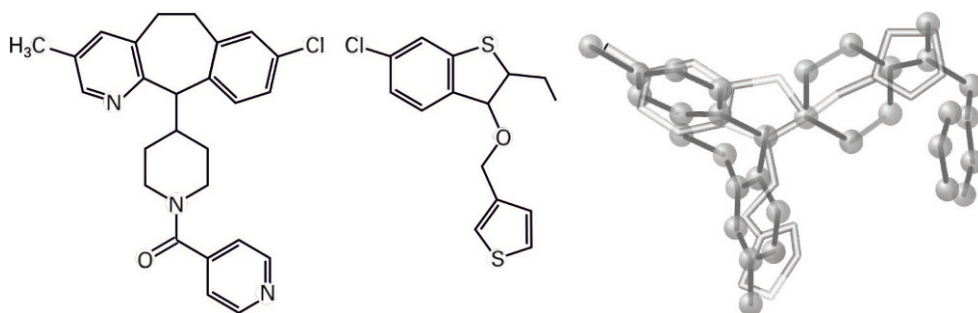


Fig. 13.2 - Scaffold hopping: Two molecules with different connectivities that 'hide' a common pharmacophore motif, clearly evidenced by the superimposed model. The compounds depicted are two farnesyl transferase inhibitors.

It is also possible to use neighbourhood models as quantitative *predictive* tools (GOLBRAIKH *et al.*, 2003), by going beyond simple virtual screening, which merely selects a set of 'supposedly active' molecules but gives no estimation of the expected activity level. On the contrary, whereas the latter compares a series of virtual molecules to one active reference compound, in the neighbourhood-model approach each compound to be predicted is compared to a whole set of experimentally characterised reference compounds. This approach selects the closest reference compounds around the compound to be predicted and extrapolates the property of this as an *average* of these reference neighbours' properties, by taking into account where necessary for each neighbour, a weighting that is inversely proportional to its distance with respect to the point to be predicted (ROLAND *et al.*, 2004).

The complementary application of neighbourhood models is the sampling of large chemical libraries allowing the choice of a representative subset of n dissimilar compounds from $N \gg n$ total molecules, by including representatives of each 'family' of bioactive compounds present. The **core** will therefore be designed in such a way that no pair of molecules m and M retained will have a dissimilarity less than a threshold $d(m, M) > d^{lim}$. This application essentially utilises fragment descriptors with the TANIMOTO metric (chapter 11; WILLETT *et al.*, 1998) and will therefore not prevent the simultaneous choosing of two molecules having the same pharmacophore motif hidden within two different skeletons. Although both run the risk therefore of hitting the same targets, this will not be perceived as a redundancy (see the previous paragraph).

13.3.3. LINEAR AND NON-LINEAR EMPIRICAL MODELS

Neighbourhood models are conceptually very close to the idea of structural space mapping: the active and inactive compounds already known serve as markers of the ‘summits’ and ‘valleys’ of bioactivity in structural space, and so permit the discovery of other active molecules in the vicinity of the summits already known (which will not prevent original discoveries from sometimes being made – see the paragraph on **virtual screening**). On the contrary, these approaches cannot be relied upon to predict the existence of a previously completely unknown summit (by screening relative to a reference pharmacophore completely novel *chemotypes* can be discovered, but the existence of other original *pharmacophore motifs* will never be deduced). In principle, these models do not formally deserve the label *predictive* in a scientific sense for which the *fundamental laws* of behaviour (also observations) are supposed to supply a complete and unambiguous description of this system, and in particular of yet unexplored regions within the space being studied. A good example is the deduction of the existence and properties of planets outside of the solar system from perturbations in the trajectory of visible celestial bodies, employing the law of gravity to model the system. An ideal procedure would consist therefore of postulating the existence of *all* the other physically possible summits of bioactivity in structural space from the available observations (supposing that these are sufficient, *i.e.* that there is a single model compatible with each of these simultaneously). Nevertheless, there is no simple analytical expression for the ‘laws’ that dictate the affinity of a molecule. To continue with the analogy, the developer of a structure-activity model finds him- or herself in the position of an astronomer constrained to deduce simultaneously the position and the mass of an unknown planet *and* the law of gravity. He or she should thus test *every imaginable function of the distance and masses of interacting bodies* (provided that he has had the intuition to presume that gravity is indeed about mass and distance, and not about electric charge!) in order to find that only the product of the masses divided by the square of the distance gives results compatible with the observations and correctly predicts the position of the new planet.

The construction of a predictive model (structure-activity relationship) starts with the (more or less arbitrary) **choice of the functional form of the law meant to relate activity to the molecular descriptors** for any molecule M :

$$P(M) = f[D_1(M), D_2(M), \dots, D_n(M)]$$

As no hypothesis can be ruled out, one may as well start with the simplest: a linear dependence (HANSCH and LEO, 1995):

$$P^{calc}(M) = f[D_1(M), D_2(M), \dots, D_n(M)] = \sum_{i=1}^n c_i D_i(M) + c_0 \quad (\text{eq. 13.2})$$

If this is a sensible choice, then the coefficients c_i can be found by multilinear regression, by imposing that the mean quadratic error between the calculated properties $P^{calc}(M)$ and the experimental values $P(M)$ of the already known molecules (the

training set) be minimal. Although the coefficients of the descriptors, which do not influence a given property, should be ‘spontaneously’ set to zero by the calibration procedure, it is sometimes necessary to *preselect* a subset of descriptors to enter into the model, the number of which should stay well below (at least by a factor of 5) the number of examples of molecules available for the calibration. This can be achieved in a deterministic or a stochastic manner. Other approaches (e.g. **Partial Least Squares** or **PLS**; WOLD, 1985) do not require preselection.

Note well: obtaining a regression equation with a good correlation coefficient is a necessary condition, but by no means sufficient. Such correlations can sometimes be obtained even between columns of random numbers, above all when the number of variables (n) is not greatly lower than the number of observations (training molecules). It is therefore advisable to check the robustness of the equation calibrated by **cross validation**. For cross validation, one or several molecules are eliminated in turn from the training set, the equation is derived again and then properties of the molecules extracted can be predicted according to this equation. Another really useful test is that of ‘randomisation’ whereby one interchanges randomly the properties of the molecules, in trying to explain the property of M based on the descriptors of another M' . In spite of such validation, it is necessary still to bear in mind that the model risks being a mere artefact (GOLBRAIKH and TROPSHA, 2002).

Good QSAR practice demands that a fraction of the available molecules is kept by for *validation*, by predicting their properties with the model obtained without any knowledge of these compounds; again a necessary step, but nevertheless still insufficient, because statistical artefacts cannot be excluded. If the test set includes molecules that are structurally close to those selected for the calibration, a good result risks being rather a consequence of this similarity. Besides, it is evident that a model cannot ‘learn’ what the training set cannot ‘teach’ it. For example, whereas the total charge of a molecule is an important descriptor for predicting the lipophilicity coefficient (partition in water/octanol) $\log P$, if the training set does not include any charged molecules, it is impossible to estimate by regression the weighting c_i associated to a column D_i filled with zeros. A training set should ideally cover *all* categories of possible structures. In the opposite case, the domain of model validity will be confined to the surroundings of the structural island used in the calibration (HORVATH *et al.*, 2005).

Lastly, if all attempts at linear modelling fail, the search can be extended to other functions $f[D_1(M), D_2(M), \dots, D_n(M)]$. Faced with this infinite choice, a bit of good physicochemical sense is sometimes beneficial (example 13.3).

Example 13.3 - a (nearly) non-linear model of apparent permeability

The passage of a drug from the intestine into the blood takes place via the intestinal wall, which is, schematically speaking, a (double) lipid membrane. One can therefore hypothesise that the lipophilicity of a compound will be a parameter having a significant impact on the rate of passage across the transmembrane. We shall include therefore the $\log P$ (chapter 12) as a potential descriptor in the prediction equation for the logarithm of a drug's

transmembrane flux ($\log Perm$). Nevertheless, the permeability does not depend in a linear way on lipophilicity: too hydrophilic a compound ($\log P \ll 0$) will never traverse the lipid barrier, whereas too hydrophobic a molecule ($\log P \gg 0$) will insert itself into the membrane and yet not cross into the blood either. This qualitative analysis allows us to establish an empirical working hypothesis stating a *parabolic* dependence of $\log Perm$ as a function of $\log P$. While minimal at the extremes of lipophilicity, the permeability will be optimal for compounds of average lipophilicity (HANSCH *et al.*, 2004).

$$\log Perm = a \cdot \log P - b \cdot \log P^2 + \text{other descriptors} \quad (\text{eq. 13.3})$$

This in fact becomes a linear model when considering the square of the partition coefficient as a new independent descriptor. Other descriptors will be necessary to take into account other phenomena that come into play from the moment of intestinal absorption – notably active transport or efflux by pumps within cells of the intestine. ▀

In the total absence of any idea about the nature of the non-linear function, neuronal networks are employed (chapter 15; GINI *et al.*, 1999). Simulating the synapses between neurons, these algorithms are capable of ‘mimicking’ a vast range of non-linear responses. However, they have the disadvantage of not being interpretable and of being very sensitive to over-training artefacts. It is not possible to list here all of the non-linear modelling techniques that have found an application in the search for structure-activity relationships, statistical methods are often imported from other fields making use of **data-mining** techniques by decision trees (chapter 15; BREIMAN *et al.*, 1984), a much-used tool in economics for risk management.

13.4. EMPIRICAL FILTERING OF DRUG CANDIDATES

To be a good drug, a compound M must simultaneously satisfy a whole **profile** of constraints with respect to numerous properties P_k . For example, an activity on the principal target (P_1) will have to be higher than a threshold p_1 , while the affinities ($P_2, P_3, \dots P_n$) for other receptors and enzymes present within targeted cells must be lower than other thresholds p_n . Furthermore, the pharmacokinetic properties ($P_{n+1}, P_{n+2}, \dots$ corresponding to solubility, permeability, metabolic stability etc.) must also fall within well-defined limits. Ideally, a medicinal chemist should have a range of models P_k^{calc} for each property P_k in question. He or she can thus visualise the zones of structural space that satisfy simultaneously all of the constraints imposed with respect to the properties, and synthesise in a targeted way only those molecules within these zones. This is, of course, a far too optimistic vision given that beyond the difficulties of obtaining such reliable predictive tools, modern biology is not yet capable of establishing an exhaustive list of all the targets that play a key role in triggering a desired response in vivo: the profile of properties P_k intended for the drug developer will therefore naturally be incomplete and imprecise.

Confronted by these formidable difficulties, one current and even more empirical line of thought leans towards the definition of the **drug-likeness** of organic molecules (AJAY *et al.*, 1998; SADOWSKY and KUBINYI, 1998). This is based on the

hypothesis that the zones of structural space to be explored as a priority are the zones already populated by known drugs, provided that one finds a choice of molecular descriptors defining a space in which these drugs are actually grouped in a consistent manner and being rather 'segregated' from other organic molecules. Such statistical studies sometimes prove to be very promising, especially when the **drug-likeness** criterion is defined based on pharmacokinetic properties (OPREA and GOTTFRIES, 2001). Nevertheless, it is extremely risky to draw conclusions too hastily about the general nature of a drug from examples of molecules used in the pharmacopeia of today, due to the unequal distribution of representatives from different therapeutic classes. For a number of novel **therapeutic orphan targets** (those without a known natural ligand) there is quite clearly no example available of an active drug.

13.5. CONCLUSION

The so-called rational approach, relying on predictive models to accelerate and to guide research into novel drugs should perhaps be entitled, more modestly, the less random approach. This is not pejorative, as every approach permitting a reduction, albeit by maybe only a few percent, of the colossal losses through investment into failed drug candidates, will provide a net competitive advantage to the pharmaceutical industry. The concept of structural space offers a context well suited to rationalising the research for novel compounds by exploiting, with statistical tools, the information collected during the research process in order to draw up a local **map of the structural regions visited onto which all of the screened molecules are projected**. This map can sometimes prove to be very useful in guiding the next steps of research by representing simply a systematisation of the structure-activity relationships, which a chemist will have seen without the aid of modelling tools, or by revealing aspects that are difficult for the human brain to grasp.

13.6. REFERENCES

- AJAY A., WALTERS W.P., MURCKO M.A. (1998) Can we learn to distinguish between drug-like and non-drug-like molecules? *J. Med. Chem.* **41**: 3314-3324
- GINI G., LORENZINI H., BENFENATI E., GRASSO P., BRUSCHI M. (1999) Predictive carcinogenicity: a model for aromatic compounds, with nitrogen-containing substituents, based on molecular descriptors using an artificial neural network. *J. Chem. Inf. Comput. Sci.* **39**: 1076-1080
- BREIMAN L., FRIEDMAN J.H., OHLSEN R.A., STONE C.J. (1984) *Classification and Regression Trees*. Wadsworth, New York, U.S.A.
- GOLBRAIKH A., TROPSHA A. (2002) Beware of q2! *J. Mol. Graph. Model.* **20**: 269-276

- GOLBRAIKH A., SHEN M., XIAO Z., XIAO Y.D., LEE K.H., TROPSHA A. (2003) Rational selection of training and test sets for the development of validated QSAR models. *J. Comput. Aided Mol. Des.* **17**: 241-253
- HANSCH C., LEO A. (1995) *Exploring QSAR Fundamentals and applications in chemistry and biology*. American Chemical Society, Washington DC, U.S.A.
- HANSCH C., LEO A., MEKAPATI S.B., KURUP A. (2004) QSAR and ADME. *Bioorg. Med. Chem.* **12**: 3391-3400
- HORTON D.A., BOURNE G.T., SMYTHE M.L. (2003) The combinatorial synthesis of bicyclic privileged structures or privileged substructures. *Chem. Rev.* **103**: 893-930
- HORVATH D. (2001a) High throughput conformational sampling and fuzzy similarity metrics: a novel approach to similarity searching and focused combinatorial library design and its role in the drug discovery laboratory. In *Combinatorial library design and evaluation: principles, software tools and applications* (GHOSE A., VISWANADHAN V. Eds) Marcel Dekker, Inc., New York: 429-472
- HORVATH D. (2001b) ComPharm – automated comparative analysis of pharmacophoric patterns and derived QSAR approaches, novel tools. In *High throughput drug discovery. A proof of concept study applied to farnesyl protein transferase inhibitor design, in QSPR/QSAR studies by molecular descriptors* (DIUDEA M. Ed.) Nova Science Publishers, Inc., New York: 395-439
- HORVATH D., JEANDENANS C. (2003) Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces – A novel understanding of the molecular similarity principle in the context of multiple receptor binding profiles. *J. Chem. Inf. Comput. Sci.* **43**: 680-690
- HORVATH D., JEANDENANS C. (2003) Neighborhood behavior of in silico structural spaces with respect to in vitro activity spaces – A benchmark for neighborhood behavior assessment of different in silico similarity metrics. *J. Chem. Inf. Comput. Sci.* **43**: 691-698
- HORVATH D., MAO B., GOZALBES R., BARBOSA F., ROGALSKI S.L. (2005) Pharmacophore-based virtual screening: strengths and limitations of the computational exploitation of the pharmacophore concept. In *Chemoinformatics in drug discovery* (OPREA T. Ed.) Wiley, New York, U.S.A.
- LIPINSKI C.A., LOMBARDO F., DOMINY, B.W., FEENEY, P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting. *Adv. Drug Deliv. Rev.* **46**: 3-26
- OPREA T.I., GOTTFRIES, J. (2001) Chemography: The art of navigating in chemical space. *J. Comb. Chem.* **3**: 157-166
- PATTERSON D.E., CRAMER R.D., FERGUSON A.M., CLARK R.D., WEINBERGER L.E. (1996) Neighborhood behavior: a useful concept for validation of molecular diversity descriptors. *J. Med. Chem.* **39**: 3049-3059
- PICKETT S.D., MASON J.S., MCLAY I.M. (1996) Diversity profiling and design using 3D pharmacophores: pharmacophore-derived queries. *J. Chem. Inf. Comput. Sci.* **36**: 1214-1223

- POULAIN R., HORVATH D., BONNET B., ECKHOFF C., CHAPELAIN B., BODINIER M.C., DEPRez B. (2001) From hit to lead. Analyzing structure-profile relationships. *J. Med. Chem.* **44**: 3391-3401
- ROLLAND C., GOZALBES R., NICOLAÏ E., PAUGAM M.F., COUSSY L., HORVATH D., BARBOSA F., REVAH F., FROLOFF N. (2004) Qsar strategy for the development of a Gpcr focused library, synthesis and experimental validation. In *Proceeding of the EuroQSAR 2004*, Istanbul, Turkey
- SADOWSKI J., KUBINYI H. (1998) A scoring scheme for discriminating between drugs and non-drugs. *J. Med. Chem.* **41**: 3325-3329
- WILLETT P., BARNARD J.M., DOWNS G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**: 983-996
- WOLD H. (1985) Partial least squares. In *Encyclopedia of Statistical Sciences* (KOTZ S., JOHNSON N.L. Eds) Wiley, New York, U.S.A., Vol. 6: 581-591
- XU J., HAGLER A. (2002) Chemoinformatics and drug discovery. *Molecules* **7**: 566-600

Chapter 14

ANNOTATION AND CLASSIFICATION OF BIOLOGICAL SPACE IN CHEMOGENOMICS

Jordi MESTRES

14.1. INTRODUCTION

A large majority of current drugs have bioactivity due to their binding to a protein, without practically exploiting any of the three other types of macromolecule within biological systems: polysaccharides, lipids and nucleic acids. It is therefore very limited today to discuss targets in general for which we can develop molecules with properties close to those of drugs. The **biological space** in the human genome, containing **proteins** capable of interacting with molecules having properties similar to those of drugs, is referred to as the **druggable genome** (HOPKINS and GROOM, 2002). We can try to estimate the size of this druggable genome by considering that the sequence similarities and the functional analogies of a gene family are often indicative of a general conservation of the active site architecture between all members of the family. It is simply assumed that if a gene family member is capable of interacting with a drug, other family members would probably be able to interact with a molecule having similar characteristics. With this reasoning, among the 30,000 genes or so in the human genome only 3,051 code for proteins belonging to families known to contain drug targets.

The fact that a protein is druggable does not necessarily imply that it is a target of therapeutic interest. The additional condition is that the druggable protein must be implicated in a disease. By imposing this condition, the number of molecular targets is reduced to 600-1,500 proteins of interest for the pharmaceutical industry. Historically, the pharmaceutical industry has tried to develop drugs against 400-500 protein targets. Unfortunately, only a small part of its efforts has manifested in molecules presenting the optimal characteristics for becoming a drug. Currently, the number of proteins targeted by marketed drugs is only 120. The biological space that remains to be explored with the potential for harbouring a therapeutic target is still significant. If we analyse the distribution of these 120 targets, we find that the majority, 88%, correspond to two main biochemical classes: **receptors** and **enzymes**, with 47% being enzymes and 41%, receptors (these are

subdivided into: 30% G protein-coupled receptors, 7% ion channels, and 4% nuclear receptors). It is therefore necessary to understand the composition and classification of these families to try to take maximum advantage of their general characteristics for the design of tests and the screening of molecular libraries directed against the whole family (BLEICHER *et al.*, 2004).

One of the current problems with annotating and classifying biological space is that no standardised classification system exists that encompasses every protein family. The ontology of the target is still an open question (chapter 1). Even in a family of proteins, different classification systems are currently found to coexist. This difficulty, which is still not resolved, greatly limits any initiative in chemogenomics (BREDEL and JACOBY, 2004) intent on integrating chemical and biological spaces based on new *in silico* methods (MESTRES, 2004). This chapter is therefore restricted to an initial overview of the topic, outlining the classification systems currently used for the main protein families of therapeutic interest.

14.2. RECEPTORS

14.2.1. DEFINITIONS

A **receptor** is defined as a molecular structure composed of a polypeptide, which **interacts specifically with a messenger, hormone, mediator, cytokine, or which ensures a specific intercellular contact**. This interaction creates a modification in the receptor which leads, for example, to the opening of a **channel** linked to it, or to the transmission by intermediate enzyme reactions to a distant effector of the receptor. Receptors are situated either at biological membranes (**membrane receptors**), or in the interior of the cell, notably in the nucleus (**nuclear receptors**). A single cell contains several different types of receptor.

Membrane receptors are composed of a section exposed to the exterior of the membrane (for plasma membrane receptors, this domain can be extracellular), where the **recognition site for the messenger molecule** is found, a transmembrane section and an intracellular section. To activate a plasma membrane receptor, the messenger molecule does not need to penetrate into the cell. The activation of membrane receptors by chemical messengers triggers modifications that can stay localised to the membrane, spread throughout the cytoplasm or reach the nucleus. In this last case, activation gives rise to a cascade of intracellular enzyme reactions continuing into the nucleus, whereupon there is modification of DNA and RNA transcription. The series of reactions taking place between activation of the membrane receptor and the cytoplasmic or nuclear effect is generally called **signal transduction**. Serving as a basis for their structural and functional characterisation, three types of **membrane receptor** can be distinguished schematically: ion channel receptors, G protein-coupled receptors and enzyme receptors.

Nuclear receptors are **transcriptional activators**, sometimes specific to particular tissues or cell types, which regulate the programmes of target gene expression by binding to specific response elements situated in their promoter regions. They **transmit the effects of steroid hormones** (oestrogens, progesterones, androgen, glucocorticoids and mineralocorticoids), **thyroid hormones**, **retinoids**, **vitamin D3** as well as activators of cell proliferation, differentiation, development and homeostasis.

14.2.2. ESTABLISHING THE 'RC' NOMENCLATURE

For receptor nomenclature, while awaiting the final recommendations of the Committee on Receptor Nomenclature and Drug Classification of the International Union of Basic and Clinical Pharmacology (NC-IUPHAR), it is necessary to follow the guidelines recently established by the working group on nomenclature of the Editorial Committee of the *British Journal of Pharmacology* (Nomenclature Working Party, 2004). A first general system for the definition, characterisation, and classification of all the known pharmacological receptors has been proposed.

In the first version of this classification system (HUMPHREY and BARNARD, 1998), each receptor receives a unique identifier referred to as an **RC number (Receptor Code)**. A complete RC number is composed of alphanumeric symbols separated by points. The two first codes refer to the structural class and subclass; the third code identifies the receptor family; the fourth code specifies the receptor type; the fifth code characterises the organism. Lastly, the sixth code determines the isoform or splice variant.

Four principal structural classes are defined:

- › **ion channel receptors**,
- › **G protein-coupled receptors**,
- › **enzyme receptors**,
- › **nuclear receptors**.

These classes are assigned the **RC numbers 1 to 4**, respectively. This suggested receptor nomenclature system may be altered in the future and it is not yet certain whether it will be adopted universally by the entire scientific community working on the different receptor classes. A second version of the classification system was recently published (HUMPHREY *et al.*, 2000) and introduces several modifications in the formalisation of this nomenclature (for example, the separation with points is used only to separate structural classes and subclasses, while the other codes are separated by a colon; a new digit is introduced between the family and the receptor type permitting specification of the receptor number) but also in the same classification.

14.2.3. ION-CHANNEL RECEPTORS

Ion-channel receptors, comprising a channel that communicates between the cytoplasm and the extracellular environment, are generally composed of several protein subunits each comprising one or more transmembrane domains with varying topology. **Their activation results in ion flow.** The messenger molecule modulates the opening of the channel and in general regulates the entrance into the cell of either Na^+ , K^+ or Ca^{2+} cations or Cl^- anions.

These ion-channel receptors are to be distinguished: voltage-gated ion channels, which are regulated by the membrane potential and a cellular depolarisation promotes their opening; and other ion-channels, whose opening is regulated by a change in the intracellular Ca^{2+} , cAMP or cGMP concentration.

The general characteristic of ion-channel receptors is that they have an instantaneous response and short duration. Numerous neurotransmitters bind to this type of receptor such as γ -aminobutyric acid to GABA-A receptors, excitatory amino acids (glutamate and aspartate) to ionotropic NMDA and kainate receptors, acetylcholine to nicotinic receptors, and ATP to P2X receptors.

In the classification system proposed by the NC-IUPHAR, the ion-channel receptors are assigned to **structural class 1**. The first version subdivides this class into 8 subclasses (HUMPHREY and BARNARD, 1998), while the second version subdivides it into 9 subclasses (HUMPHREY *et al.*, 2000). Drastic changes to the classification and the nomenclature have been introduced. For example, neurotransmitter receptors have changed from RC 1.8 to 1.9. With regard to the nomenclature, the same receptor can be identified with an entirely different RC. For example, following the first version of the nomenclature system, the human serotonin 5-HT_{3A} ion channel is identified as:

1.1.5HT.01.HSA.00

whereas in the second version, the identifier is:

1.1:5HT:1:5HT3A:HUMAN:00

The subunit of the human P2X₁ receptor is identified in the first version as:

1.4.NUCT.01.HSA.00

and according to the second version, as:

1.3.2:NUCT:1:P2X1:HUMAN:00

The nature and importance of the changes from one version to the other has certainly not helped to spread and establish this classification system.

As an alternative to the system proposed by the NC-IUPHAR, the NC-IUBMB (Nomenclature Committee of the International Union of Biochemistry and Molecular Biology) recommended a general classification of membrane transport proteins (<http://www.chem.qmul.ac.uk/iubmb/mtp/>) based on the classification system developed at the University of California in San Diego (<http://tcdb.ucsd.edu/tcdb/>). In this

system, all transport proteins (including ion channels) are classified by five digits and letters: the first is a number that denotes the transporter class (ion-channel, primary transporter etc.); the second is a letter that corresponds to the transporter subclass; the third is a number that defines the transporter family; the fourth is a number that specifies the subfamily in which the transporter is found; and the fifth digit designates the substrate transported. According to this nomenclature system proposed by the NC-IUBMB, the serotonin human 5-HT_{3A} ion channel is identified by:

1.A.9.2.1

and the subunit of the human P2X₁ receptor by:

1.A.7.1.1

14.2.4. G PROTEIN-COUPLED RECEPTORS

G protein-coupled receptors (GPCRs) are so called because their activity requires the presence of **guanosine diphosphate (GDP)**, which is phosphorylated to give **guanosine triphosphate (GTP)**, this phosphorylation is coupled to the prior transfer of protons as a source of energy. This family comprises **more than a thousand members**, which are activatable by a very wide variety of chemical messengers, of which the large majority are neuropeptides. These receptors have a single polypeptide chain and comprise an **extracellular section** harbouring the **binding site** for the messenger, a hydrophobic **transmembrane section** composed of seven helices (the polypeptide chain crosses the membrane seven times) and an **intracellular section in contact with proteins** which ensures the **transfer and amplification of the signal** received by the receptor to enzymes (adenylcyclase, phospholipase C, and guanylate cyclase) whose activity is thereby regulated. Each G protein is heterotrimeric, *i.e.* composed of three different subunits, α , β and γ , the latter two forming a heterodimeric complex. In order to bind to G proteins, the initially inactive receptor must be activated by its agonist. When activated by their ligands, GPCRs catalyse the exchange of GDP to GTP, which activates both the G $_{\alpha}$ and G $_{\beta\gamma}$ proteins, then becoming able to modulate the activity of different intracellular effectors (enzymes, channels, ion exchangers).

In the classification system proposed by the NC-IUPHAR, G protein-coupled receptors are assigned to **structural class 2**. Different from channel receptors, the second version (HUMPHREY *et al.*, 2000) conserves the number and the code of the three subclasses defined in the first version (HUMPHREY and BARNARD, 1998): rhodopsin, secretin receptor and metabotropic glutamate/GABA_B receptor. The changes to the nomenclature from one version to the other are purely a formality. For example, the rat serotonin 5-HT_{1A} receptor is identified as:

2.1.5HT.1A.RNO.00

(and 2.1:5HT:1:5HT1A:RAT:00, according to the second version)

The rat acetylcholine muscarinic receptor M1 is encoded by:

2.1.ACH.M1.RNO.00

(and 2.1:ACH:1:M1:RAT:00, according to the second version).

Despite the existence of a general classification system proposed by the NC-IUPHAR the scientific community working in this field still uses alternative classification systems. The superfamily of G protein-coupled receptors is divided into six main groups, identified by a sequential reference. The use of numbers (BOCKAERT and PIN, 1999) or characters (HORN *et al.*, 2001) for the identification of different groups depends on the classification system adopted (1-4 and A-D). **Family 1**, or **class A**, contains the group of receptors related to rhodopsin; **family 2**, or **class B**, groups the receptors related to secretin; **family 3**, or **class C**, corresponds to metabotropic receptors for glutamate/pheromone; and **family 4**, or **class D**, contains fungal pheromone receptors. The two groups that remain are annotated differently depending on the classification system. In the numerical system (BOCKAERT and PIN, 1999), receptors of the 'frizzled' and 'smoothened' type are classified as **family 5**, while the cAMP receptors are simply referred to as the **cAMP family**. On the other hand, in the alphanumeric system (HORN *et al.*, 2001), the cAMP receptors are classified as **class E**, whereas receptors of the 'frizzled' and 'smoothened' type are referred to directly as the **frizzled/smoothened family**. The reader is invited to consult the website <http://www.gpcr.org/7tm/> to find details of the alphanumeric classification which has been adopted for the construction of the GPCRDB, the **G protein-coupled receptor database** (HORN *et al.*, 2001).

14.2.5. ENZYME RECEPTORS

The family of **enzyme receptors** groups receptors comprising an **associated enzymatic activity**, activated by the binding of a messenger ligand. Their activation by a messenger modulates this activity, which can one be of various types including: tyrosine kinase, serine/threonine kinase, tyrosine phosphatase or guanylate cyclase, and may be intrinsic or indirectly associated to the receptor. These receptors are composed of one or more subunits each possessing a hydrophobic transmembrane domain. This is the case for example with receptors for insulin, growth factors and atrial natriuretic peptide.

In the classification system proposed by the NC-IUPHAR, enzyme receptors are assigned to **structural class 3**. The second version of the nomenclature (HUMPHREY *et al.*, 2000) keeps the number and the code of the four subclasses as defined in the first version (HUMPHREY and BARNARD, 1998).

14.2.6. NUCLEAR RECEPTORS

Nuclear receptors form a family of transcriptional regulators that are essential for embryonic development, metabolism, differentiation and cell death. Aberrant signalling caused by nuclear receptors induces the dysfunction of cell proliferation,

reproduction and of metabolism, leading to various pathologies such as cancer, infertility, obesity and diabetes (GRONEMEYER *et al.*, 2004). These transcription factors are regulated by ligands; in some cases, their ligands are unknown. In general, receptors with unknown ligands are called **orphan receptors**. The importance of nuclear receptors in human pathology motivates the search for the natural ligand.

Nuclear receptors comprise three principal domains: an amino-terminal transactivation domain (AF-1) having a variable sequence and length and recognised by coactivators and other transcription factors; a highly conserved DNA-binding domain (DBD), which has a structure referred to as a zinc-finger domain since zinc atoms, by bonding to its histidine and cysteine residues, create the appearance of fingers; and a carboxyterminal hormone- or ligand-binding domain (LBD) having a generally well-conserved architecture but which diverges sufficiently thereby ensuring the selective recognition of ligands. The LBD contains the AF-2 activation function.

A systematic nomenclature for nuclear receptors has been proposed by the NC-IUPHAR. In receptor classification, this family of transcription factors is assigned to **structural class 4**, which is divided into two subclasses: 4.1 for non-steroid receptors, and 4.2 for steroid receptors (HUMPHREY and BARNARD, 1998). This nomenclature has not yet been adopted by the whole scientific community.

An alternative nomenclature system specific to nuclear receptors has been proposed (Nuclear Receptors Nomenclature Committee, 1999). In this system, nuclear receptors are grouped according to functional criteria with a three-character identifier. The first digit specifies the subfamily (the six principal subfamilies are assigned an identifier from 1 to 6). All of the nuclear receptors in these subfamilies contain the conserved DBD and LBD domains. However, there are also unusual nuclear receptors that contain only one of these two conserved domains. These receptors are grouped into a seventh subfamily identified by the digit 0. The second character of this nomenclature system is a capital letter, which defines the group in the subfamily; the third digit is an identifier for a particular nuclear receptor.

14.3. ENZYMES

14.3.1. DEFINITIONS

Enzymes are proteins that speed up thousands of times the chemical reactions of metabolism taking place in the cellular or extracellular medium (chapter 5). They act at low concentrations and remain intact at the end of the reaction, serving as biological catalysts. Enzymatic catalysis occurs because enzymes 'bind better' at the transition states of the chemical reaction than to the substrate itself, which produces a substantial reduction in the activation energy of the reaction and thus an acceleration of the reaction rate (GARCIA-VILOCA *et al.*, 2004). A large number of reactions catalyzed by enzymes are responsible for the metabolism of small molecules (HATZIMANIKATIS *et al.*, 2004).

The sequence of enzyme reactions producing a set of metabolites from metabolite precursors and cofactors defines a metabolic pathway. The length of the pathway is the number of biochemical reactions between the precursor and the final metabolite in this pathway. The definition of a metabolic pathway is not unique and the number and length of the pathways can vary depending on the different databases habitually used for studying them. The most frequently used databases currently are:

- › KEGG (KANEHISA *et al.*, 2004; <http://www.genome.jp/kegg/>),
- › MetaCyc (KRIEGER *et al.*, 2004; <http://metacyc.org/>),
- › BRENDA (SCHOMBURG *et al.*, 2004; <http://www.brenda.uni-koeln.de/>),
- › IntEnz (FLEISCHMANN *et al.*, 2004; <http://www.ebi.ac.uk/intenz/>).

14.3.2. THE 'EC' NOMENCLATURE

The classification of enzymes according to their function is based on numerical identifiers comprising four numbers separated by points (TIPTON and BOYCE, 2000). These identifiers are classically known as the **EC numbers (Enzyme Commission numbers)**.

- » The **first number** specifies the enzyme's **class**; there are six classes based on the reaction type that they catalyse:
 - › oxidoreductases,
 - › transferases,
 - › hydrolases,
 - › lyases,
 - › isomerases,
 - › ligases,corresponding to the EC numbers 1 to 6 respectively.
- » The **second number** refers to the **subclass** of the enzyme according to the molecule or functional group involved in the reaction. For example, for the oxidoreductases the subclass indicates the type of group oxidised by the 'donor' (for example 1.1 for the CH–OH group and 1.5 for the CH–NH group), whereas for the transferases the subclass indicates the type of transfer produced (for example, 2.1 indicates the transfer of a single carbon atom and 2.3 the transfer of an acyl group).
- » The **third number** specifies the enzyme's '**sub**'-subclass thus defining the reaction even more precisely. For example, for the oxidoreductases the sub-subclass defines the acceptor (for example, 1.-.1 signifies that the acceptor is NAD or NADP and 1.-.2, a cytochrome), whereas for transferases the sub-subclass gives more information about the group transferred (for example, 2.1.1 is used to classify methyltransferases and 2.1.4, the amidinotransferases).
- » Lastly, the **fourth number** determines the **particular enzyme** in the sub-subclass (for example, 1.1.2.3 refers to L-lactate dehydrogenase and 2.1.1.45, to thymidylate synthase).

The current catalogue of enzymes contains 3804 enzymes classified into 222 subclasses, 63 subclasses, and 6 classes. The potential identification of new enzymes as well as additional information gathered about already existing enzymes can affect established enzyme nomenclature and classification. Consequently, this catalogue is revised periodically by the NC-IUBMB which generates annually revised documents. This information is accessible at <http://www.chem.qmw.ac.uk/iubmb/enzyme/>.

14.3.3. SPECIALISED NOMENCLATURE

Specialised nomenclature has also been developed for enzyme families that are involved in particular metabolic pathways (for instance, lipid metabolism, <http://www.plantbiology.msu.edu/lipids/genesurvey/>) or displaying structural similarities and functional analogies (such as enzymes acting upon sugars in the Carbohydrate-Active-enZYMes or CAZY database, <http://afmb.cnrs-mrs.fr/CAZY/>). These approaches show how the structuring of biological information is not always trivial and that it is guided by the scientific context. It is probable that one nomenclature may be better than another for qualifying a target in a given scientific project, or even that none of the currently available data-structuring systems is satisfactory.

14.4. CONCLUSION

Despite the efforts to order molecular biological entities, starting off most simply with proteins, we are still quite far away from having a general system for annotating and classifying biological space. One of the main problems is the existence of different classification systems that have been established for a long period and fully accepted by the scientific community working in these particular fields. The definition of a global classification system, or of a set of classification systems that ‘cross-talk’ unambiguously within the whole of biological space, will never be established until the moment we have a better understanding of the functional and structural characteristics of the different protein families. To fulfil this objective, it is very important that the current initiatives in structural genomics (STEVENS *et al.*, 2001) supply in the years to come a considerable and representative amount of data relating to different protein families. This structural information should be deposited in the largest publically accessible collection of protein structures, which in its current form, is the **Protein Data Bank (PDB)** (BERMAN *et al.*, 2000) and which the reader can consult at the website <http://www.rcsb.org/pdb/>. To date, the PDB contains over 28,000 structures. The family of enzymes is the best structurally characterised of all families of pharmaceutical interest. The first enzyme structure (subtilisin, EC 3.4.21.62) was solved and deposited in the PDB in 1972 (PDB code: 1sbt). Since then, the presence of enzyme structures in the PDB has increased considerably and in December 2004, there were 13,877 such structures, constituting nearly 50% of the total population of structures deposited in the PDB. The family of nuclear receptors is the second best characterised family as much in number of

structures as in their diversity. The first structure of a nuclear receptor (RXRa, NR 2.B.1) was solved and deposited in the PDB in 1996 (PDB code: 1lbd). In December 2004, there were 150 nuclear receptor structures deposited, including 27 DBDs and 123 LBDs. Unfortunately, the methodological difficulties in over-expressing, purifying, stabilising and crystallising membrane proteins are yet more numerous. Consequently, the number and diversity of G protein-coupled receptor structures and ion-channel receptors in the PDB is currently very limited. The reader is invited to consult the website <http://cgl.imim.es/fcp/> for an analysis of the structural representation of protein families in the PDB.

This chapter describes the state of progress and characterisation of the target, defined in molecular detail for proteins. Targets defined functionally in terms of metabolic or cellular integration, though, are still without a consensual view (chapter 1). There is still a wide scope to define the **biological space of targets** at the level of **simple molecular structures** (nucleic acids, such as DNA regions, metabolites such as reaction intermediates trapped in molecular cages), of **complex molecular structures** (multi-protein complexes, metabolic pathways) or even at the level of **integrated functions on the scale of the cell or whole organism** – and this represents an important challenge for the future.

14.5. REFERENCES

- BERMAN H.M., WESTBROOK J., FENG Z., GILLILAND G., BHAT T.N., WEISSIG H., SHINDYALOV I.N., BOURNE P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**: 235-242
- BLEICHER K.H., BÖHM H.J., MÜLLER K., ALANINE A.I. (2003) Hit and lead generation: beyond high-throughput screening. *Nat. Rev. Drug Discov.* **2**: 369-378
- BOCKAERT J., PIN J.P. (1999) Molecular tinkering of G protein-coupled receptors: an evolutionary success. *EMBO J.* **18**: 1723-1729
- BREDEL M., JACOBY E. (2004) Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **5**: 262-275
- FLEISCHMANN A., DARSOW M., DEGTYARENKO K., FLEISCHMANN W., BOYCE S., AXELSEN K.B., BAIROCH A., SCHOMBURG D., TIPTON K.F., APWEILER R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res.* **32**: 434-437
- GARCIA-VILOCA M., GAO J., KARPLUS M., TRUHLAR D.G. (2004) How enzymes work: analysis by modern rate theory and computer simulations. *Science* **303**: 186-195
- GRONEMEYER H., GUSTAFSSON J.Å., LAUDET V. (2004) Principles for modulation of the nuclear receptor superfamily. *Nat. Rev. Drug Discov.* **3**: 950-964
- HATZIMANIKATIS V., CHUNHUI L., IONITA J.A., BROADBELT L.J. (2004) Metabolic networks: enzyme function and metabolite structure. *Curr. Opin. Struct. Biol.* **14**: 300-306

- HOPKINS A.L., GROOM C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* **1**: 727-730
- HORN F., VRIEND G., COHEN F.E. (2001) Collecting and harvesting biological data: the GPCRDB and NucleaRDB information systems. *Nucleic Acids Res.* **29**: 346-349
- HUMPHREY P.P.A., BARNARD E.A. (1998) International Union of Pharmacology. XIX. The IUPHAR Receptor Code: a proposal for an alphanumeric classification system. *Pharmacol. Rev.* **50**: 271-277
- HUMPHREY P.P.A., BARNARD E.A., BONNER T.I., CATTERALL W.A., DOLLERY C.T., FREDHOLM B.B., GODFRAIND T., HARMER A.J., LANGER S.Z., LAUDET V., LINBIRD L.E., RUFFOLO R.R., SPEDDING M., VANHOUTTE P.M., WATSON S.P. (2000) The IUPHAR Receptor Code. In *The IUPHAR Compendium of Receptor Characterization and Classification*, 2nd Edition, IUPHAR Media, London: 9-23
- KANEHISA M., GOTO S., KAWASHIMA S., OKUNO Y., HATTORI M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32**: 277-280
- KRIEGER C.J., ZHANG P.F., MUELLER L.A., WANG A., PALEY S., ARNAUD M., PICK J., RHEE S.Y., KARP P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **32**: 438-442
- MESTRES J. (2004) Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr. Opin. Drug Discov. Dev.* **7**: 304-313
- Nomenclature Working Party (2004) Nomenclature Guidelines for Authors. *Br. J. Pharmacol.* **141**: 13-17
- Nuclear Receptors Nomenclature Committee (1999) A unified nomenclature system for the nuclear receptor superfamily. *Cell* **97**: 161-163
- SCHOMBURG I., CHANG A., EBELING C., GREMSE M., HELDT C., HUHN G., SCHOMBURG D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.* **32**: 431-433
- STEVENS R.C., YOKOYAMA S., WILSON I.A. (2001) Global efforts in structural genomics. *Science* **294**: 89-92
- TIPTON K., BOYCE S. (2000) History of the enzyme nomenclature system. *Bioinformatics* **16**: 34-40

MACHINE LEARNING AND SCREENING DATA

Gilles BISSON

15.1. INTRODUCTION

In all living beings, to differing degrees and with the help of extremely varying mechanisms (genetic, chemical or cultural), one observes an aptitude for acquiring new behaviour through their interaction with the environment. The objective of **machine learning** is to study and put into effect such mechanisms using artificial systems: robots, computers etc. Of course, beyond this very general view, the objectives of machine learning are often more pragmatic and the following two definitions describe quite well the type of activities grouped within this discipline:

- › “*Learning consists of the construction and modification of representations or models from a series of experiments*” (MICHALSKI, 1986)
- › “*Learning consists of improving the performance of a system with a given task based on experiments*” (MITCHELL, 1997)

Both of these definitions contain the key word **experiment**, which must be understood to mean the possibility of having a representative set of observations (a **sample**) of the phenomenon we want to model and/or the process to be optimised. The increasing interest we see today towards methods in machine learning can be readily justified.

Indeed, the general use of computers and the implementation of automated processes for data acquisition, such as automated chemical library screening, facilitates the creation and exchange of databases and more generally of ever larger quantities of information. Also, in every sphere of activity an increasing need arises for intelligent systems capable of assisting humans in carrying out demanding or routine tasks, in particular for the analysis of data flows produced by scientific experiments. However, when one works in complex and poorly formalised domains it is often difficult, indeed impossible, to describe beforehand in a precise and optimal manner the computational processing that needs to be undertaken. The process of analysis must therefore be constructed around **heuristic processes to find a solution** (*heuristic*: a technique consisting of using a set of empirical rules to solve a problem more rapidly, with the risk of not obtaining an optimal solution). We

find ourselves thus pushed to design and develop **generic learning tools** (in the sense that the methods proposed are not specific to a particular field) that are able to analyse autonomously the data arising from the environment and, based on this, able to make decisions and find correlations between the data and the knowledge handled.

We can use machine-learning techniques to characterise non-exhaustively, for instance, families of concepts, to classify observations, to optimise mathematical functions and to acquire control knowledge (example 15.1).

Example 15.1

Let us imagine a pretend screening experiment in which the toxicity of 8 molecules is tested, these molecules being represented by a set of 4 descriptors (fig. 15.1). The aim is to construct a predictive **SAR (Structure-Activity Relationship)** model that is expressed as a **decision tree** of smallest size. In such a tree, which is easy to create with the help of an algorithm such as C4.5 (QUINLAN, 1993), each node corresponds to a descriptor and the branches correspond to the values that this descriptor can take; the leaves of the tree correspond to the predicted activity.

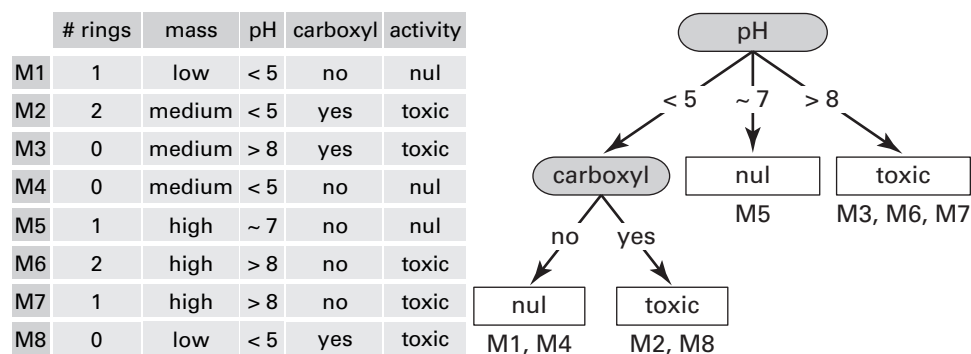


Fig. 15.1 - The table contains the results of a fictitious screening experiment in which the molecules tested are described by four descriptors: the number of cycles, the mass, the pH in solution and the presence of a carboxyl radical. The decision tree was constructed from these data and allows prediction of the molecules' activity by using the pH and carboxyl descriptors.

The problem dealt with in example 15.1 is trivial; however, when working with several tens or hundreds of thousands of molecules described by hundreds of thousands of descriptors and for which the degree of activity is not known with certainty, it is easy to imagine the theoretical and algorithmic complexity of the computational problem to be solved. It is of note that the construction of a decision tree is certainly not the only technique used in machine learning. Among the intensively studied approaches at the beginning of this millenium are **inductive logical programming (ILP)**, **BAYESIAN** and **MARKOV** approaches, **large-margin separators**, **genetic algorithms**, **conceptual clustering**, **grammatical inference** and **reinforcement learning**. It is also necessary to emphasise the links maintained between machine learning and data analysis and statistics. The reader wishing to learn more about this field of study is referred to WITTEN and EIBE (1999) for a

working description of the methods in machine learning and to the books by MITCHELL (1997), and CORNUÉJOLS and MICLET (2002) for a description of the theoretical and methodological foundations of the discipline. Lastly, the work by RUSSELL and NORVIG (2003) presents the methods for heuristic exploration and more general information about the field of Artificial Intelligence.

Thus, due to the diversity of approaches and the complexity of the problems treated, machine learning finds itself at the crossroads of numerous disciplines such as mathematics, human sciences and artificial intelligence. Besides, the fields supposedly more distanced from computer science – such as biology and physics – are also a major source of ideas and methods. However, whatever the paradigm adopted, the setup of a machine learning system remains very similar; the remainder of this chapter will explore this further.

15.2. MACHINE LEARNING AND SCREENING

No sooner than we become interested in the acquisition of new knowledge (as opposed to the optimisation of existing knowledge) are we confronted by **two principal issues** which are: **supervised learning** and **unsupervised learning**. In the first case the set S of *example* data to be analysed is described by a set of pairs $\{(X_i, U_i)\dots\}$ where each i is characterised by a description X_i and a label U_i denoting the class to which it belongs, or more generally the concept to be characterised. The machine-learning step consists therefore of constructing a *set of hypotheses*, h , (or a function when the data are numerical, which essentially involves regression) which establishes a relationship between the class of individuals as a function of the description associated with it, such as: $U_i = h(X_i)$. This set of hypotheses allows a **classifier** to be constructed, *i.e.* a computer program that enables later prediction of a class of new examples. Very often, when there are several possible U_i labels, we are dealing, without a loss of generality, with a two-label problem having a separation between the *positive examples*, which illustrate the concept U_i about which we seek to learn a characteristic, and the other $U_{j \neq i}$ called *negative examples* or *counter-examples*. It is important to take into account the negative examples as they permit the search to be focussed on the discriminatory properties. In the case of unsupervised learning – the term *classification* is also used – the U_i labelling does not exist and the system only has the single X_i descriptions. The aim is thus to construct a **partition** of S in which the different groups produced are consistent (the individuals placed in the same group resemble each other) and contrasting (the groups are well differentiated).

Armed with these definitions, we can now ask ourselves what, in the case of data arising from a high-throughput screen, are the problems that may potentially benefit from machine-learning techniques? For the moment, we shall not discuss the problem of representing these data computationally.

Figure 15.2 shows the results obtained from a real screening experiment. The horizontal axis corresponds to the plate number containing the compounds analysed and the vertical axis, the intensity of the *bioactivity signal*. Each grey point represents the results obtained for each molecule in the chemical library. The black and pale grey diamonds (lower down) represent the values respectively of the positive and negative *control* molecules in the plate, which act as reference points. In this context, the machine-learning examples X_i correspond to descriptions of the molecules tested and the U_i labels to the signal which is expressed either in numerical form, or more often in a discrete form (active and inactive molecules). As we can see, a first round of data normalisation is generally necessary to standardise the signal intensities (chapter 4), but again this will not be discussed here.

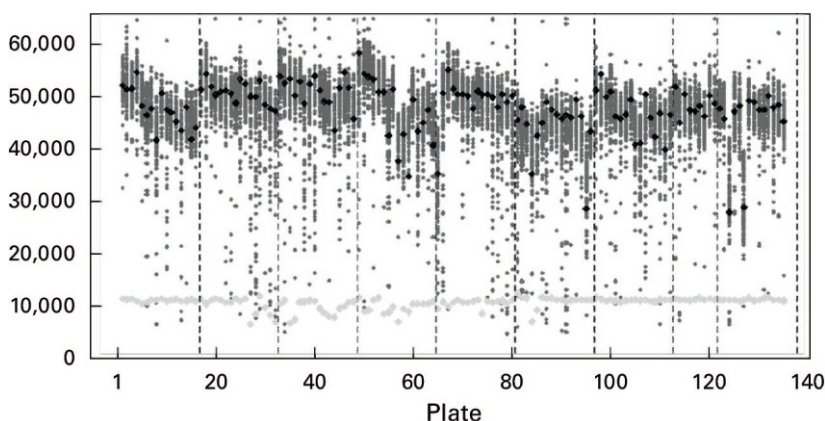


Fig. 15.2 - Graphical representation of the results produced by chemical library screening. The data shown constitutes a whole dataset.

The selection of **hits** in the **primary screen** (see chapter 1) is decided relative to a given activity threshold. Now, the measurements are characterised by a large uncertainty (variability of the phenomenon studied, tests carried out at an average concentration etc.) and the use of a threshold induces a certain number of *false positives* (molecules kept by mistake) and *false negatives* (molecules rejected by mistake). In a way, the presence of false positives is not such a problem as they will be mostly eliminated during the **secondary** screening (see chapter 1) the purpose of which is precisely to verify the activity of each hit. The same is not true of the false negatives as this entails a loss of information, which is all the more problematic as the proportion of active molecules in the screen becomes lower (on the order of 2 to 5%). It would therefore be useful to learn to find again these false negatives in the chemical library afterwards by using the results of the primary and secondary screens as training data according to the mechanism described in figure 15.3.

The idea here is to use confirmed molecules (Hit 2) as positive examples, and those retained during the first screening but not confirmed during the second (Hit 1) as negative examples. The description of molecules has to contain, in addition to the

physicochemical data, information about the experimental screening conditions (distribution of hits in the plate, date of screening, concentration tested etc.) since false positives can arise as a consequence of the experimental conditions. The classifier learnt is then used with the part of the chemical library that was not retained in order to detect possible candidate molecules FN_i , which should then be validated by a complementary screen. In this approach the classifier learnt is specific to the screen undertaken, this process must therefore be integrated into the platform management software.

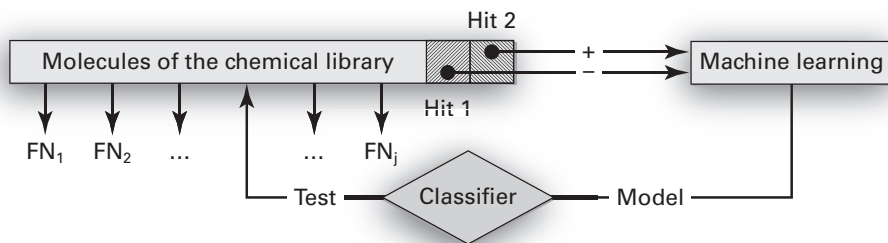


Fig. 15.3 - Construction of a classifier

in order to identify the examples wrongly classified as negative

In this process, the confirmed molecules (Hit 2) are taken to be positive examples; the molecules retained from the primary screening, but not confirmed after the second (Hit 1) are taken as negative examples. The description of molecules must include, in addition to structural data, information about the experimental conditions of the screening (distribution of hits in the plate, screening date etc.) since false positives can result from experimental problems. The classifier learnt is then used with the whole chemical library to detect possible candidate molecules, which then have to be confirmed with a complementary screen.

Furthermore, in the context of unsupervised machine learning and, at a push, semi-independently of the problem of a given screening, it can be interesting to use classification techniques to organise the molecules from one or more chemical libraries into families in order to obtain a *mapping of structural space* (chapter 13) or indeed to build libraries of recurrent motifs as proposed by the SUBDUE system and its extensions (COOK and HOLDER, 1994; GONZALEZ *et al.*, 2001). These approaches are notably interesting for the extraction of relevant and significant substructures permitting the molecules to be rewritten in a more concise form and thus helping to set up SAR models (DESHPANDE *et al.*, 2003). However, it is clear that **the main problem posed by an analysis of the screening results is precisely this automatic development of a SAR model which relates the molecule's behaviour to its properties and structure**. In this context, we are again dealing with an instance of supervised learning where the active molecules correspond to positive examples and the others, the vast majority, to negative examples. The aim is to supply the experimenter with a set of hypotheses, typically expressed in the form of molecule fragments, capable of explaining the experimental results and above all of facilitating the synthesis of other compounds.

This problem has been studied for tens of years in **ILP** (inductive logical programming), a subfield of machine learning focussed on *structured representations*. The advantage of this technique is the ability **to represent the molecules directly in the form of sequences or graphs**, without having to redescribe them with the help of **fingerprints** (or structural keys; chapter 11), *i.e.* a list of predefined fragments, as is typically the case with **QSAR** analyses (these methods generally rely on classical statistical techniques such as regression and the representation of models and draw on the global properties of molecules – geometric, topological, electronic, thermodynamic – about which the reader can learn more online [Codessa]). In ILP, the work was initiated by the group of S. MUGGLETON with the PROGOL system (SRINIVASAN *et al.*, 1999; MARCHAND-GENESTE *et al.*, 2002) which uses a representation in logic of the data and knowledge. The most recent systems combine these approaches with kernel methods coming from **SVM** (FROEHLICH *et al.*, 2005; LANDWEHR *et al.*, 2006). In parallel, other systems which use representation languages such as SMILES/SMART, designed by and for chemists, have also been proposed, the most well known being WARMR (DESHAPE and DE RAEDT, 1997), MOLFEA (HELMA *et al.*, 2003b) and CASCADE (OKADA, 2003). The reader will find a short summary of these approaches in the review by STERNBERG and MUGGLETON (2003).

On the whole, while the results are remarkable (HELMA *et al.*, 2000) they are limited most of the time to the *rediscovery of results* that are more or less well known. Thus, FINN *et al.* (1998) identified pharmacophores on inhibitors of angiotensin-converting enzyme which had been partially set out by MAYER (1987); similarly, the MOLFEA system again found that AZT is an active molecule in a dataset [HIV-Data-Set-1997] containing the results of screening 32000 molecules. The same goes for contests like the **Predictive Toxicology Challenge, PTC** (HELMA and KRAMER, 2003a; [Helma-PredTox]), which allows a comparison of the different techniques using datasets for which the characteristics are entirely known. This criticism must be tempered by recognising that these works are relatively recent and they do tackle complex problems. Besides, as we shall see when describing the methodology for using this kind of system, it is necessary to understand that the value of machine-learning techniques cannot be measured by using the final result as the sole yardstick.

15.3. STEPS IN THE MACHINE-LEARNING PROCESS

Figure 15.4 schematises the key steps in using a system for machine learning. It is important to note that at present this is in general not a ‘push-button’ process. Quite the opposite, the use of such a system is only relevant as an iterative procedure for the acquisition and formalisation of knowledge necessitating many exchanges (working on the time-scale of a month) between the machine-learning expert and the subject expert – here, a biologist and/or a chemist.

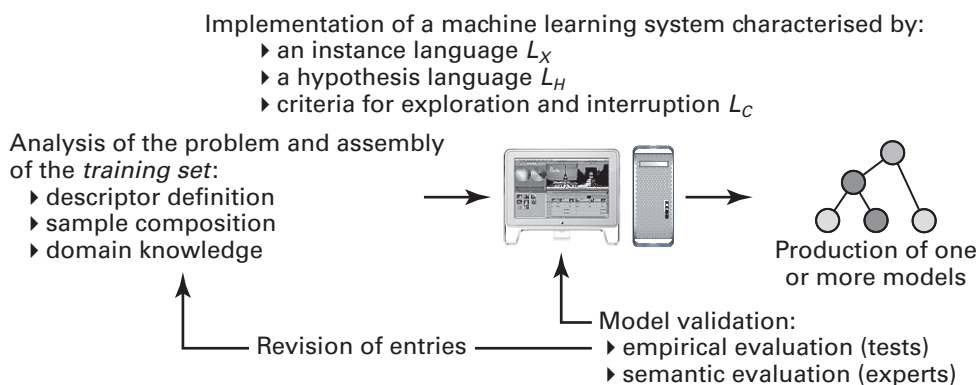


Fig. 15.4 - Using a machine-learning system

The process is intrinsically iterative linking the different phases: description of the problem, construction and validation of the models, and then revision of the entries.

In this scheme, the data collection phases and the use of a machine-learning system can be more or less closely coupled. Thus, if one decides to use a specific tool, the information can be directly encoded by using the specific features of its representation languages, L_X and L_H (discussed below). Conversely, if we work within a wider scope or when seeking to explore different machine-learning paradigms there is a need to create a database containing the training data, then to translate these data into an *ad hoc* formalism, which is not always a trivial task. To simplify the point we shall consider below to be working within a pre-defined system.

15.3.1. REPRESENTATION LANGUAGES

As shown in [figure 15.4](#), leaving aside algorithmic processes, a machine-learning system is characterised by the languages used to communicate with the user. By user, we mean here the team formed by (at least) the expert in machine learning and the expert in bio-chemoinformatics.

We thus have:

- ▶ the instance language L_X that permits a representation of the elements in the training set or, in the case of screening, a description of the molecules in the chemical library,
- ▶ the hypothesis language L_H that serves on the one hand to formulate hypotheses and hence, in the end, the model expressing the knowledge acquired (and which will form the basis for the classifier allowing classification of new examples), and on the other, the domain knowledge which has been explained by the experts,
- ▶ lastly, the set of criteria used by the system to control its learning and to decide when to stop is described in the language L_C , also called *bias language*. This last language can be considered to be either constituted of a simple set of parameters permitting decisions to be made about the characteristics of the model learned (for example: number of rules, size of the rules) or, on the contrary, a true language enabling complex control of the learning process.

With regards to L_X and L_H , we can define schematically several paradigms (table 15.1), with a divide between **structured and vectorial representations** and for the latter, a distinction between **numerical and symbolic approaches**.

Table 15.1 - Principal modes of representation for the examples and knowledge in machine learning

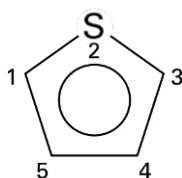
	Vectorial representations	Structured representations
Examples (L_X)	Numerical ▶ Table of data $M(i, j)$	▶ Relationship graph
	Symbolic ▶ Propositional logic: <i>molecular_weight=167,</i> <i>cycle_number=6,</i> <i>contains_Cu=true...</i>	▶ Predicate logic: <i>bond(m1,c1,o,double),</i> <i>bond(m1,c1,c2,simple),</i> <i>aldehyde(m1)</i>
Knowledge (L_H)	Numerical ▶ Parameters of the model learned $[a_1, ..., a_n]$	▶ Relationship graph
	Symbolic ▶ Decision tree ▶ Decision rules: <i>if(molecule_weight<500), (logP>5)...</i> <i>then(drug_candidate=true)</i>	▶ PROLOG clauses: <i>mutagen_molecule (M:</i> <i>bond(M,A1,A2,double),</i> <i>in_a_cycle(M,A2,5),</i> <i>atom_type(A1,Cu)...</i>

In the vectorial representations, the examples are described in the form of a set of properties that can be expressed either in a data matrix where each line represents an example and each column a descriptor, or with propositional logic in the form of a conjunction of properties: 'attribute = value'. The models learned (or the domain knowledge) are expressed in a similar way. In structured representations each example corresponds to the description of a collection of objects which is either assimilable to a labelled graph, given attributes and possibly oriented if the relationships between the objects are non-commutative, or to a symbolic description by using predicate logic. It is the same for knowledge that is often expressed as clauses in the PROLOG language.

It is important to underline that all of these representations do not have the same degree of expressivity (example 15.2). Thus, if it is easy to describe the 2D (or 3D) structure of a molecule in the context of structured representations, in a vectorial representation the user is obliged to describe the molecules with the help of a finite set of **dictionary-based fingerprints** (chapter 11), which limits the possibility of learning since the system cannot generate new ones. This richness of expression however has a corollary: the more high-level the representation language L_H , the more the learning process is complex and (potentially) long.

Example 15.2 - representation of the examples in the case of Thiophene

Here are a few ways of representing this molecule. In the case of representations in the form of predicates or propositions of course there are other ways to define the language descriptors.



▶ Representation in the form of predicates:

*bond(C1,S),bond(S,C2),bond(C2,C3)...**atom(S,sulphur)...*

▶ Representation in SMILES format:

S2C=CC=C2

▶ Representation in the form of propositions:

#cycle5=1,contains_sulphur=true, sulphur_position=,2...

To limit this complexity, as we have previously seen, several systems use the language **SMILES** (Simplified Molecular Input Line Systems) by WEININGER (1988) which permits representation of molecules in the intermediate form of 1D sequences (example 15.2). Besides, from an algorithmic point of view many systems use ‘propositionalisation’ allowing the transformation, with certain limitations, of structural data into vectorial data. This is true for the STILL system (SEBAG and ROUVEIROL, 2000) which achieved good results with the PTC dataset, even if it is detrimental to the readability of the models built. Other examples of similar research include the works by ALPHONSE and ROUVEIROL (2000), KRAMER *et al.* (2001), FLACH and LACHICHE (2005) or approaches like **SVM** (Support Vector Machine; VAPNIK, 1998), by using specific kernels (GÄRTNER, 2002) to process the structural data.

15.3.2. DEVELOPING A TRAINING SET

The goal of this step is to develop a description of the problem to be submitted to the learning system. This description implements three types of information. First of all, it is necessary to define the set of *descriptors* used to describe the examples. In the case of chemistry, the possibilities are vast ranging from the general physicochemical properties of the molecule to the description of the structure in 1D, 2D or 3D form (chapter 11; FLOWER, 1998; TODESCHINI and CONSONNI, 2000; KING *et al.*, 2001; BESALU *et al.*, 2002). The choice of representation will depend on several constraints: the sort of information accessible (for instance, the 3D structures of molecules are not always known), the possibility of using the L_X and L_H languages of the selected system and obviously the kind of problem to be solved, which infers a set of presumably appropriate properties. This choice of descriptors is fundamental as it determines what is ‘learnable’ in the model. Thus, in the field of medicinal chemistry certain fundamental properties of pharmacophores like the presence of hydrophilic/hydrophobic radicals and of protein donors/acceptors etc. must be explicitly supplied to the machine learning system in one form or another as they cannot be easily gleaned from the structure. If the descriptor database is too large, ‘selection of variables’ methods can be used (LIU and YU, 2005; [LIU-YU, 2002]) to determine the most informative descriptors.

The second step consists of defining, in parallel to the choice of representation language, the *examples* that will be used in the learning process. In the case of screening this choice is relatively obvious. Since a predefined chemical library is employed, the examples naturally are molecules. However, if this chemical library is too large, it may be necessary to carry out a pre-sampling in which a subset of negative examples (*i.e.* molecules that are not hits) are kept in order to speed up the learning process.

Finally, the assembly of *strategic knowledge*, which guides machine learning, is an important step. Definition of this knowledge with the L_c language enables the system to converge more quickly towards the correct hypotheses (example 15.3).

However, this information is not always known beforehand and one of the roles of the learning process is precisely to facilitate its acquisition by establishing dialogue between the experts.

Example 15.3 - expression of constraints on the model sought

In the search carried out by PROGOL, based on screening for small-molecule inhibitors of angiotensin-converting enzyme (FINN *et al.*, 1998), the authors used the following constraints in order to limit the search space: each 'candidate solution' (fragment) must contain a hydrogen donor *A*, a hydrogen acceptor *B* such that the distance between *A* and *B* equals 3 Å (± 1 Å), and it must have a Zinc site.

LIPINSKI's 'rule' (LIPINSKI *et al.*, 2001), which expresses a set of properties generally borne out by drug candidates, is also a good example of constraints that are appropriate to introduce into the system when the component sought has a therapeutic objective. ▀

15.3.3. MODEL BUILDING

Once the training set has been developed, the learning system can be implemented. Fundamentally, building a SAR model can be described as a *heuristic search* process (fig. 15.5) in the space of all possible hypotheses *H* which is implicitly defined by the system's L_H language. Of course, the size of this space is potentially gigantic: for structured representations it corresponds to a description space of all imaginable molecular fragments. The aim of machine learning is thus to search the minimal set of hypotheses termed *complete* (which satisfy the positive examples of *X*) and *consistent* (which reject the negative examples of *X*), in agreement with the strategic constraints expressed with L_C . Once the criteria for stopping the system have been fulfilled, the set of best hypotheses found makes up the final model returned to the experimenter (example 15.4).

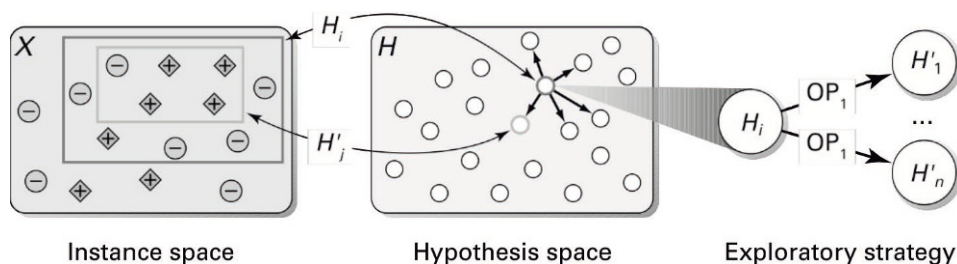


Fig. 15.5 - In the case of the 'generate and test' type of approach, the search process is as follows: at any given instant, the machine-learning system has one (or several) hypotheses H_i and it generates new H'_n by applying certain transformation rules Op_n which are specific to it. At each iteration, it conserves the best hypothesis (or hypotheses) H'_j selected most often based on the criterion of 'coverage', in other words, the ratio between the number of recognised positive and negative examples, as well as the description's compactness.

Let us note that for screening data exhibiting a pronounced imbalance between the number of positive and negative examples in the chemical library, the process of selecting hypotheses with learning algorithms by coverage can require certain adjustments. For example, the method developed by SEBAG *et al.* (2003) can be used.

This method replaces the usual performance criterion based on the error rate, by that based on the **optimisation of the area under the curve** (also called a **Receiver Operating Characteristic** or **ROC curve**), which is classic in the analysis of medical data expressing the inverse relationship between the *sensitivity* and the *specificity* of a test.

Example 15.4 - the FOIL system and its function

FOIL (QUINLAN, 1990) is a classic system in ILP that builds its models by successive specialisations of its hypotheses. These are represented in the form of PROLOG clauses that correspond to the description of molecular fragments. The algorithm is structured around two iterations: the purpose of iteration [1] is to build as many clauses as necessary in order to cover all positive molecules (the activity is not necessarily 'explainable' by the description of a single fragment). Iteration [2] allows each fragment to be built incrementally.

Programme = $\emptyset$

P = set of predicates present in the active molecules

Iteration [1] While the set P is not empty:

N = set of predicates present in the inactive molecules

C = concept_to_learn ($X_1, \dots, X_n$):-

Iteration [2] While the set N is not empty:

Build a set of candidate predicates $\{L_1 \dots L_k\}$

Evaluate the coverage I of the fragments $C \wedge L_i$ in the sets P and N

$$\text{Couv}(C \wedge L_i) = \log_2 \left(\frac{|P_{\text{Couvert}}|}{|P_{\text{Couvert}}| + |N_{\text{Couvert}}|} \right)$$

Add the selected term L_c to the clause C

Remove from N the fragments of negative molecules rejected by C

Add the learned clause C to the Program

Remove from P the fragments of positive molecules covered by C

Here is a fictitious example of iteration [2] to learn the concept of mutagen molecule M :

Initially : mutagen(M):-

Iteration 1: mutagen(M):-bond($M, A1, A2, \text{double}$)

Iteration 2: (M):-bond($M, A1, A2, \text{double}$), in_a_ring($M, A2, 5$)

Iteration 3: mutagen(M):-bond($M, A1, A2, \text{double}$), in_a_ring($M, A2, 5$),
atom_type($A1, \text{Cu}$)

This last clause can be interpreted as follows: *a molecule is a mutagen if it possesses an aromatic ring of 5 carbons in which one of the atoms is linked to a copper atom by a double bond.* 

15.3.4. VALIDATION AND REVISION

Lastly, when the learning step is complete, it is necessary to evaluate the model (and therefore the classifier) that has been built. This evaluation can be carried out on two different levels: *statistical* and *semantic*.

- » Regarding the **statistical evaluation**, the classic means consists of measuring the classifier's prediction rate, *i.e.* its capacity to predict the label U_i of example X_i (see section 15.2). Two types of error rate can be distinguished and are presented in [figure 15.6](#).
 - » Firstly, the **learning error** corresponds to the error made by the classifier with the examples used in learning. Contrary to what one might think, this error is

not always nil. In many cases the system only learns one ‘approximate’ model, which is far from being a failing, notably if the data entered contained uncertainty (referred to as *noise*) in the values of the descriptors or in the class labels; this is typically the case in screening.

- › Secondly, the **generalisation error** corresponds to the error made by the classifier with the examples contained in a new sample not studied during the learning.

In practice however, it is not always possible to benefit from an independent database of new examples; for screening, notably if the quantity of negative examples permit this type of validation, it is no longer the case for positive examples which can be quite few in number. A **cross validation** must therefore be undertaken by splitting up the set of positive and negative examples into N subsets $\{N_1 \dots N_j\}$, then by building iteratively M_i models of different classifiers. Each model is built with the $\{N_1 \dots N_j\} - N_i$ subsets and tested on the examples from the N_{th} subset.

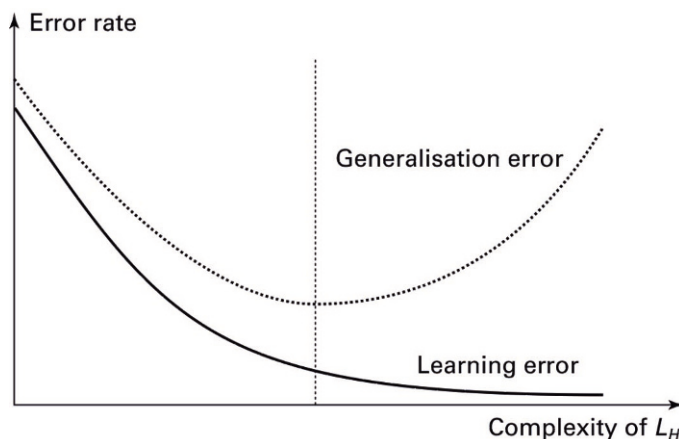


Fig. 15.6 - Evolution of the machine-learning and the generalisation error rates, as a function of the complexity of the L_H representation language used

The more complex the language, the more the system learns an exact model. However, beyond a certain threshold, whereas the performance in machine learning continues to improve, that of generalisation worsens. This paradox is known as the ‘bias-variance’ compromise and corresponds to the fact that with too complex a language (*i.e.* where the hypothesis space H is too vast) the information supplied by the examples (which are constant in number) is no longer enough to direct the system towards a model that is really predictive. It is therefore important to work at the ‘right level’ of representation.

- › Lastly, the **semantic evaluation** aims to understand the significance of the model developed by the system and therefore to judge its relevance; this evaluation can only be reasonably carried out by experts in the field. With this perspective, it is clear that the learning systems working with symbolic representations (as opposed to numerical representations) facilitate this communication. It is thus relatively easy to transform a PROLOG clause into a 2D scheme representing the molecular fragment judged to be significant by the system.

Besides the (important) aspect of acquiring the result, the statistical and semantic evaluations aim above all to identify what information is missing or incomplete in order to **improve the modelling of the problem** carried out with the training set. As a function of the analysis performed during the evaluation phase, in agreement with experts in the field, various actions may thus be undertaken, e.g. addition or removal of descriptors, modification of learning parameters, introduction of new constraints to control the generation of hypotheses etc.

15.4. CONCLUSION

Today, machine learning offers a large number of methods to discriminate or to categorise data. These methods permit a rapid exploration of large databases of experimental results, which is typically the case for screening data, and to build predictive and/or explanatory models. From a practical point of view, the majority of systems are either available commercially, or as *free software* on the Internet. For instance the **WEKA platform** (WITTEN *et al.*, 2005; [Weka]) is a toolbox integrating numerous learning methods (at the moment all are vectorial) coupled to services for the visualisation and manipulation of data streams.

However, in domains requiring complex expertise, it is clear that the use of learning methods demands a significant and sustained investment of time and people to prepare the data, to calculate the model, to evaluate the results and to set up the subsequent iterations. This being the case, **one of the basic advantages** of a procedure centred on machine-learning techniques **resides in the very process of modelling and formalisation that these approaches require**. Due to the systematic nature of algorithms, the user is in fact compelled to remove all of these ambiguities and to express the set of his or her hypotheses and knowledge explicitly. Currently, many projects aim to go further in this approach to the modelling of screening results, either by trying to implement virtual screening methods (BAJORATH, 2002, and SEIFERT *et al.*, 2003; see also chapter 16) which may possibly rely on considerable use of machine-learning methods ([Accamba]), or by trying to integrate into the algorithms control of the experimental process itself for the analysis of molecules (KING *et al.*, 2004).

15.5. REFERENCES AND INTERNET SITES

[Accamba]: website of the Accamba project: <http://accamba.imag.fr/>

ALPHONSE E., ROUVEIROL C. (2000) Lazy propositionalisation for Relational Learning. In *Proc. of the 14th European Conference on Artificial Intelligence (ECAI-2000)*, IOS Press, Berlin: 256-260

BAJORATH J. (2002) Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discov.* 1: 882-894

BESALÚ E., GIRONÉS X., AMAT L., CARBÓ-DORCA R. (2002) Molecular quantum similarity and the fundamentals of QSAR. *Acc. Chem. Res.* **35**: 289-295

[Codessa]: website giving a list of descriptors organised by type:

<http://www.codessa-pro.com>

COOK D.J., HOLDER L.B. (1994) Substructure Discovery Using Minimum Description Length and Background Knowledge. *J. Artif. Intell. Res.* **1**: 231-255

CORNUEJOLS A., MICLET L. (2002) *Apprentissage Artificiel*. Eyrolles, Paris

DEHASPE L., DE RAEDT L. (1997) Mining association rules in multi-relational databases. In *Proc. of ILP'97 workshop*, Springer Verlag, Berlin-Heidelberg-New York: 125-132

DESHPANDE M., KURAMOCHI M., KARIPYS G. (2003) Frequent Sub-Structure-Based Approaches for Classifying Chemical Compounds. Technical Report. In *Proc. of IEEE Int. Conference on Data Mining (ICDM03)* IEEE Computer Society Press, Melbourne, Floride

FINN P., MUGGLETON S.H., PAGE D., SRINIVASAN A. (1998) Pharmacophore discovery using the Inductive Logic Programming system PROGOL. *Machine Learning* **30**: 241-271

FLACH P., LACHICHE N. (2005) Naive Bayesian Classification of Structured Data. *Machine Learning* **57**: 233-269

FLOWER D.R. (1998) On the properties of bit string-based measures of chemical similarity. *J. Chem. Inf. Comput. Sci.* **38**: 379-386

FRÖHLICH H., WEGNER J., SIEKER F., ZELL R. (2005) Optimal Assignment Kernels for Attributed Molecular Graphs. In *Proc. of Int. Conf. on Machine Learning (ICML)*: 225-232

GÄRTNER T. (2003) A survey of kernels for structured data. *ACM SIGKDD Explorations Newsletter* **5**(1): 49-58

GONZALEZ J., HOLDER L., COOK D. (2001) Application of graph based concept learning to the predictive toxicology domain. In *PTC Workshop at the 5th PKDD*, Université de Freiburg

HELMA C., GOTTMANN E., KRAMER S. (2000) Knowledge Discovery and Data Mining in Toxicology. *Stat. Methods Med. Res.* **9**: 329-358

HELMA C., KRAMER S. (2003a) A survey of the Predictive Toxicology Challenge 2000-2001. *Bioinformatics* **19**: 1179-1182

HELMA C., KRAMER S., DE RAEDT L. (2003b) The Molecular Feature Miner MolFea. In *Proc. of the Beilstein Workshop 2002*, Beilstein Institut, Frankfurt am Main

[Helma-PredTox]: website offering data and tools for the prediction of toxicological properties: <http://www.predictive-toxicology.org/>

[HIV-Data-Set-1997]: website offering a public dataset of screening results the AIDS Screening Results, (May' 97 Release):

http://dtpws4.ncifcrf.gov/DOCS/AIDS/AIDS_DATA.HTML

- KING R.D., MARCHAND-GENESTE N., ALSBERG B. (2001)
A quantum mechanics based representation of molecules for machine inference.
Electronic Transactions on Artificial Intelligence **5**: 127-142
- KING R.D., WHELAN K.E., JONES F.M., REISER P.G., BRYANT C.H., MUGGLETON S.H., KELL D.B., OLIVER S.G. (2004) Functional genomic hypothesis generation and experimentation by a robot scientist. *Nature* **427**: 247-252
- KRAMER S., LAVRAC N., FLACH P. (2001) Propositionalization approaches to relational data mining. In *Relational Data Mining* (DZEROSKI S., LAVRAC N. Eds)
Springer Verlag, Berlin-Heidelberg-New York
- LANDWEHR N., PASSERINI A., RAEDT L. D., FRASCONI P. (2006) kFOIL: Learning Simple Relational Kernels. In *Proc. of Twenty-First National Conference on Artificial Intelligence (AAAI-06)*, AAAI, Boston
- LIPINSKI C.A., LOMBARDO F., DOMINY, B.W., FEENEY P.J. (2001) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development setting. *Adv. Drug Deliv. Rev.* **46**: 3-26
- [Liu-Yu-2002]: Features selection for data mining: a survey:
<http://www.public.asu.edu/~huanliu/sur-fs02.ps>
- LIU H., YU L. (2005) Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Trans. on Knowledge and Data Engineering* **17**: 1-12
- MARCHAND-GENESTE N., WATSON K.A., ALSBERG B.K., KING R.D. (2002) A new approach to pharmacophore mapping and QSAR analysis using Inductive Logic Programming. Application to thermolysin inhibitors and glycogen phosphorylase b inhibitors. *J. Med. Chem.* **45**: 399-409
- MAYER D., MOTOC I., MARSHALL G. (1987) A unique geometry of the active site of angiotensin-converting enzyme consistent with structure-activities studies.
J. Comput. Aided Mol. Des. **1**: 3-16
- MICHALSKI R.S. (1986) Understanding the nature of learning: Issues and research directions. In *Machine Learning: An Artificial Intelligence Approach*, Vol. II,
Morgan Kaufmann, San Francisco, CA : 3-25
- MITCHELL T. (1997) *Machine learning*. Mc Graw Hill, New York.
- OKADA T. (2003) Characteristic substructures and properties in chemical carcinogens studied by the cascade model. *Bioinformatics* **19**: 1208-1215
- QUINLAN J.R. (1993) *C4.5: Programs for Empirical Learning*.
Morgan Kaufmann, San Francisco, CA
- QUINLAN J.R. (1990) Learning logical definitions from relations.
Machine Learning **5**: 239-266
- RUSSELL S.J., NORVIG P. (2003) *Artificial Intelligence: a modern approach*.
Prentice-Hall, Upper Saddle River, New Jersey

- SEBAG M., ROUVEIROL C. (2000) Resource-bounded Relational Reasoning: Induction and Deduction Through Stochastic Matching. *Machine Learning* **38**: 41-62
- SEBAG M., AZÉ J., LUCAS N. (2003) Impact Studies and Sensitivity Analysis in Medical Data Mining with ROC-based Genetic Learning. In *Proc. of IEEE Int. Conference on Data Mining (ICDM03)*, IEEE Computer Society Press, Melbourne, Floride: 637-640
- SEIFERT M., WOLF K., VITT D. (2003) Virtual high-throughput *in silico* screening. *Biosilico* **1**: 143-149
- SRINIVASAN A., KING R.D., MUGGLETON S. (1999) The role of background knowledge: using a problem from chemistry to examine the performance of an ILP program. In *Technical Report PRGTR -08-99*, Oxford University Computing Laboratory, Oxford
- STERNBERG M.J.E., MUGGLETON S.H. (2003) Structure-activity relationships (SAR) and pharmacophore discovery using inductive logic programming (ILP). *QSAR and Combinatorial Science* **22**: 527-532
- TODESCHINI R., CONSONNI V. (2000) *Handbook of Molecular Descriptors* (MANNHOLD R., KUBINYI H., TIMMERMAN H. Eds.) Wiley-VCH, Weinheim
- VAPNIK V. (1998) *The Statistical Learning Theory*. John Wiley, New York
- WEININGER D. (1988) SMILES: a chemical language and information system. 1. Introduction and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **28**: 31-36
- WITTEN I.H., EIBE F. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd Edition, Morgan Kaufmann, San Francisco, CA
- [Weka]: site Weka: <http://www.cs.waikato.ac.nz/~ml/weka/index.html>

VIRTUAL SCREENING BY MOLECULAR DOCKING

Didier ROGNAN

16.1. INTRODUCTION

The growing number of genomic targets of therapeutic interest (HOPKINS and GROOM, 2002) and macromolecules (proteins, nucleic acids) for which a three-dimensional structure (3D) is available (BERMAN *et al.*, 2000) makes the techniques of virtual screening increasingly attractive for projects aiming to identify bioactive molecules (WALTERS *et al.*, 1998; LENGAUER *et al.*, 2004). By **virtual screening**, we mean all **computational research undertaken using molecular data banks to aid the selection of molecules**. The search can be carried out with different types of constraints (physicochemical descriptors, pharmacophore, topology of an active site) and must end with the selection of a low percentage (1-2%) of molecules present in the initial **chemical library (ligand data bank)**. Here we are going to deal with the diverse integrated strategies capable of bringing success in virtual screening based on the 3D structure of the protein target.

16.2. THE 3 STEPS IN VIRTUAL SCREENING

All virtual screening can be broken down into three steps of equal importance:

- › setting up the initial chemical library,
- › the screening itself,
- › the selection of a list of virtual hits.

It is of note that errors in each of these three steps will have significant consequences which generally manifest in an increase in the rate of false positives and false negatives. It is important therefore to be very careful at each step.

16.2.1. PREPARATION OF A CHEMICAL LIBRARY

Choice of chemical library

Two types of chemical library can be used in virtual screening: physical collections (available molecules) and virtual collections (molecules needing to be synthesised).

Every pharmaceutical company now has at its disposal proprietary collections in physical (microplates) and electronic forms containing up to several million molecules. Furthermore, diverse collections are available commercially (BAURIN *et al.*, 2004) and constitute a major source of bioactive molecules notably for the academic world.

In this chapter we compare commercial and public libraries of molecules, in an attempt to present a more complete picture. For an insight into the commercially available screens, the reader can refer to quite an exhaustive list described on the website <http://www.warr.com/links.html#chemlib>. The reader is also invited to consult the documentation relating to each of these chemical libraries for a description of their usefulness and originality.

A global analysis of these collections shows that the percentage of drug candidates from each of these screening sources is moderate (CHARIFSON and WALTERS, 2002; *fig. 16.1a*). What is their molecular diversity? The diversity of these collections is poor but highly dependent on their origin. The collections arising from combinatorial chemistry are vast but not so diverse in their molecular scaffolds (*fig. 16.1b*). Only a small number of collections (such as the French National Chemical Library – *Chimiothèque Nationale*; see chapter 2) offers a larger size/diversity ratio. It is of note that this kind of libraries also tends to be most similar to those collections that harbour proven drug candidates (SHERIDAN and SHPUNGIN, 2004). The choice of a chemical library is therefore crucial and depends on the project. Rather than selecting a single library, it is more appropriate to choose from among the possible sources the most diverse molecules in terms of molecular scaffolds and to avoid redundancy as much as possible.

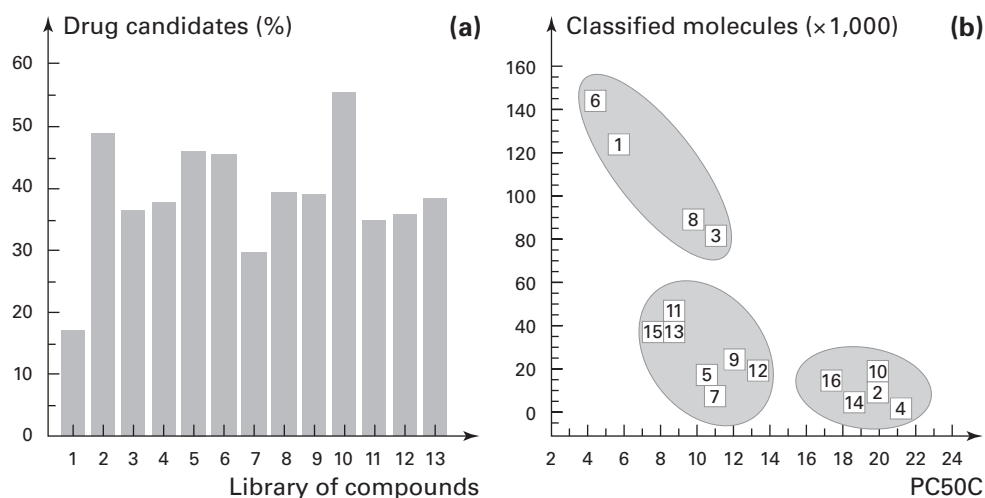


Fig. 16.1 - Analysis of chemical library screens

(a) Potential for finding drug candidates in commercial (1 to 7 and 9 to 13) and public (8: *Chimiothèque Nationale*) library screens - **(b)** Diversity of molecular scaffolds: percentage of scaffolds covering 50% of drug candidates (PC50C metric) from a chemical library (1 to 3 and 5 to 16: commercial chemical libraries; 4: *Chimiothèque Nationale*)

Filtering and final preparation of the chemical library

In order to select only the molecules of interest, it is common to filter the chemical library with a certain number of descriptors (fig. 16.2) in order to keep only potential drug candidates (CHARIFSON and WALTERS, 2002).

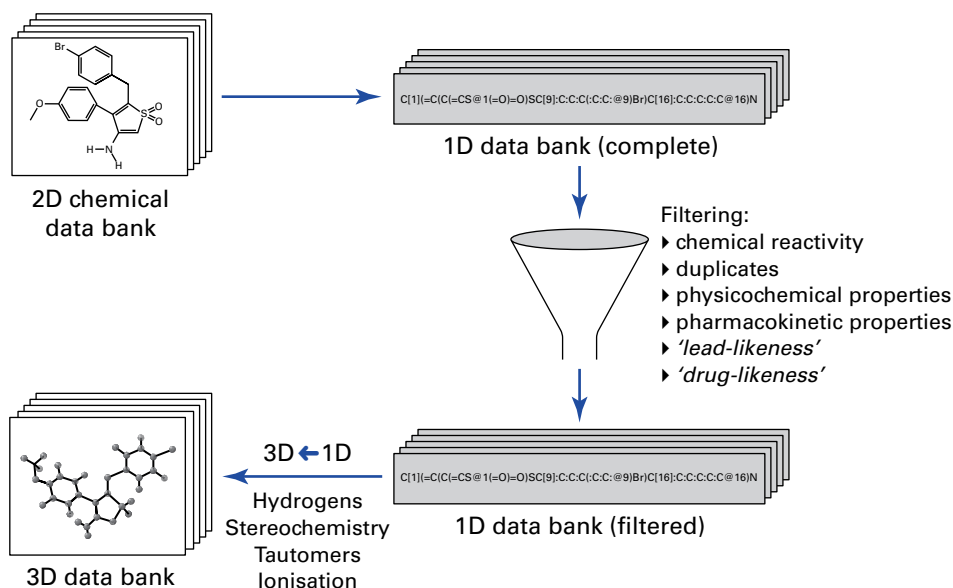


Fig. 16.2 - Different phases in the preparation of a chemical library

The terms *lead-likeness* and *drug-likeness* designate the properties generally accepted (according to statistical models) to be leads or drugs.

These filters are designed to eliminate the chemically reactive, toxic or metabolisable molecules, and those displaying inadequate physicochemical properties (e.g. not conforming to LIPINSKI's rule; see chapter 8) likely to be so-called permissive hits, or capable of interfering with the experimental screening test (e.g. fluorescent molecules; see chapter 3). In any event, it is important to adapt the filtering level to the project specifications. This filtering should be strict if one wishes to identify hits for a target already well studied in the past and for which a number of ligands exist. The filtering will be smoother if the aim of virtual screening is to discover the first putative ligands for an orphan target (*i.e.* having no known ligand). The last step in the preparation is therefore to convert this 1D format into a 3D format by including a complete atom representation. We suggest below some links to diverse tools for the manipulation of chemical libraries (example 16.1).

Example 16.1 - principal tools for the design/management of virtual chemical libraries

Name	Editor	Internet Site	Function
ISIS/base	MDL	http://www.mdli.com	archiving
ChemOffice	Cambridge Soft	http://www.cambridgesoft.com	archiving
Filter	Openeyes	http://www.eyesopen.com	filtering

Cliff	Molecular Network	http://www.mol-net.de	filtering
Pipeline Pilot	SciTegic	http://www.scitegic.com	filtering and automation of procedures
Marvin	Chemaxon	http://www.chemaxon.com	archiving
Ligprep	Schrodinger	http://www.schrodinger.com	filtering

16.2.2. SCREENING BY HIGH-THROUGHPUT DOCKING

High-throughput docking consists of predicting both the active conformation and the relative orientation of each molecule in the selected chemical library relative to the target of interest. Broadly speaking, the search focusses on the **active site** however it may have been determined experimentally (by site-directed mutagenesis for example). It is important to take into account at this point the throughput that needs to be achieved and it is necessary to consider the best compromise between speed and precision. As a general rule, high-throughput docking requires a rate bordering on 1-2 minutes/ligand. A number of docking programs are available (TAYLOR *et al.*, 2002; example 16.2).

Example 16.2 - main programs for molecular docking

Name	Editor	Internet Site
AutoDock	Scripps	http://www.scripps.edu/mb/olson/doc/autodock/
Dock	UCSF	http://dock.compbio.ucsf.edu/
FlexX	BioSolveIT	http://www.biosolveit.de/FlexX/
Fred	OpenEyes	http://www.eyesopen.com/products/applications/fred.html
Glide	Schrödinger	http://www.schrodinger.com/Products/glide.html
Gold	CCDC	http://www.ccdc.cam.ac.uk/products/life_sciences/gold/
ICM	Molsoft	http://www.molsoft.com/products.html
LigandFit	Accelrys	http://www.accelrys.com/cerius2/c2ligandfit.html
Surflex	Biopharmics	http://www.biopharmics.com/products.html

These methods all use the principle of steric complementarity (Dock, Fred) or molecular interactions (AutoDock, FlexX, Glide, Gold, ICM, LigandFit, Surflex) to place a ligand in the target's active site. In general, the protein is considered to be rigid whereas the ligand's flexibility is relatively well accounted for up to about 15 rotatable bonds.

Three principles are routinely applied when dealing with a ligand's flexibility:

- › a group of ligand conformations is first calculated and these are docked as a rigid body into the site (e.g. Fred),
- › the ligand is constructed incrementally, fragment by fragment (e.g. Dock, FlexX, Glide, Surflex),
- › a more or less exhaustive conformational analysis of the ligand is conducted so as to generate the most favourable conformations for docking (e.g. ICM, Gold, LigandFit).

In general, several ligand dockings are generated and classified by order of decreasing probability according to a scoring function (GOHLKE and KLEBE, 2001) which tries at best to approximate the free energy of the bond between the ligand and its target (*i.e.* its affinity). Or more simply it classifies the chemical library molecules by their interaction energy with the site on the protein target. The precision of molecular docking tools can be evaluated by comparing experimental predictions and solutions with a representative set of protein-ligand complexes from the **Protein Data Bank** (<http://www.rcsb.org/pdb/>). Experience shows that when considering the 30 most probable solutions, a ligand is generally well docked in about 75% of cases (KELLENBERGER *et al.*, 2004). The 'best solution' (closest to the experimental solution) is not always the one predicted as the most probable by the scoring function (only in about 50% of cases) which hugely complicates the predictive analysis of docking solutions (KELLENBERGER *et al.*, 2004). There are many reasons to explain these imperfections, which may prove to be more or less complicated to overcome (example 16.3).

Example 16.3 - main causes of error during molecular docking

Cause of error	Treatment
Active site devoid of cavities	impossible?
Flexibility of the protein	very difficult
Influence of water	very difficult
Imprecision of the scoring functions	very difficult
Uncommon interactions	difficult
Flexibility of the ligand	difficult
Pseudosymmetry of the ligand	difficult
Bad set of coordinates (protein)	easy
Bad atom types (ligand, protein)	easy

This is the reason why molecular docking remains a difficult technology to put into practice because it must be applied and constantly adapted according to the context of the project. It is nevertheless possible to give some guidelines depending on the type of protein, the active site and the ligand(s) to be docked (KELLENBERGER *et al.*, 2004).

When applied to virtual chemical-library screening, molecular docking must not only supply the most precise conformations possible for each ligand in the library, but also be able to classify ligands in decreasing order of predicted affinity so as to enable selection of the hits of interest. This remains one of the major challenges of theoretical chemistry, as it is best to predict with a precision and speed compatible with the screening rate the enthalpic (an easy task) and entropic (a much more difficult task) components of the free energy of binding for each ligand (GOHLKE and KLEBE, 2001). Numerous methods for the prediction of free energy exist. Their precision is however related to the rate with which they can be applied ([fig. 16.3](#)). Thermodynamic methods give a relatively high precision but are only applicable to a ligand pair (prediction of free-energy differences). On the contrary, empirical

functions can be used at higher throughput ($> 100,000$ molecules) but with only average precision (on the order of 7 kJ/mol, or in terms of affinity, one and a half pK units).

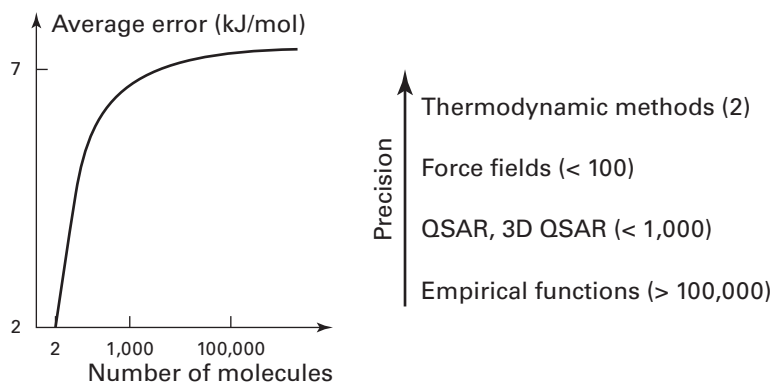


Fig. 16.3 - Methods for predicting a ligand's free energy of binding (affinity)

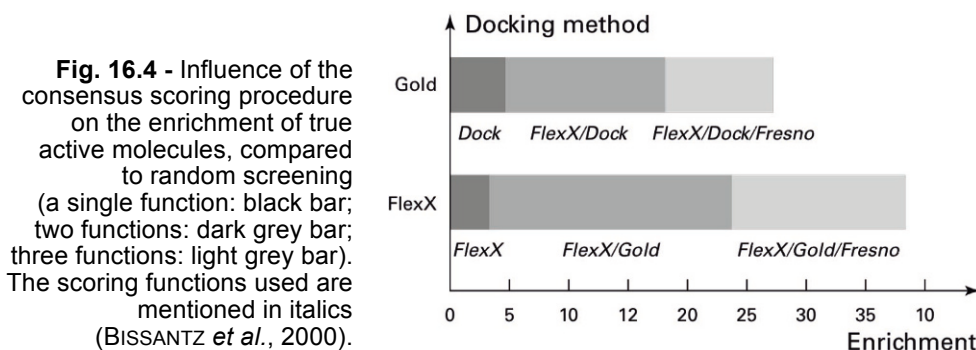
Many studies show that it is impossible to predict with any precision the affinity of chemically diverse ligands (FERARRA *et al.*, 2004). **It is reasonable to hope to be able to discriminate between ligands of nanomolar, micromolar and millimolar affinity**, which is probably sufficient to identify hits in a chemical library but insufficient to optimise them. From the moment the hit selection is made on the basis of docking scores, whatever these may be, **virtual screening by molecular docking will inevitably therefore produce many false positives and above all false negatives, which clearly distinguishes this method from experimental high-throughput screening, which identifies more exhaustively the true positives.**

16.2.3. POST-PROCESSING OF THE DATA

Accepting the fact that the scoring functions are imperfect, the best strategy to increase the rate of true positives during virtual screening consists of trying to detect the false positives. This is only possible by analysing the screening output with an additional chemoinformatics method. Several solutions are possible. The simplest one involves scoring once again the dockings obtained with the help of scoring functions different from that used during the docking. Each function has its imperfections and so through a consensual analysis (CHARIFSON *et al.*, 1999) false positives are detected by identifying the hits not in common to two or three functions relying on different physicochemical principles (fig. 16.4).

The selection of hits scored among the top 5% of different functions allows the final selection to be enriched in true positives (CHARIFSON *et al.*, 1999; BISSANTZ *et al.*, 2000). This method offers the advantage of adjusting a screening strategy with respect to the known experimental data. It suffices to prepare a test chemical library or a small number of true active molecules (about ten, for example) and to mix this with a large number of supposedly inactive molecules (e.g. a thousand),

then to dock the chemical library by means of diverse docking tools, and to rescore the docking obtained with different scoring functions. A systematic analysis of the enrichment of true positives is done by calculating the number of true active molecules in the many selection lists determined by single or multiple scoring. The screening strategy (docking/scoring) giving the best enrichment can then be applied to the full-scale screen. Despite these advantages, this technique cannot be applied in the absence of experimental data (knowledge of several chemically diverse true, active compounds). In this instance, it is necessary to apply more general strategies for eliminating false positives: detection of ligands insufficiently embedded (STAHL and BÖHM, 1998); refinement by energy minimisation of the docking conformations (TAYLOR *et al.* 2003); consensus docking using diverse tools (PAUL and ROGNAN, 2002); docking onto multiple conformations of the target (VIGERS and RIZZI, 2004); rescoring multiple dockings (KONTOYIANNI *et al.*, 2004). For most, these approaches are quite complicated to set up and are not guaranteed to be widely applicable to many screening projects.



A more simple but efficient post-processing strategy consists of applying a statistical treatment to the molecules in a chemical library grouped in a 'quasi-phylogenetic' manner by molecular scaffold (NICOLAOU *et al.*, 2002). Rather than being interested in the individual scores, it suffices to look at their distribution within homogenous chemical classes. This **enables no longer the molecules but the molecular scaffolds to be picked out, supposedly enriched sufficiently in virtual hits** (fig. 16.5) and thus the identification of false negatives (badly docked and/or scored active molecules).

Regardless of the method, the final selection of molecules to rank for experimental evaluation first includes an examination of the individual 3D interactions of each virtual hit with the receptor as well as a study of the availability of the molecules from the respective suppliers, if a commercial collection was screened. Depending on the time lapse between downloading the electronic catalogue of the chemical library and placing the order, the percentage of molecules becoming unavailable increases significantly (about 25% after three months). Bringing these commercial chemical libraries up to date is therefore an absolute necessity in order to guarantee the maximum availability of the ligands chosen by virtual screening.

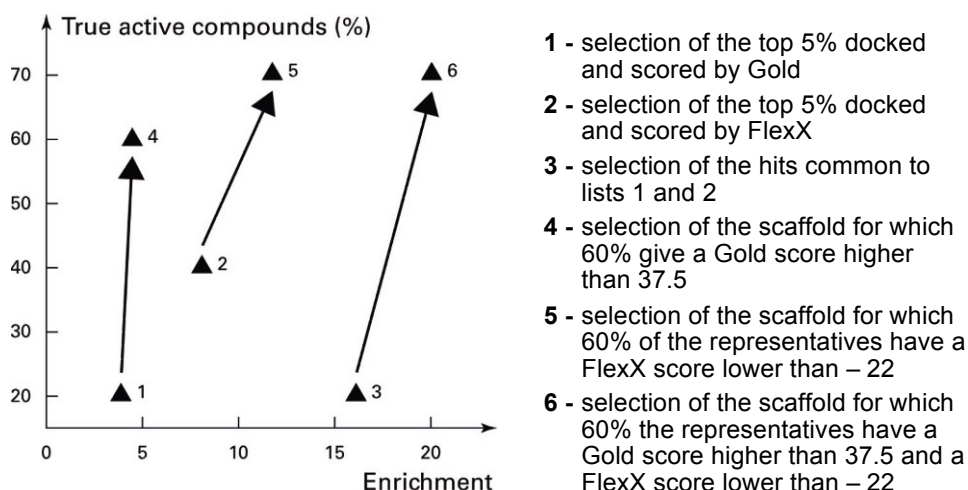


Fig. 16.5 - Influence of the data-analysis strategy on the enrichment of active compounds relative to random screening, from the same docking dataset (10 antagonists of the vasopressin V1a receptor seeded in a database of 1,000 molecules; BISSANTZ *et al.*, 2003). The molecular scaffolds were calculated with the software ClassPharmer (Simulations Plus, Lancaster, USA). The arrows indicate the recorded gain in the selection of true active molecules by analysing the molecular scaffolds (singletons excluded).

16.3. SOME SUCCESSES WITH VIRTUAL SCREENING BY DOCKING

Many examples of successful virtual screens have been described in the last few years (example 16.4). Based on the high-resolution crystal structures of proteins or nucleic acids, it is generally possible to obtain experimentally validated hit rates of around 20-30%, using chemical libraries of varying sizes and diversity, but always filtered beforehand as indicated above.

Example 16.4 - examples of successes in virtual chemical-library screening

Molecular target	Chemical library	Size	Hit rate	Reference
Bcl-2	NCI	207 K	20%	ENYEDY <i>et al.</i> , 2001
HCA-II	Maybridge/Leadquest	90 K	61%	GRÜNEBERG <i>et al.</i> , 2001
Er α	ACD-Screen	1500 K	72%	SHAPIRA <i>et al.</i> , 2001
GAPDH	Comb. Lib.	2 K	17%	BRESSI <i>et al.</i> , 2001
PTP1B	Pharmacia	230 K	35%	DOMAN <i>et al.</i> , 2002
β -Lactamase	ACD	230 K	5%	POWERS <i>et al.</i> , 2002
BCR-ABL	Chemdiv	200 K	26%	PENG <i>et al.</i> , 2003
XIAP	Chinese Nat. Lib.	8 K	14%	NIKOLOVSKA <i>et al.</i> , 2004
Aldose reductase	ACD	260 K	55%	KRAEMER <i>et al.</i> , 2004
Chk-1 kinase	AstraZeneca	550 K	35%	LYNE <i>et al.</i> , 2004
Ribosomal A-site	Vernalis Collection	900 K	26%	FOLOPPE <i>et al.</i> , 2004

Virtual screening with homology models remains more difficult because of the uncertainty generated by the model itself (OSHIRO *et al.*, 2004). Notable progress has however been documented, particularly in the field of G protein-coupled receptors, where several retrospective (BISSANTZ *et al.*, 2003; GOULDSON *et al.*, 2004) and prospective (BECKER *et al.*, 2004; EVERS and KLEBE, 2004) studies have shown that it is possible to enrich significantly the hit lists with true active molecules (example 16.5). It is important nevertheless to adjust well the 3D model of the receptor as a function of the ligand type sought (agonist, inverse agonist, neutral antagonist).

Example 16.5 - state of the art of what is possible in virtual screening by docking

What is possible:

- ▶ screen around 50,000 molecules/day
- ▶ discrimination of the true active compounds from molecules chosen at random
- ▶ obtaining hit rates of 10-30%
- ▶ identification of around 50% of the true active molecules
- ▶ selectivity profiling of different targets

What remains difficult:

- ▶ prediction of the exact orientation of the ligand
- ▶ prediction of the exact affinity of the ligand
- ▶ discrimination the true active molecules from the chemically similar inactive ones
- ▶ identification of 100% of the true active molecules
- ▶ accounting for the flexibility of the target

16.4. CONCLUSION

Virtual chemical library screening by docking has become a method routinely used in chemoinformatics to identify ligands for targets of therapeutic interest. It is necessary to remember that this technology is very sensitive to the 3D coordinates of the target and in spite of everything generates numerous false negatives. Just as important as the screening itself are the phases of chemical library preparation and searching through the results to detect potential false positives so as to improve the hit rate, which can reach 30% in favourable cases. Rather than focussing on the hit rate, it is more interesting to consider the number of new chemotypes in the ligands identified and validated by screening. With this in mind, this tool is a natural complement to the medicinal chemist for suggesting molecular scaffolds likely to lead quickly to focussed chemical libraries of greater use. The progress yet to be made in the prediction of ADME/Tox (**A**bsorption, **D**istribution, **M**etabolism, **E**xcretion and **T**oxicity) properties should allow significant enhancement of the potential of this chemoinformatics tool.

16.5. REFERENCES

- BAURIN N., BAKER R., RICHARDSON C., CHEN I., FOLOPPE N., POTTER A., JORDAN A., ROUGHLEY S., PARRATT M., GREANEY P., MORLEY D., HUBBARD R.E. (2004) Drug-like annotation and duplicate analysis of a 23-supplier chemical database totalling 2.7 million compounds. *J. Chem. Inf. Comput. Sci.* **44**: 643-651
- BECKER O.M., MARANTZ Y., SHACHAM S., INBAL B., HEIFETZ A., KALID O., BAR-HAIM S., WARSHAVIAK D., FICHMAN M., NOIMAN S. (2004) G protein-coupled receptors: *in silico* drug discovery in 3D. *Proc. Natl Acad. Sci. USA* **101**: 11304-11309
- BERMAN H.M., WESTBROOK J., FENG Z., GILLILAND G., BHAT T.N., WEISSIG H., SHINDYALOV I.N., BOURNE P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.* **28**: 235-242
- BISSANTZ C., FOLKERS G., ROGNAN D. (2000) Protein-based virtual screening of chemical databases. I. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **43**: 4759-4767
- BISSANTZ C., BERNARD P., HIBERT M., ROGNAN D. (2003) Protein-based virtual screening of chemical databases. II. Are homology models of G Protein-Coupled Receptors suitable targets? *Proteins* **50**: 5-25
- BRESSI, J.C., VERLINDE C.L., ARONOV A.M., SHAW M.L., SHIN S.S., NGUYEN L.N., SURESH S., BUCKNER F.S., VAN VOORHIS W.C., KUNTZ I.D., HOL W.G., GELB M.H. (2001) Adenosine analogues as selective inhibitors of glyceraldehyde-3-phosphate dehydrogenase of Trypanosomatidae via structure-based drug design. *J. Med. Chem.* **44**: 2080-2093
- CHARIFSON P.S., CORKERY J.J., MURCKO M.A., WALTERS W.P. (1999) Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **42**: 5100-5109
- CHARIFSON P.S., WALTERS W.P. (2002) Filtering databases and chemical libraries. *J. Comput. Aided Mol. Des.* **16**: 311-323
- DOMAN T.N., MCGOVERN S.L., WITHERBEE B.J., KASTEN T.P., KURUMBAIL R., STALLINGS W.C., CONNOLLY D.T., SHOICHET B.K. (2002) Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B. *J. Med. Chem.* **45**: 2213-2221
- ENYEDY I.J., LING Y., NACRO K., TOMITA Y., WU X., CAO Y., GUO R., LI B., ZHU X., HUANG Y., LONG Y.Q., ROLLER P.P., YANG D., WANG S. (2001) Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening. *J. Med. Chem.* **44**: 4313-4324
- EVERS A., KLEBE G. (2004) Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. *J. Med. Chem.* **47**: 5381-5392
- FERRARA P., GOLHKE H., PRICE D.J., KLEBE G., BROOKS C.L. (2004) Assessing scoring functions for protein-ligand interactions. *J. Med. Chem.* **47**: 3032-3047

- FOLOPPE N., CHEN I.J., DAVIS B., HOLD A., MORLEY D., HOWES R. (2004) A structure-base strategy to identify new molecular scaffolds targeting the bacterial ribosome A-site. *Bioorg. Med. Chem.* **12**: 935-947
- GOHLKE H., KLEBE G. (2001) Statistical potentials and scoring functions applied to protein-ligand binding. *Curr. Opin. Struct. Biol.* **11**: 231-235
- GOULDSON P.R., KIDLEY N.J., BYWATER R.P., PSAROUDAKIS G., BROOKS H.D., DIAZ C., SHIRE D., REYNOLDS C.A. (2004) Toward the active conformations of rhodopsin and the beta2-adrenergic receptor. *Proteins* **56**: 67-84
- GRÜNEBERG S., WENDT B., KLEBE G. (2001) Subnanomolar inhibitors from computer screening: a model study using human carbonic anhydrase II. *Angew Chem. Int. Ed. Engl.* **40**: 389-393
- HOPKINS A.L., GROOM C.R. (2002) The druggable genome. *Nat. Rev. Drug Discov.* **1**: 727-30
- KELLENBERGER E., RODRIGO J., MULLER P., ROGNAN D. (2004) Comparative evaluation of eight docking tools for docking and virtual screening accuracy. *Proteins* **57**: 225-242
- KONTOYIANNI M., MCCLELLAN L.M., SOKOL G.S. (2004) Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **47**: 558-565
- KRAEMER O., HAZMANN I., PODJARNY A.D., KLEBE G. (2004) Virtual screening for inhibitors of human aldose reductase. *Proteins* **55**: 814-823
- LENGAUER T., LEMMEN C., RAREY M., ZIMMERMANN M. (2004) Novel technologies for virtual screening. *Drug Discov. Today* **9**: 27-34
- LYNE P.D., KENNY P.W., COSGROVE D.A., DENG C., ZABLUDOFF S., VENDOLOSKI J.J., ASHWELL S. (2004) Identification of compounds with nanomolar binding affinity for checkpoint kinase 1 using knowledge-based virtual screening. *J. Med. Chem.* **47**: 1962-68
- NICOLAOU C.A., TAMURA S.Y., KELLEY B.P., BASSETT S.I., NUTT R.F. (2002) Analysis of large screening data sets via adaptively grown phylogenetic-like trees. *J. Chem. Inf. Comput. Sci.* **42**: 1069-1079
- NIKOLOVSKA-COLESKA Z., XU L., HU Z., TOMITA Y., LI P., ROLLER P.P., WANG R., FANG X., GUO R., ZHANG M., LIPPMAN M.E., YANG D., WANG S. (2004) Discovery of embelin as a cell-permeable, small-molecular weight inhibitor of XIAP through structure-based computational screening of a traditional herbal medicine three-dimensional structure database. *J. Med. Chem.* **47**: 2430-2440
- OSHIRO C., BRADLEY E.K., EKSTEROWICZ J., EVENSEN E., LAMB M.L., LANCTOT J.K., PUTTA S., STANTON R., GROOTENHUIS P.D. (2004) Performance of 3D-database molecular docking studies into homology models. *J. Med. Chem.* **47**: 764-767
- PAUL N., ROGNAN D. (2002) ConsDock: a new program for the consensus analysis of protein-ligand interactions. *Proteins* **47**: 521-533
- PENG H., HUANG N., QI J., XIE P., XU C., WANG J., WANG C. (2003) Identification of novel inhibitors of BCR-ABL tyrosine kinase via virtual screening. *Bioorg. Med. Chem. Lett.* **13**: 3693-3699

- POWERS R.A., MORANDI F., SHOICHET B.K. (2002) Structure-based discovery of a novel, noncovalent inhibitor of AmpC beta-lactamase. *Structure* **10**: 1013-1023
- SHERIDAN R.P., SHPUNGIN J. (2004) Calculating similarities between biological activities in the MDL Drug Data Report database. *J. Chem. Inf. Comput. Sci.* **44**: 727-740
- STAHL M., BÖHM H.J. (1998) Development of filter functions for protein-ligand docking. *J. Mol. Graph. Model.* **16**: 121-132
- TAYLOR R.D., JEWSBURY P.J., ESSEX J.W. (2002) A review of protein-small molecule docking methods. *J. Comput. Aided Mol. Des.* **16**: 151-166
- TAYLOR R.D., JEWSBURY P.J., ESSEX J.W. (2003) FDS: flexible ligand and receptor docking with a continuum solvent model and soft-core energy function. *J. Comput. Chem.* **24**: 1637-1656
- VIGERS G.P., RIZZI J.P. (2004) Multiple active site corrections for docking and virtual screening. *J. Med. Chem.* **47**: 80-89
- WALTERS W.P., STAHL M.T., MURCKO M.A. (1998) Virtual screening – an overview. *Drug Discov. Today* **3**: 160-178
- WASZKOWYCZ B., PERLINS T.D.J., SYKES R.A., LI J. (2001) Large-scale virtual screening for discovery leads in the postgenomic era. *IBM Sys. J.* **40**: 361-376

APPENDIX

BRIDGING PAST AND FUTURE?

Chapter 17

BIODIVERSITY AS A SOURCE OF SMALL MOLECULES FOR PHARMACOLOGICAL SCREENING: LIBRARIES OF PLANT EXTRACTS

*Françoise GUERITTE, Thierry SEVENET,
Marc LITAUDON, Vincent DUMONTET*

17.1. INTRODUCTION

The term ‘biodiversity’ refers to the diversity of living organisms. This diversity of Life is represented as trees (called ‘taxonomic trees’) following the classification principles first proposed by Aristotle, then rigorously put forward by Linnaeus and connected to natural evolution by Darwin (in neo-Darwinian terms, trees are then called ‘phylogenetic trees’). Beyond the unifying chemical features that characterise living entities (nucleotides, amino acids, sugars, simple lipids etc.), some important branches in the Tree of Life – like plants, marine invertebrates and algae, insects, fungi and bacteria etc. – are known to be sources of innumerable drugs and bioactive molecules. The exploration of this biodiversity was initiated in prehistoric times and is still considered a mine for the future. To allow access to libraries of extracts sampled in this biodiversity, a methodology has been designed following the model defined originally for single-compound chemical libraries. Thus ‘extract libraries’ have been developed to serve biological screening on various targets. There are far fewer extract-libraries than chemical libraries. The positive results obtained from these screenings do not straightforwardly allow the identification of a bioactive molecule, since extracts are mixtures of molecules, but they can orientate research projects towards the discovery of novel active compounds that can be potential drug leads.

The development of extract libraries is an important connection between traditional pharmacopeia and modern high-throughput technologies and approaches. Inquiries into folk uses were the source of the first medicines. Since very ancient times, humans (from hunters and gatherers to farmers) have been trying to use resources in their environment to feed, cure and also to poison. Ancient written records are found in many civilizations (clay tablets of Mesopotamia, Ebers Papyrus from

Egypt, Chinese Pen t'saos). The first chemical studies of the Plant Kingdom (pharmacognosy: the study of medicines derived from natural sources) were pioneered in France: in the XIXth century, pharmacists were able to isolate pure bioactive products; however their chemical structures were determined one century later. DEROSNE purified narcotine and analgesic morphine from opium, the thick latex of poppy (1803); PELLETIER and CAVENTOU isolated strychnine from *Strychnos* in 1820 and antimalarial quinine from Peruvian *Cinchona*; LEROUX isolated salicin, an antipyretic glycoside, in 1930 from the trunk bark of *Salix* spp, a common tree that 'grows in water and never catches cold'. Cardiotoxic digitalin was crystallised from *Digitalis purpurea* by NATIVELLE in 1868 and colchicine from *Colchicum autumnale* by HOUDÉ in 1884. The tremendous development of chemistry in the XXth century allowed, after structural elucidation of the active principles, the synthesis of analogues, which were more active, less toxic and easier to produce. The first achievement in that field was the preparation by FOURNEAU, in 1903, of the synthetic local anaesthetic, stovaine, modelled on the natural alkaloid cocaine.

Until the 1990s, research into natural products was essentially oriented by chemotaxonomic guidelines (alkaloids from Apocynaceae and Rutaceae, acetogenins from Annonaceae, saponins from Sapindaceae and Symplocaceae). Facing the current need for new medicines and for chemogenomic tools, a careful inventory of the biological activity of plant extracts, lichens and marine organisms would be invaluable, making use of automated extraction and fractionation technologies and automated biological screening. New strategies to find novel bioactive molecules from extract libraries and particularly from plant-extract libraries have been initiated in a series of research centers like the Institute of Natural Products Chemistry, (Institut de Chimie des Substances Naturelles, ICSN), CNRS (Gif-sur-Yvette, France), the experience from which has been used to write the present chapter.

If we take into account the number of living organisms in the Plant Kingdom (about 300,000 species), the search for new medicines requires the broadest screening capacity. For example, the screen set up in the sixties, by the United States Department of Agriculture and the National Cancer Institute cooperative program, to evaluate the potential anticancer activity of more than 35,000 plants has resulted in the discovery of few but key lead compounds used as therapeutic agents such as vinblastine and taxol. Chemical studies of vinblastine and taxol then led to the discovery of Navelbine[®] and Taxotere[®] respectively, at the Institute of Natural Products Chemistry. Automated technologies provide solutions to generate rapidly and efficiently such a biological inventory of plant biodiversity.

In this chapter, we describe how the systematic chemical exploration of biodiversity can be put into practice. As detailed through a series of examples, and by contrast with other works introduced in this book, this gigantic task requires the collaboration of scientists from multiple disciplines and backgrounds, and the unprecedented cooperation of countries, some providing their natural landscape as a mine of biodiversity, others providing their technologies as mining tools.

17.2. PLANT BIODIVERSITY AND NORTH-SOUTH CO-DEVELOPMENT

The highest levels of biodiversity observed in the Plant Kingdom are encountered in tropical and equatorial areas. Regions in central and eastern Africa, south-eastern Asia, the Pacific islands or southern America are the richest. Some regions have unique gems, like Madagascar where plants reach 75% of endemism. With few exceptions, countries in these parts of the world have no biodiversity protective policy or real means to fight against 'biopirates'. Since the adoption of the Biodiversity Convention in Rio de Janeiro in 1992, the developing countries are internationally protected by a set of rules enacted in a series of agreements such as the Manila Declaration, the Malacca Agreement, the Bukit Tinggi Declaration and the Phuket Agreement. Following these agreements, plants growing in developing countries cannot be collected without the consent of local partners, and without their benefitting academically and financially. If any scientific results come out of bioscreening, the original country where samples were collected should be associated to any related benefits.

In Europe, national research institutions have independently signed agreements with governmental or academic institutions from countries where plants are collected. Programs of systematic prospecting and collections have been established, for instance between France (Institute of Natural Products Chemistry) and Malaysia, Vietnam, Madagascar, Uganda (fig. 17.1). All of these countries were willing to develop research programs on their floras, by collaborating through missions, short stays, theses, or postdoctoral positions, in the framework of partnerships. Since 1995, about 6,700 plants were collected in the partner countries leading to the development of a unique library of 13,000 extracts.

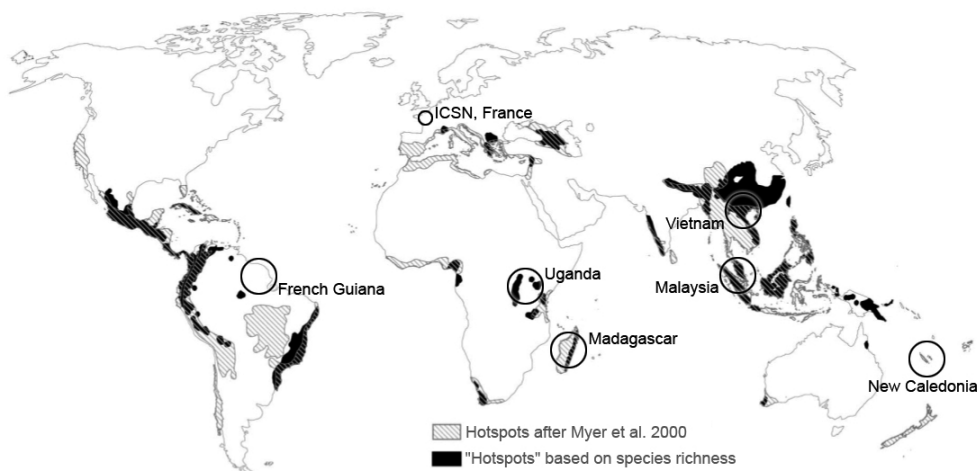


Fig. 17.1 - Cooperation between the Institute of Natural Products Chemistry, CNRS (Gif-sur-Yvette, France) and overseas partners (Hotspots in dark, from MUTKE, 2005)

17.3. PLANT COLLECTION: GUIDELINES

In the current global effort to investigate the biodiversity of the Plant Kingdom, the field collections occur mainly in primary rain forests in tropical and equatorial areas, but also in dry forests (e.g. Madagascar) or mining scrublands (e.g. in New Caledonia). Depending on their relative abundance, their protection status in the International Union for Conservation of Nature's threatened species lists, and local legislation for national parks and reserves, a permit for collection may sometimes be required. In the field, plant collection is managed by a botanist for the primary identification in order to minimize plant duplication and to focus on pre-selected species.

Chemical composition is not uniform in a plant; different parts are therefore collected separately. Common parts are leaves, trunk bark, stems for shrubs or aerial parts for herbs, and, when possible, fruits, flowers or seeds, roots or root bark. The minimum amount of fresh material required for extraction and characterization of the active constituents is one to five kilograms. It corresponds to a small branch of a big tree, a shrub, or a few specimens collected in the surroundings for bushes or more for herbs. For each species collected, at least three herbarium specimens are kept: one for the local herbarium, one for the French Herbarium Museum, and one for the world specialists of the given family, if a more precise identification is needed (fig. 17.2, left). The collection identification number, collected parts, short botanical description, environment, estimation of abundance, drawings (fig. 17.2, right) together with pictures and GPS coordinates are also noted down for each sample. This low-tech, low-throughput registration of collected samples is essential to help identification and recollection.

Guidelines for the selection and collection of plants have evolved to embrace as much chemical diversity as possible. Thirty years ago, at the beginning of the research program in New Caledonia, the selection was only based on the collection of alkaloid-bearing plants, these chemicals being well known for their pharmacological activities. Then, the interest was widened to ethnopharmacological data or observations of plant-insect interactions. Taking into account the miniaturisation and automation of biological assays, a taxonomically oriented collection was preferred. Various types of soil are submitted to the inventory (*i.e.* in New Caledonia: peridotitic, micaceous and calcareous soils). All fertile and original plants could be collected, sometimes with indications of traditional uses (which is often the case in Madagascar or Uganda) or other properties. Thus, in Uganda an additional approach was followed by the CNRS, the National Museum of Natural History and Ugandan authorities based on the unusual plant feeding by chimpanzees that might be related to self-medication (zoopharmacognosy). Before extraction, plants are air-dried, avoiding damage caused by direct sun rays, or spread in homemade drying installations, and turned upside down every day. When dried, the material is crushed to obtain a powder to facilitate solvent extraction.

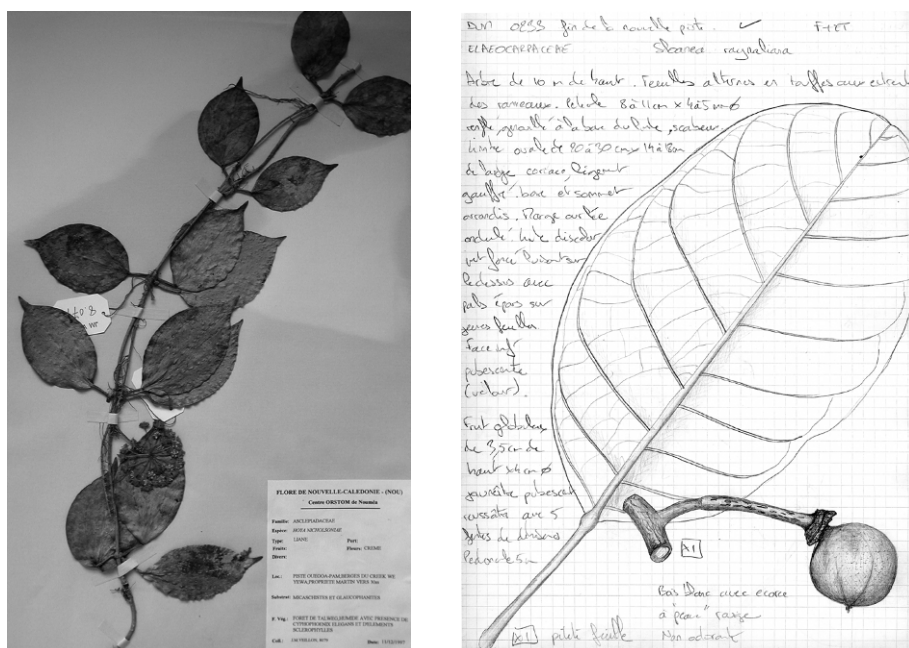


Fig. 17.2 - Herbarium specimen (left), field notes and drawing (right)

17.4. DEVELOPMENT OF A NATURAL-EXTRACT LIBRARY

17.4.1. FROM THE PLANT TO THE PLATE

Based on the example of the natural-extract library of the Institute of Natural Products Chemistry, for each bilateral partnership, about 200 plants are collected every year, giving 400 plant parts, each one being extracted with ethyl acetate. The choice of the extraction solvent was guided by the need to avoid the enrichment in polyphenols and tannins, which often give false positive results in bioactivity screenings. After concentration, extracts become gummy solids or powders. Again, tannins are removed by filtration (on a polyamide cartridge). Then the extracts are dissolved in DMSO (see chapter 1) and the solutions are distributed in 96-well mother plates, which will serve to make the daughter plates submitted for biological analysis. The microplates are gathered and stored at -80°C . At the time of writing, the natural-extract library obtained following this procedure is constituted of more than 13,000 extracts coming from about 6,700 plants.

17.4.2. MANAGEMENT OF THE EXTRACT LIBRARY

A database stores the information relating to the plants that have been collected, the extracts obtained from the different parts of the plants, the corresponding microplates in which the extracts have been distributed and the results of the screening

with biological assays etc. The botanical data including th taxonomical identification with a reference number, the location (GPS coordinates whenever possible) and the date of harvest, the part of the plant collected (bark, leaves, seeds, roots etc.) are included in the database and pictures showing the plants in their natural environment are displayed. On the other side, the reference for the extract is linked with one part of the plant, the type of solvent used, the plate reference and the position in the plate. The data relating to the biological assays (targets, pharmacological domain, unit, results etc.) are uploaded in the database as soon as the tests are completed and validated (fig. 17.3).



Pôle Substances Naturelles

DUM-0505 (Résultats)



Retour

Historique Récolte

Famille :
Genre :
Espèce :
Sous-Espèce :
Variété :
Pays :
Lieu :
Date :
Protection :
Fichier :

PROTEACEAE
Sleumerodendron
austrocaledonicum

Nouvelle-Calédonie
Aoupinié, Nouvelle route d'exploitation, 518 m, UTM 533777/7659537
2005-04-20
non



ICSN : ACHE 10 µg/mL (Résultat : % Inh) - Système nerveux central

Détection

PP	Solvant	Extrait	Position	Composition	Résultats	Date	Description		
Feuille	AcOEt	LN-1960-f	PSN-348	A07	LFSN-0082	0.00	0.00	2007-08-02	IC50 tacrine 93nM
Ecorce	AcOEt	LN-1962-f	PSN-348	B07	LFSN-0082	0.00	10.00	2007-08-02	IC50 tacrine 93nM
Fruit	AcOEt	LN-1964-f	PSN-348	C07	LFSN-0082	0.00	0.00	2007-08-02	IC50 tacrine 93nM
Bois	AcOEt	LN-1966-f	PSN-348	D07	LFSN-0082	10.00	10.00	2007-08-02	IC50 tacrine 93nM

ICSN : Cytotoxicité KB 1 µg/mL (Résultat : % Inh) - Oncologie

Détection

PP	Solvant	Extrait	Position	Composition	Résultats	Date	Description		
Feuille	AcOEt	LN-1960-f	PSN-348	A07	LFSN-0082	18.00	11.00	-3.00	2007-08-30
Ecorce	AcOEt	LN-1962-f	PSN-348	B07	LFSN-0082	49.00	60.00	33.00	2007-08-30
Fruit	AcOEt	LN-1964-f	PSN-348	C07	LFSN-0082	22.00	7.00	23.00	2007-08-30
Bois	AcOEt	LN-1966-f	PSN-348	D07	LFSN-0082	48.00	37.00	38.00	2007-08-30

Fig. 17.3 - Database for the management of the natural-extract library of the Institute of Natural Products Chemistry. For the botanical description: *Famille* = Family, *Genre* = Genus, *Espèce* = Species, *Sous-espèce* = Sub-species, *Variété* = Cultivar, *Pays* = Country, *Lieu* = Collection place. For the recorded bioactivities, assays have been developed in different therapeutic fields: *Système nerveux central* = Central nervous system, *Oncologie* = Oncology.

17.5. STRATEGY FOR FRACTIONATION, EVALUATION AND DEREPLICATION

17.5.1. FRACTIONATION AND DEREPLICATION PROCESS

In the past, the isolation of natural products was the main bottleneck in the natural products field. Tedious purifications were often performed with the main and sole purpose of structural characterisation. Nowadays, the characterisation of the bioactivity of previously known or novel compounds is necessarily driven by the implementation of various bioassays. In this context the rapid identification of already known compounds, a process called dereplication, together with the detection of the presence of novel compounds in extracts is essential.

A rapid and automated preliminary fractionation of the filtered extract constitutes therefore the first important step in the isolation process, as it determines the continuation or interruption of the study, depending on the results of the biological assays. At this point, the objective is either the discovery of novel bioactive compounds with original scaffolds or the recording of an interesting bioactivity for a known compound, which had not been previously tested with the studied target.

Several methods can be applied for fractionating a crude extract. Some methods include simple separations using a silica-phase cartridge with various solvents leading to 3 or 4 fractions, while others are much more sophisticated using the hyphenated techniques of HPLC-SPE-NMR (high-performance liquid chromatography, HPLC, coupled with solid-phase extraction, SPE, and nuclear magnetic resonance, NMR), LC/MS (liquid chromatography, LC, coupled with mass spectrometry, MS), LC/CD (liquid chromatography, LC, coupled with circular dichroism, CD) leading to a large number of fractions or sometimes directly leading to pure compounds in minute quantities in the best case.

As discussed in chapter 3, biological assays require specific miniaturisation developments and some statistical analyses, which cannot be achieved on a one-extract basis. It is therefore necessary to duplicate microplates to test the fractions containing the bioactive compounds in various parallel bioassays. But more often, at this stage, the fractions are still complex and could contain mixed chemical entities, present in low or high amounts. It is important to note that during the preparation of microplates, the fractions are not weighed. They are successively dried and dissolved in a given amount of DMSO in order to get what is called a 'virtual' or 'equivalent' concentration of 10 mg/mL, identical to the concentration of the original 96-well mother microplates. Accurately weighed and filtered extracts are also placed as controls in the microplate, at a 10 mg/mL concentration.

If a bioactivity is measured for a particular extract during the primary biological screening, the results observed in fractionated samples should be consistent. This consistency is particularly meaningful for IC_{50} values, reflecting the efficiency of a

compound in an extract. An example is given for two New-Caledonian plants possessing strong cytotoxicity in three cancer cell lines (table 17.1). An analysis of the results showed the extremely good correlation existing between the IC₅₀ obtained for the crude extract and one active fraction of the standard HPLC fractionation. Acetogenins and flavones were isolated and characterised for *Richella obtusata* (Annonaceae) and *Lethedon microphylla* (Thymeleaceae), respectively.

Table 17.1 - Consistency of the bioactivity detected for crude and fractionated natural extracts

Bioactivity was assayed on 3 cancer cell lines (murine leukaemia P388, lung cancer NCI-H460 and prostate cancer DU-145) for the crude extract (CE) and fractions (F1 to F9) after a standard HPLC fractionation. Bioactivity is given as the IC₅₀ in µg/mL. EtOAc, ethyl acetate.

	F1	F2	F3	F4	F5	F6	F7	F8	F9	CE
<i>Richella obtusata</i> from EtOAc fruit extract										
P388	Not active	Not active	14.2	2.7	0.21	1.1	2.3	Not active	Not active	0.1
NCI-H460	Not active	Not active	7.3	4.7	0.29	1.0	3.5	Not active	Not active	0.2
DU-145	Not active	Not active	7.0	5.8	3.7	5.3	5.6	Not active	Not active	3.6
<i>Lethedon microphylla</i> from EtOAc leaf extract										
P388	7.4	1.1	10.2	Not active	Not active	Not active	Not active	Not active	Not active	1.4
NCI-H460	1.4	0.2	3.2	Not active	Not active	Not active	Not active	Not active	Not active	0.1
DU-145	2.2	0.34	4.8	Not active	Not active	Not active	Not active	Not active	Not active	0.26

The advantage of the automatic procedure is that it requires little handling and offers the possibility of fractionating a large number of extracts in a reasonable time. However three difficulties can arise:

- › bad resolution of peaks, for instance with alkaloids (the addition of trifluoroacetic acid or triethylamine can improve the separation of the basic compounds);
- › precipitation in the injection loop with apolar products;
- › an activity split between several fractions due to the activity of several compounds of different bioactivity.

Once the biological activity has been confirmed in a particular fraction, a third step can be decided leading to the isolation of the active compounds. Classical chromatographic methods are used for this purpose.

LC/MS-coupled methods can provide certain information without the isolation of pure compounds. For example, when applied to the detection of turriane phenolic

compounds in *Kermadecia* extracts, under atmospheric pressure chemical ionisation negative-ion mode, an LC/MS-MS analysis of the quasimolecular peak $[M-H]^-$ of kermadecin A revealed the presence of an ion at $m/z = 369$ corresponding to the loss of a fragment of 108 amu, suggesting the loss of the dimethylpyran ring. In addition, in APCI positive-ion mode, LC/MS-MS analysis of kermadecin A indicated the presence of another ion at $m/z = 297$, resulting from the loss of a fragment supposed to be a 13-carbon aliphatic chain. These fragmentations were systematically observed for compounds containing such moieties (fig. 17.4), an observation that was useful for detecting the presence of this structure in complex mixtures.

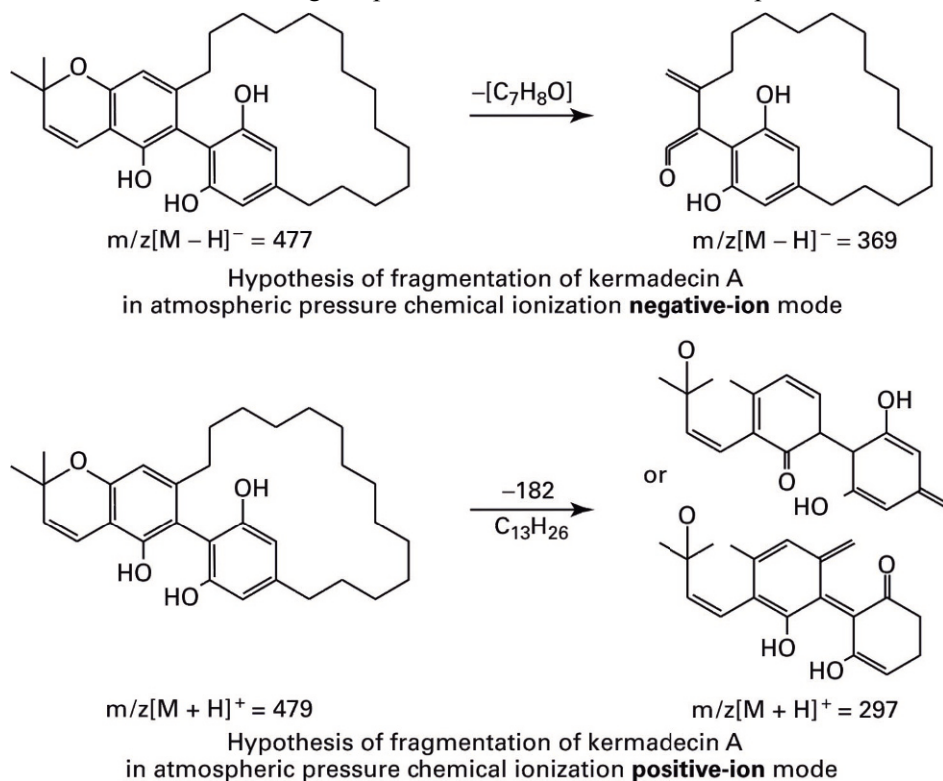


Fig. 17.4 - Characterisation of compounds from extracts by LC/MS (liquid chromatography coupled with mass spectrometry) and fragmentation
 In this example, the combination of mass spectrometry in negative – or positive – ion mode allowed the identification of kermadecin A by the detection of ionised products with specific masses. A mixture containing this compound and treated accordingly in negative- or positive-ion mode will give rise to peaks at the corresponding masses.

17.5.2. SCREENING FOR BIOACTIVITIES

In the last few decades, *in vitro* high-throughput screening (HTS) has been adopted by most of the big pharmaceutical companies as an important tool for the discovery of new drugs. Selection of the most suitable targets is the most crucial issue in this approach (chapters 1 and 2). Current targets are mainly defined in therapeutic

fields like oncology, diabetes, obesity, neurodegenerative diseases and antivirals. In academic groups, screening is conducted on a smaller scale and targets are more related to research projects and the search for biological tools.

The strategy of ICSN, the Institute of Natural Products Chemistry, comprises four steps:

- › biological screening,
- › fractionation,
- › dereplication,
- › isolation of the active constituents.

To carry out rapidly and efficiently a biological inventory of plant biodiversity, biological screening on cellular, protein and enzyme targets have been developed. *In vitro* assays have been miniaturised and automated to allow broad screening. Biological screening is performed either by an academic platform or in the context of a partnership with other academic or industrial groups. For the cytotoxicity screening at the ICSN, a cell line of the nasopharynx adenocarcinoma is routinely used. Other cell lines, including non-tumour cells, can be used to explore the selectivity of the compounds. A collaboration with the Laboratory of Parasitology at the National Museum of Natural History, Paris, allows a systematic focus on antiplasmodial activity using synchronised cultures of *Plasmodium falciparum*, the causative agent in malaria. Biological screening generates numerous 'hits' depending on the concentration chosen for the assays and the threshold value fixed. In some cases such as with antiplasmodial activity, the observed hits are often correlated to cytotoxicity. The goal is to have a good index of selectivity and the remaining question is whether or not to choose slightly cytotoxic extracts as good candidates for antiplasmodial activity.

Screening of enzymatic targets includes acetylcholinesterase inhibition activity (an enzyme from *Torpedo californica*) using colorimetric detection of the 2-nitro-5-thiobenzoate anion. This enzyme is involved in neurodegenerative diseases like ALZHEIMER's disease. Research projects with other public laboratories are exploring the domain of kinase inhibitors.

The domain of agriculture protection is also investigated, as the demand for new herbicides, insecticides and fungicides is considerable. Miniaturised *in vivo* assays with whole target organisms are now possible and are an integral part of the screening process.

17.5.3. SOME RESULTS OBTAINED WITH SPECIFIC TARGETS

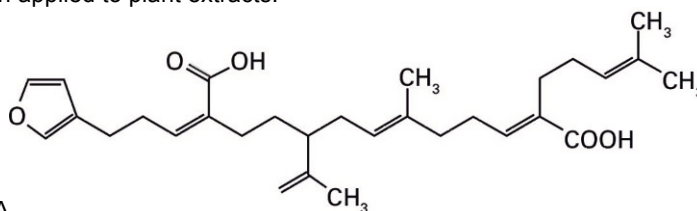
Peroxisome proliferator-activated gamma-receptor

The peroxisome proliferator-activated receptor (PPAR) is a member of the nuclear hormone receptor superfamily of ligand-activated transcription factors that are related to the retinoid, steroid and thyroid hormone receptors. PPAR- γ is an isoform that has attracted attention since it became clear that agonists to this isoform could

play a therapeutic role in diabetes, obesity, inflammation and cancer. The most described endogenous ligands of PPAR- γ is a prostaglandin, and most known ligands of the PPAR family are lipophilic compounds. In an effort to find new naturally occurring PPAR- γ ligands, a series of 1,200 plant extracts, prepared from species belonging to the New Caledonian and Malaysian biodiversity, was screened. The binding affinity of the compounds towards PPAR- γ was evaluated by competition against an isotopically labelled reference compound (rosiglitazone). Several Sapindaceae belonging to the genus *Cupaniopsis*, and several Winteraceae of the genus *Zygogynum* collected in New Caledonia, exhibited strong binding activity (examples 17.1 and 17.2).

Example 17.1 - linear triterpenes from *Cupaniopsis* spp., Sapindaceae from New Caledonia

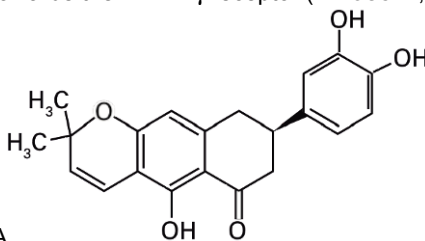
Cupaniopsis trigonocarpa, *C. azantha* and *C. phallacrocarpa* contain linear triterpenes, named cupaniopsins, of which 5 exhibit a strong binding activity towards the PPAR- γ receptor. The most active is cupaniopsin A (BOUSSEROUEL *et al.*). *Cupaniopsis* species are well represented in South East Asia, particularly in New Caledonia, and it was the first time that such linear triterpenes were isolated from the Plant Kingdom, thanks to this new strategy of dereplication applied to plant extracts.



Cupaniopsin A

Example 17.2 - phenyl-3-tetralones from *Zygogynum* spp., Winteraceae from New Caledonia.

The Winteraceae family is considered by botanists to be very primitive. Four species of the genus *Zygogynum*, namely *Z. stipitatum*, *Z. acsmithii*, *Z. pancheri* (2 varieties) and *Z. bailloni*, contain phenyl-3-tetralones named zygolones and analogues, which also exhibit a strong binding activity towards the PPAR- γ receptor (ALLOUCHE, *et al.*, 2008).



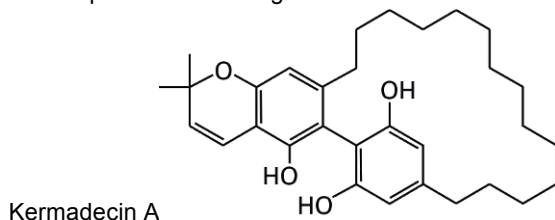
Zygolone A

Cytotoxicity against tumour cells

A number of plant extracts show a significant positive inhibitory activity on an adenocarcinoma tumour cell line. An example is the discovery of cytotoxic molecules from the Proteaceae family (example 17.3).

Example 17.3 - new cytotoxic cyclophanes from *Kermadecia* spp, Proteaceae from New Caledonia.

The study of *Kermadecia elliptica*, an endemic New Caledonian species belonging to the Proteaceae family was carried out following its potent cytotoxicity against adenocarcinoma (KB) cells (JOLLY *et al.*, 2008). A bioassay and LC/MS-directed fractionations of the EtOAc extract provided 8 new cyclophanes, named kermadecins A-H. In an initial step using this strategy the phytochemical investigation of *K. elliptica* led to the isolation of 3 new compounds named kermadecins A-C present in minute quantities in the plant, but clearly present in the cytotoxic fraction 6 (t_R 42 to 50 minutes) of the standard HPLC fractionation. Kermadecins A and B exhibited a strong cytotoxic activity. These compounds belong to the turriane family. Turrianes were first isolated in the 1970s from two closely related Australian Proteaceae, *Grevillea striata* and *G. robusta*. An LC/MS method was then used to detect and to direct further purification leading to the kermadecins D-H. A preliminary LC/APCI-MS (see § 17.5.1) study of kermadecins A-C proved to be particularly efficient due to the low polarity of this kind of compound and the presence of phenols which gave reliable ionisations in both positive and negative ion modes.

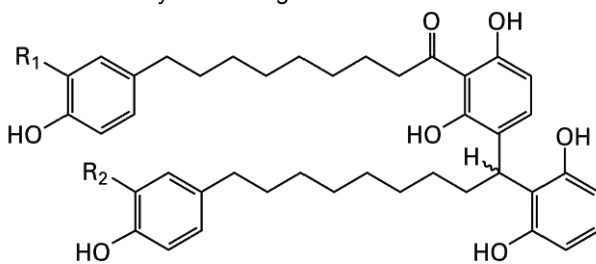


Anticholinesterase activity

An anticholinesterase bioassay has allowed the systematic screening of a large number of plants at the Institute of Natural Products Chemistry, among which Myristicaceae (nutmeg family) from Malaysia (example 17.4).

Example 17.4 - anticholinesterase alkylphenols from *Myristica crassa*, a plant collected in Malaysia.

A significant acetylcholinesterase inhibitory activity was observed for the ethyl acetate extracts from the leaves and the fruits of several Myristicaceae collected in Malaysia (MAIA *et al.*, 2008). As the strongest inhibition was observed for the extract of the fruits of *Myristica crassa*, this species was selected for further investigation. This study was accomplished with the aid of HPLC-ESI-MS and NMR analysis, and led to the isolation and identification of 3 new acylphenol dimers, giganteone C and maingayones B and C, along with the known malabaricones B and C and giganteone A. As little as 2 g of crude extract were sufficient to undertake this study and 50 mg for the standard HPLC fractionation.



Maingayone B: $R_1 = OH$, $R_2 = H$; Maingayone C: $R_1 = H$, $R_2 = OH$

17.5.4. POTENTIAL AND LIMITATIONS

In this chapter, the interest of screening plant extracts for the discovery of new active molecules is illustrated. It is believed that studying biodiversity will contribute not only to the knowledge of plant components but mainly to the isolation of compounds that can interact with specific cellular or enzymatic targets and lead to potential drugs in various pharmacological and therapeutic domains. Natural products in general, and those synthesized by plants in particular, possess a high chemical diversity and biological specificity. To date, these characteristics have not been found with computational and combinatorial chemistry, nor by human design. Who could have imagined the complex structures and the anticancer properties of the alkaloid vinblastine, the diterpene taxol or the macrocyclic epothilone? These compounds, provided as examples, are produced by plants or microorganisms and are probably used as chemical defences, although the real cause for their biosynthesis is not really known. Plants produce a large varied range of products with structures belonging to different series such as terpenes, alkaloids, polyketides, glycosides, flavonoids etc. This chemical diversity found in natural products has not been exploited entirely for its biological diversity: 'old' (known) products may interact with new biological targets and new isolated compounds may possess interesting biological properties. For that reason, it seems important to study, as far as we can, living organisms for their potential activities. The strategies adopted at the Institute of Natural Products Chemistry as well as in other research centers worldwide, allow the exploration of tropical plants, which contain molecules having complex structures. Thanks to the official cooperation programs with colleagues from Malaysia, Vietnam, Uganda and Madagascar and those from New Caledonia and French Guiana, a number of plant extracts is at our disposal to be screened against cellular and enzymatic targets. One important point to note is that these collaborations also lead to the training of students from these countries, with mutual benefit, capacity-building effects and cooperation with developing nations.

As far as the proposed extraction strategy is concerned, the use of ethyl acetate as the extraction solvent, in order to remove polyphenols and tannins that possess unspecific interactions with protein targets, avoids the isolation of more polar compounds that might possess biological activity. This choice was justified by the fact that hydrophilic compounds are often difficult to handle as potential drugs, and furthermore that it was not reasonable to increase the number of extracts when considering the limited capacity of research teams. Nevertheless, taking into account ethnomedicinal information, the extraction process can be adapted based on local use by traditional practitioners. Another possible limitation is related to the extract itself, which is defined as a complex mixture of natural products. A strong UV absorption or a specific fluorescence emission of some compounds can interfere with some methods of detection designed for miniaturized assays, leading to wrong interpretations.

17.6. CONCLUSION

This chapter reports a sweeping change in the field of classical phytochemistry, in which focussed searches of different chemical categories were previously preferred (alkaloids, acetogenins, saponins etc.), rather than an extensive exploration, which is now made possible. The novel technologies and strategies allow an increase in yield, although a standardised method of dereplication is needed. It is now possible to isolate minor compounds from plants and to elucidate their structure with minute amounts of products. The strategies exposed here need to be improved, as well as the biological screening, but the preliminary results observed are noteworthy. Given the potential of biodiversity to produce sophisticated, original and most importantly, bioactive compounds, the future challenge lies therefore in the protection of biodiversity, and in increasing our current capacity to investigate the chemical diversity it might provide. This would definitely bridge the past, *i.e.* traditional pharmacopeia, and the present, *i.e.* technology, and be probably more rational for the introduction of small molecules to the environment, as part of green chemistry objectives.

17.7. REFERENCES

- ALLOUCHE N., MORLEO B., THOISON O., DUMONTET V., NOSJEAN O., GUÉRITTE F., SÉVENET T., LITAUDON M. (2008) Biologically active tetralones from New Caledonian *Zygogynum* spp. *Phytochemistry* **69**: 1750-1755
- BOUSSEROUËL H., LITAUDON M., MORLEO B., MARTIN M.-T., THOISON O., NOSJEAN O., BOUTIN J., RENARD P., SÉVENET T. (2005) New biologically active linear triterpenes from the bark of three new-caledonian *Cupaniopsis* sp. *Tetrahedron* **61**: 845-851
- CBD (Convention on Biodiversity (1992) <http://www.cbd.int/convention/convention.shtml>
- JOLLY C., THOISON O., MARTIN M.-T., DUMONTET V., GILBERT A., PFEIFFER B., LÉONCE S., SEVENET T., GUERITTE F., LITAUDON M. (2008) Cytotoxic turrianes of *Kermadecia elliptica* from the New Caledonian rain forest. *Phytochemistry* **69**: 533-540
- MAIA A., SCHMITZ-AFONSO I.M.-T., LAPRÉVOTE O., GUÉRITTE F., LITAUDON M. (2008) Acylphenols from *Myristica crassa* as new acetylcholinesterase inhibitors. *Planta Medica* **74**: 1457-1462
- MUTKE J., BARTHLOTT W. (2005) Patterns of vascular plant diversity at continental to global scales. *Biol. Skr.* **55**: 521-531.

Absorption

→ see *ADME*.

Activation

Stimulation or acceleration of a biological process.

To learn more: chapters 1 and 5. → see also *Inhibition*.

Activator

A substance that stimulates or accelerates a biological process.

To learn more: chapters 1 and 5. → see also *Effector*; *Inhibitor*; *Biological phenomenon*; *Bioactivity*.

Activity

In biology, it designates the dynamic effect, the process, the change induced by the components of the living world.

To learn more: chapters 1 and 5. → see also *Protein*; *Enzyme*; *Metabolism*.

In QSAR, it designates the effect of a molecule on its target. The ambiguity with the biological sense (a molecule would then be ‘active’ on a biological ‘activity’) has led to the use of the term ‘bioactivity’, which is preferable.

To learn more: chapter 12. → see also *QSAR*; *Bioactivity*.

In UML, it represents a state in which a real-world task or a software process is carried out.

To learn more: chapter 6. → see also *UML*.

Activity diagram

In UML, a diagram showing the flow of activities.

To learn more: chapter 6. → see also *UML*; *Activity*.

Actor

In UML, this represents a role that the user plays with respect to a system.

To learn more: chapter 6. → see also *UML*.

ADME

ADME is an acronym for Absorption, Distribution, Metabolism and Excretion. It describes the efficacy of a pharmacological compound in an organism according to these four criteria. **Absorption**: before a compound acts, it must penetrate into the blood circulation, generally by crossing the intestinal mucosa (intestinal absorption). Entry into the target organs and cells must also be ensured. This can be a big problem with some natural barriers such as the blood-brain barrier. **Distribution**: the compound has to be able to circulate in the blood until it reaches its site of action. **Metabolism**: the compounds must be chemically destroyed once they have exerted their effect, otherwise they would accumulate in tissues and continue to interfere with natural processes. In some cases, chemical modifications taking place within cells are necessary prerequisites for an exogenous molecule to adopt its active form. **Excretion**: the compound metabolised must be excreted so as not to accumulate in the body to toxic doses.

→ see also *ADME-Tox*; *QSAR*.

ADME-Tox

ADME evaluation, accompanied by an evaluation of the molecule's possible toxicity at high doses.

→ see also *ADME*; *QSAR*.

Amino acid

Molecule constituting the main building block of proteins. Twenty exist in proteins (Alanine, Arginine, Asparagine, Aspartic acid, Cysteine, Glutamic acid, Glutamine, Glycine, Histidine, Isoleucine, Leucine, Lysine, Methionine, Phenylalanine, Proline, Serine, Threonine, Tryptophan, Tyrosine, Valine). The amino acids are linked in chains by peptide bonds. The type of amino acid is encoded by RNA and DNA by a combination of 3 bases termed a triplet or codon (e.g. ATG or TGC).

→ see also *Protein*; *Gene*.

Analysis

In the context of programming, the step of investigating a problem and its needs.

Anion

Molecule harbouring one or more negative charges.

To learn more: chapter 11. → see also *Cation*.

Antibody

Protein complex produced by some blood cells (white cells) and whose dual function consists both of recognising and binding an 'antigen' molecule arising from a foreign body, and of activating other cells participating in the body's immune defence. The specific recognition of an antigen makes an antibody the ideal tool ideal for labelling the antigen in question, referred to as immunolabelling. In practice, an interesting biological structure (a target for example) is injected into the blood of a small animal (mouse or rabbit) whereupon it acts as an antigen stimulating the accumulation in the blood of a specific antibody (referred to as an anti-target antibody). By collecting the anti-target antibody and coupling it to a fluorochrome it is possible for example to visualise the target in a cellular context after immunolabelling and detection of the fluorescent complex of target plus antibody.

To learn more: chapters 3 and 8. → see also *Cytoplot*.

Apoptosis

Also called 'programmed cell death'. It corresponds to a sort of gentle cell death by 'implosion', which does not cause damage to its environment, contrary to necrosis, a violent death by 'explosion' of the cell. The malfunctioning of apoptosis can lead to the immortalisation of cells normally destined to die, thus inducing the formation of cancerous tumours.

→ see also *Cancer*.

Aromatic ring

An aromatic ring is a cyclic group of atoms that possesses $(4n + 2)\pi$ electrons. To be aromatic, all of the compound's π electrons must be located in the same plane.

To learn more: chapter 11.

Automatic learning

Automatic learning is a field of study in artificial intelligence. Its purpose is to construct a computational system capable of storing and ordering data. By extension, any method allowing construction of a model of the reality from the data. Based on a formal description of molecules and the results of pharmacological screening, automatic learning is a means to deduce QSAR-type rules.

To learn more: chapter 15. → see also *Molecular descriptor*; *QSAR*.

Arity

→ see *Multiplicity*.

Assay / Test

Practical implementation of a protocol for testing samples, resulting in the emission of a signal. This signal allows the measurement of a biological phenomenon.

To learn more: chapters 1 and 6. → see also *Biological phenomenon*; *Test protocol*; *Model*; *'Biological-Target Molecule' project*; *Reaction medium*; *Signal*; *Bioactivity*.

Association

In UML, a particular relationship between two classes.

To learn more: chapter 6. → see also *Class*; *UML*.

Attribute

In UML, a named characteristic or property of a class.

To learn more: chapter 6. → see also *Class*; *UML*.

Background

Value rounded to 'zero' relative to the measured signal. The background can, in certain cases, be taken as the reference signal.

To learn more: chapter 4. → see also *Biological inactivity control*; *Test*.

Base

In chemistry, a base is a product that, when mixed with an acid, gives rise to a salt. Two definitions exist. According to the definition of Joannes BRONSTED and Thomas LOWRY, a base is a chemical compound that tends to capture a proton from a complementary entity, an acid. The reactions that take place between an acid and a base are called acid-base reactions. Such a base is termed a BRONSTED base. By the definition of LEWIS, a base is a species that can, during the course of a reaction, share a pair of electrons (a doublet). It is therefore a nucleophilic species, which possesses a non-bonding electron pair in its structure.

To learn more: chapter 11.

In biology, a base is in particular a nitrogenous molecule and a component of the nucleotides in DNA (adenine, guanine, cytosine and thymine) or ribonucleotides in RNA (adenine, guanine, cytosine and uracil).

→ see also *Nucleic acids*; *Nucleotides*.

Bioactive molecule (false)

Molecule having a proven effect on the measured signal but acting on a molecular target other than that studied or involving another biological mechanism.

To learn more: chapter 1. → see also *Bioactivity*; *Test*.

Bioactive molecule (true)

Molecule whose effect on the biological phenomenon is caused by the direct interaction with the biological target studied.

To learn more: chapter 1. → see also *Bioactivity*; *Test*.

Bioactivity (Biological activity)

Characterises a molecule that has a measurable effect on the biological phenomenon studied, dependent on the test protocol used.

To learn more: chapter 1. → see also *Test*.

Bioactivity control

Particular mixture for which the measured signal is equivalent to the expected signal for a highly bioactive molecule.

To learn more: chapter 1. → see also *Bioactivity*; *Test*.

Bioavailability

Describes a pharmacokinetic property of small molecules in their usage in the whole organism (as a drug), namely the fraction of the dose reaching the blood circulation, and by extension the cellular target. The bioavailability must be taken into consideration during the calculation of doses for administrative routes other than intravenous.

Biological inactivity

Characteristic of a molecule that does not have a measurable effect on the biological phenomenon studied, depending on the test protocol used.

To learn more: chapter 1. → see also *Bioactivity*; *Test*.

Biological inactivity control

Particular mixture for which the measured signal is equivalent to the expected signal for a molecule having no effect on the biological phenomenon (a molecule known to lack activity). An example molecule would be the solvent used for the chemical library, at the final concentration used.

To learn more: chapter 1. → see also *Bioactivity*; *Test*.

Biologically inactive molecule (false)

Molecule interacting with the target but not having a visible effect on the measured signal during the test.

To learn more: chapter 1. → see also *Bioactivity*; *Test*.

Biologically inactive molecule (true)

Molecule that does not interact with the biological target.

To learn more: chapter 1. → see also *Bioactivity*; *Test*.

Biological phenomenon

Event that is produced by the activity of the biological target.

To learn more: chapter 1. → see also *Effector*; *Activator*; *Inhibitor*.

'Biological-Target / Molecule' project

Analysis of the bioactivity of chemical library molecules on a biological phenomenon, by running the screening platform. Such a project is divided into a certain number of tasks, with the aim of identifying inhibitors or activators of a biological phenomenon in order to characterise the target.

To learn more: chapter 6. → see also *Test*; *Bioactivity*; *Model*.

Blank

→ see *Biological inactivity control*.

Cancer

Cancer is a general term to describe diseases characterised by anarchic cell proliferation within normal tissue. These cells are all derived from the same clone, a cancer initiator cell, which has acquired certain characteristics enabling it to divide indefinitely and may become metastatic.

Cation

Molecule harbouring one or more positive charges.

To learn more: chapter 11. → see also *Anion*.

Chip

Miniature device allowing parallel micro-experiments. While DNA chips are well known, chips permitting bioactivity measurements, and therefore pharmacological screens, are currently under development in numerous technological research centres. This book underlines the potential that these technological advances can bring (but does not illustrate their applications).

Chemical library

Collection of small molecules contained in multi-well plates.

To learn more: chapters 1, 2, 8, 10, 11 and 13. → see also *Molecule*; *Well*; *Multi-well plate*.

Chirality

A chiral molecule is composed of one or more atoms (most often carbon) linked to other atoms (generally by four bonds) in three-dimensional space possessing distinct groups in each corner of this space, which confers asymmetry upon the molecule. In chemistry, stereoisomers are distinguished as either enantiomers or diastereoisomers. Enantiomers are two three-dimensional molecules, each the mirror image of the other, and not superimposable (as for a person's two hands, for example). Diastereoisomers are molecules that are not mirror images of each other.

To learn more: chapter 11. → see also *Isomer*.

Chromosomes

Structures observable under a microscope within dividing cells, chromosomes carry and transfer hereditary characteristics. Composed of DNA and proteins, they harbour genes. In eukaryotes (living organisms whose cells have a nucleus), they are present in the nucleus of cells as homologous pairs, with two copies of each chromosome. Humans possess 23 pairs per cell. The germ cells of the gonads only contain a single copy of each pair.

Class diagram

In UML, a diagram showing the classes, their interrelationships and their relationships to other elements of the modelling process.

To learn more: chapter 6. → see also *UML*; *Class*.

Combinatorial chemistry

Fast generation of a chemical library of compounds identified as pure (parallel syntheses) or as a mixture, by using chemical reactions able to assemble several different fragments in only a few steps and by exploiting the different possible combinations.

In addition to reaction efficiency, combinatorial chemistry also includes technologies destined to simplify the steps of synthesis (notably purification, automation, miniaturisation etc.) in order to improve productivity. These products are then destined for biological screening.

To learn more: chapter 10.

Conceptual class

In UML, a set of objects or concepts that share the same attributes, relationships and behaviours.

To learn more: chapter 6. → see also *UML*.

Cytoblot

Immunodetection of a molecular target in cells by a specific antibody permitting, amongst other things, the detection of changes affecting the target, such as post-translational modifications (chemical modification of a protein after its initial synthesis), its cellular abundance, conformational changes etc. These enable visualisation of microscopic aspects of the cellular phenotype.

To learn more: chapters 3 and 8. → see also *Antibody*.

Dalton

In biochemistry, the name given to the atomic mass unit (symbol Da). It is equal to 1/12 of the mass of an atom of carbon 12, which is 1.66×10^{-27} kg.

Design

In the context of software programming, the step of developing a design solution satisfying the needs.

To learn more: chapter 6. → see also *UML*.

Diastereoisomer

→ see *Chirality*.

Distribution

In biology, transport of a molecule in the body.

→ see also *ADME*.

In robotics, dispensing of solutions and compounds in different wells of a multi-well plate.

→ see also *Well*; *Multi-well plate*.

DNA (deoxyribonucleic acid)

Molecule making up chromosomes. It is composed of two complementary chains, spiraling around each other (double helix). Each strand is a chain of nucleotides. A nucleotide, the elementary building block of DNA, is composed of three molecules: a simple sugar, a phosphate group and one of four nitrogenous bases, which are adenine, guanine, cytosine and thymine (A, G, C and T). The two DNA strands couple together in a double helix at the centre of which the bases pair up due to complementarity: A with T, and C with G.

→ see also *Nucleic acids*; *Nucleotide*; *Chromosome*; *Gene*.

Docking

Prediction of the binding mode of a small molecule to its macromolecular target.

To learn more: chapter 16. → see also *Virtual screening*; *Molecular modelling*.

Domain model

In UML, a set of class diagrams.

To learn more: chapter 6. → see also *Class*; *Modelling*; *UML*; *Ontology*.

DOS

Chemistry: Diversity-Oriented Synthesis; (not to be confused with the computing term, Disk Operating System).

To learn more: chapter 10. → see also *Diversity-oriented synthesis*.

Drug design

Set of *in silico* design techniques for novel medicinal substances with the aim of optimising their interactions with a given target.

EC50

Effective concentration at 50% of the total effect. The experimenter defines with a dose-effect curve, the concentration of molecule for which 50% of the bioactivity is observed, whether inhibition or activation, or indeed any measurable effect.

To learn more: chapter 5. → see also *IC50*.

Effector

Substance – activator or inhibitor – exerting an effect on a biological phenomenon.

To learn more: chapter 1. → see also *Activation*; *Inhibition*; *Bioactivity*; *Biological phenomenon*.

ELISA

An ELISA test (Enzyme-Linked Immunosorbent Assay) is an immunological test, utilising antibodies to detect and to assay a target (to which the antibody is specific) in a biological sample.

To learn more: chapter 3. → see also *Antibody*.

Enantiomer

→ see *Chirality*.

Enzymatic screening

Screening whereby the test permits measurement of an enzyme activity in a suspension containing a partially or fully purified enzyme.

To learn more: chapters 1 and 5. → see also *Screening*.

Enzyme

Protein or protein complex catalysing a chemical reaction of metabolism.

To learn more: chapters 5 and 14. → see also *Protein*; *Activity*; *Target*; *Enzyme activity*; *Metabolism*.

Excretion

Set of natural biological processes permitting the elimination of organic matter.

→ see also *ADME*.

Force field

A force field is a series of equations modelling the potential energy of a group of atoms in which a certain number of parameters is determined experimentally or evaluated theoretically. It describes the interactions between bonded and non-bonded atoms.

To learn more: chapter 11. → see also *Molecular modelling*.

Genes

DNA segments within chromosomes, they conserve and transmit hereditary characteristics. A gene is an element of information, characterised by the order in which the nucleic acid bases are linked and containing the necessary instructions for the cellular production of a particular protein. A gene is said 'to code' for a protein.

→ see also *DNA*; *Chromosome*; *Protein*.

Genome

Set of genes in an organism. The genome of a cell is formed from all of the DNA that it contains.

→ see also *Gene*.

Genotype

Set of genetic information of an individual organism or a cell.

→ see also *Phenotype*.

High-content screening

Screens that measure several parameters for each sample and from which one attempts to extract more detailed information regarding the biological effects of the molecules.

To learn more: chapters 8 and 9. → see also *Screening*; *Signal*.

High-throughput screening

Screening performed with a large number of samples (from several thousands to several millions), requiring automation of both the tests (in general miniaturised, carried out in multi-well plates) and signal detection.

To learn more: chapter 1. → see also *Screening*; *Miniaturisation*.

Hit

Molecule coming from a chemical library, whose effect on a biological target under experimental study conditions, is identified by 'screening' the entire chemical library. In any given screen, the number of hits obtained depends on the number of molecules present in the chemical library and on the experimental conditions (notably of the molecular concentration used).

To learn more: chapter 1. → see also *Hit candidate*.

Hit candidate

Bioactive molecule kept for having a signal higher than a bioactivity threshold.

To learn more: chapters 1 and 3. → see also *Hit*.

HTS

→ see *High-throughput screening*.

Hydrogen bond

A hydrogen bond is a weak chemical bond. It is an electrostatic dipole-dipole interaction which is generally formed between the hydrogen of a heteroatom and a free electron pair carried by another heteroatom. The hydrogen linked to an electronegative atom bears a fraction of very localised positive charge allowing it to interact with the dipole produced by the other electronegative atom, which thus functions as a hydrogen-bond acceptor.

To learn more: chapters 11, 12 and 13.

IC50

Concentration of the inhibitory compound at 50% of the total inhibition.

To learn more: chapter 5. → see also *EC50*.

Immunolabelling

→ see *Antibody*.

Implementation

In the context of software programming, the step of putting into practice a design solution.

To learn more: chapter 6.

Information system

System that produces and manages information to assist humans with the functions of making and executing decisions. The system comprises computer hardware and software programs.

To learn more: chapter 6.

Inhibition

Phenomenon of halting, blocking or slowing down a biological process.

To learn more: chapters 1 and 5. → see also *Activation*.

Inhibitor

A substance that blocks or interferes with a biological process.

To learn more: chapters 1 and 5.

Isomers

Two molecules are isomers when they have an identical atomic composition but a different molecular arrangement.

To learn more: chapters 11 and 13. → see also *Chirality*; *Tautomer*.

Ligand

In chemistry, a ligand is an ion, an atom or functional group linked to one or more central atoms or ions (often metals) by non-covalent bonds (typically, coordination bonds). An assembly of this sort is termed a 'complex'.

In biochemistry, a ligand is a molecule interacting specifically and in a non-covalent manner with a protein, called the target. Originally, ligand referred to a natural compound binding to a specific receptor, however, this term is also employed to mean a synthetic compound acting on a target competitively or not with respect to the natural ligand. When the molecule bound is converted by an enzyme's catalytic activity, the term 'substrate' is used. Quantification of ligand binding calls upon a huge variety of techniques. In conventional biochemistry, radioactive forms of a ligand ('hot' ligands) together with a variable proportion of non-radioactive ligand ('cold' ligand) allow the quantity of bound ligand to be measured by competition assay.

→ see also *Receptor*.

Metabolism

The group of molecular conversions, chemical reactions and energy transfers that take place continuously in the cell or living being.

→ see also *Biological phenomenon*; *Phenotype*; *ADME*.

Miniaturisation

Design and simplification of a test making it measurable in a multi-well plate and adapted so as to be manageable using a liquid-handling robot and other peripheral devices.

To learn more: chapter 3. → see also *Well*; *Multi-well plate*.

Model

Formalised structure, used to account for a set of interrelated phenomena.

To learn more: chapter 6. → see also *Modelling*; *Domain model*; *UML*; *Ontology*.

Modelling

Building of models.

To learn more: chapter 6. → see also *Model*.

Molecular descriptor

Object characterising the information contained in a molecule allowing analysis and manipulation using computational tools.

To learn more: chapters 11, 12 and 13. → see also *Virtual screening*; *Molecular modelling*.

Molecular modelling

Empirical method permitting the experimental results to be adequately reproduced using simple mathematical models of atomic interactions.

To learn more: chapter 11. → see also *Force field*.

Molecule

The smallest part of a chemical substance that can exist independently; molecules are composed of two or more atoms.

To learn more: chapter 11.

mRNA (messenger RNA)

RNA molecule whose role consists of transmitting the information contained in the sequence of bases of one strand of a DNA molecule (therefore its genetic code) to the cellular machinery that manufactures proteins (chains of amino acids).

→ see also *RNA*; *Gene*; *Protein*.

Multiplicity

In UML, indicates the number of objects likely to participate in a given association.

To learn more: chapter 6. → see also *UML*.

Multi-well plate

A plate generally having the dimensions: 8 × 12 cm with a depth of 1 cm, in which rows of individual depressions, called wells (normally 24, 48, 96 or 384), are arranged. These plates, disposable after use, can be handled by robots. The different reagents, biological extracts and chemical library molecules are dispensed in this plate type during screening. The effect of each molecule in each well is measured with appropriate signal detection methods.

To learn more: chapter 1. → see also *Well*; *Reaction medium*; *Chemical library*; *Test*; *Screening*; *Well function*.

Nucleic acids

Biological macromolecules harbouring hereditary information, comprising genes. Two types exist depending on the nature of the constituent sugar molecule: DNA (component of chromosomes) and RNA. Nucleic acids are characterised by their particular sequence of nucleotides.

→ see also *DNA*; *RNA*; *Nucleotide*.

Nucleotide

Basic motif of DNA comprising three chemical elements: one of four nitrogenous bases (A, C, G or T), the sugar deoxyribose, and a phosphate group. In RNA the sugar is ribose and another base, uracil (U), replaces thymine (T).

→ see also *Nucleic acids*; *DNA*; *RNA*.

Object

An instance of a real entity (e.g. a person, a thing etc.).

To learn more: chapter 6. → see also *Class*; *UML*; *Ontology*.

Ontology

Ontology was originally a field of Philosophy aiming to study the nature of being. Ontology has been used for several years in Knowledge Engineering and Artificial Intelligence for structuring the concepts within these fields. The concepts are grouped together and considered as elementary blocks allowing expression of the domain knowledge covered by them. In practice, an ontology is conceived in a simple form to be a 'structured vocabulary' and in its more sophisticated form as a 'schema, which shows unambiguously everything known about a given subject of study', while featuring the semantic relationships. Ontologies are useful for sharing knowledge, creating a consensus, constructing systems of knowledge-bases and ensuring the interoperability between different computing systems. They are therefore essential in multidisciplinary domains such as chemogenomics. Numerous ontology projects are in process such as gene, cell or indeed target ontology.

To learn more: chapters 1, 6 and 14. → see also *Class*; *UML*; *Object*.

Parallel synthesis

The compounds are manufactured in distinct reactors simultaneously (by automation, for example) in one or several steps in order to be able to develop a unique product (in the ideal case). Reactions are therefore chosen which show good chemo- and stereoselectivity. At the end, each well of the plate destined for screening can only contain a single product.

To learn more: chapter 10.

Pharmacophore

A pharmacophore is made up of the pharmacologically active part of a molecule acting as a model. Pharmacophores are therefore groups of active atoms used in drug design.

To learn more: chapters 11 and 13. → see also *Pharmacophore point*.

Pharmacophore point

Pharmacophore points are atoms or groups of atoms in a molecule which, due to their particular arrangement in the molecule, acquire specific interaction properties: typically hydrogen bond donors or acceptors, anions, cations, aromatic and hydrophobic centres.

To learn more: chapters 11 and 13. → see also *Pharmacophore*.

Phenotype

Set of apparent characteristics of a cell or an individual. These characteristics result from the interaction of genetic factors and the external environment.

To learn more: chapters 8 and 9. → see also *Genotype*.

Phenotypic screening

Screening in which the test allows measurement of a complex phenotypic trait of cells or whole organisms.

To learn more: chapters 8 and 9. → see also *Screening*; *Antibody*; *Cytoblot*.

Protein

A complex molecule whose backbone is formed by the linkage of amino acids, and having functions as varied as catalysis (enzymes), the recognition of foreign bodies (antibodies) or energy transport (e.g. globin associated to iron, as in haemoglobin).

→ see also *Amino acids*; *Gene*; *Target*.

QSAR

Acronym for Quantitative Structure-Activity Relationship. From the measured bioactivity of a set of molecules sharing certain structural properties, a QSAR analysis aims to deduce a quantitative correlation linking bioactivity with (data or properties relative to) the structure of molecules.

To learn more: chapters 12 and 15.

Reaction medium

Contained in a well, the reaction medium is a solution composed of a number of reagents as well as the biological target, relating to which the experimenter wishes to study a particular phenomenon.

→ see also *Well*; *Multi-well plate*.

Receptor

In biochemistry, a receptor is a protein to which a neurotransmitter – a hormone (or more generally, a ligand) – binds specifically, and thus induces a cellular response.

→ see also *Ligand*.

Reference signal

Signal measured in the absence of the test molecule. For example, in the case of screening, this would mean in the absence of a small molecule.

To learn more: chapters 1, 3 and 4. → see also *Signal*.

RNA (ribonucleic acid)

Molecule very similar to DNA but containing most commonly a single strand, formed from a backbone made of phosphate and ribose sugars, along which the bases (adenine, cytosine, guanine or uracil) are attached in a linear sequence.

→ see also *mRNA*; *Nucleic acids*; *DNA*; *Gene*.

Screening

Carrying out tests (screens) to measure the bioactivity or biological inactivity of each molecule towards the biological target, at a known concentration under the experimental study conditions. Screening is a task within a 'Biological-Target / Molecule' project.

To learn more: chapter 1. → see also *Bioactivity*; *Test*; *High-throughput screening*; *High-content screening*; *Enzymatic screening*; *Phenotypic screening*; *Virtual screening*; 'Biological-Target / Molecule' project.

Signal

Measurable property of the biological phenomenon in the conditions specified by the protocol. This signal can be the absorption (absorbance) or emission (fluorescence) of light. The signal in a broader sense could also be an image of a cell. In this case, signal detection requires an analysis of the image obtained.

To learn more: chapters 1, 3 and 4. → see also *Test protocol*; *Model*; 'Biological-Target / Molecule' project; *Reaction medium*; *Signal*; *Test*.

SQL

Acronym for Structured Query Language, a language of structured requests used to generate and to interrogate relational databases.

→ see also *Database*.

Stereoisomer

Isomer possessing chirality. Stereoisomers have the same KEKULÉ structural formula but have a different spatial arrangement of chemical groups. They can be related as enantiomers or as diastereoisomers.

→ see also *Chirality*.

Target (pharmacological)

What one aims to reach pharmacologically. It is important to define well what one means by a target: it could simply refer to the molecular target (a protein); the target can also be a complex structure (e.g. a subcellular compartment such as an organelle, a complex of several proteins, a whole cell characterised by its phenotype) or a dynamic and complex process that one aims to destabilise (for instance, a metabolic pathway).

To learn more: chapter 1 and 14. → see also *Ontology*; *Protein*; *Metabolism*; *Phenotype*; *Bioactivity*.

Task (in a project)

Action performed within the framework of a project, producing a deliverable: numerical results or a list of bioactive molecules, for instance.

To learn more: chapter 7. → see also '*Biological-Target / Molecule*' project; *Deliverable*.

Tautomer

Tautomers are isomers of compounds in equilibrium in a tautomerisation reaction, which involves the simultaneous migration of a proton and a double bond.

To learn more: chapter 11. → see also *Isomer*.

Test protocol

Exhaustive description of the solutions, experimental conditions (temperature and incubation time), and processes to be planned and executed.

To learn more: chapter 3. → see also *Model*; '*Biological-Target / Molecule*' project; *Reaction medium*; *Signal*; *Test*.

Topology

The topology of a molecule is the description of the set of interatomic connections of which it is composed. It is in fact its two-dimensional structure without taking into account either the atom or bond types.

To learn more: chapter 11.

Use case

In UML, the complete set of actions, initiated by an actor and which the system executes to bring a benefit to this actor.

To learn more: chapter 6. → see also *Actor*; *UML*.

Use-case diagram

In UML, a diagram showing the use cases, their interrelationships and their relationships to the actors.

To learn more: chapter 6. → see also *UML*; *Use case*.

UML

Acronym for Unified Modelling Language, it is a standardised notation in modelling.

To learn more: chapter 6. → see also *Actor*; *Class*; *Implementation*; *Model*; *Object*; *Ontology*.

Virtual screening

Screening driven by computational methods. Searching an electronic chemical library for molecules satisfying the constraints imposed by specific physicochemical properties, a pharmacophore or the topology of a binding site.

To learn more: chapter 16. → see also *Screening*.

Well

Small depression in a multi-well plate.

To learn more: chapter 6. → see also *Well function*.

Well function

Basic purpose for which the reaction conditions have been established in a given well (for example: control, sample).

To learn more: chapter 6.

THE AUTHORS



Samia ACI

Research officer, CNRS

Centre of Molecular Biophysics, CNRS, Orléans, France



Caroline BARETTE

Research engineer, CEA

Laboratory of Large-Scale Biology,

Centre for Bioactive Molecules Screening,

Institute of Life Sciences Research and Technologies, CEA, Grenoble, France



Gilles BISSEAU

Research officer, CNRS

TIMC IMAG Laboratory - Joseph Fourier University,

Grenoble Institute of Applied Mathematics, France



Alain CHAVANIEU

Lecturer at Montpellier University

Centre for Structural Biochemistry,

CNRS - Servier - Montpellier University - INSERM,

Faculty of Pharmacy, Montpellier, France



Jean CROS

Professor at Toulouse University

Institute for Pharmacology and Structural Biology,

CNRS - Pierre Fabre,

Centre for Pharmacology and Health Research, Toulouse, France

**Benoît DÉPREZ**

*Professor at the Faculty of Pharmacy of Lille,
Correspondent of the National Academy of Pharmacy,
Former director of the Lead Discovery Department of the company Devgen
Director of Inserm lab U761 "Drug Discovery",
Pasteur Institute of Lille, University of Lille, Lille, France*

**Vincent DUMONTET**

*Research engineer, CNRS
Institute for Natural Products Chemistry,
CNRS - Gif Research Center, Gif-sur-Yvette, France*

**Gérard GRASSY**

*Professor at the University of Montpellier,
Correspondent of the National Academy of Pharmacy,
Professeur Agrégé in Pharmacy
Centre for Structural Biochemistry,
CNRS - Servier - University of Montpellier - INSERM,
Faculty of Pharmacy, Montpellier, France*

**Françoise GUERITTE**

*Research director, INSERM
Institute for Natural Products Chemistry,
CNRS - Gif Research Center, Gif-sur-Yvette, France*

**Marcel HIBERT**

*Professor at the Strasbourg Faculty of Pharmacy,
Director of the French National Chemical Library, CNRS Silver Medallist
Laboratory of Therapeutic Innovation,
CNRS - Strasbourg University,
Faculty of Pharmacy, Illkirch, France*

**Dragos HORVATH**

*Research officer, CNRS
Former director of the Molecular Modelling Department of the company Cerep
InfoChemistry Laboratory
UMR 7177 CNRS - Strasbourg University,
Institute of Chemistry, Strasbourg, France*

**Martine KNIBIEHLER**

Research engineer, CNRS
Institute of Advanced Technologies in Life Sciences
CNRS - University of Toulouse III - INSAT
Centre Pierre Potier / ITAV - Canceropôle Toulouse, France

**Laurence LAFANECHÈRE**

Director of research, CNRS
Institut Albert Bonniot, Molecular Ontogenesis and Oncogenesis,
INSERM - CNRS - CHU - EFS - Joseph Fourier University,
and Centre for Bioactive Molecules Screening,
Institute of Life Sciences Research and Technologies, CEA, Grenoble, France

**Marc LITAUDON**

Research engineer, CNRS
Institute for Natural Products Chemistry,
CNRS - Gif Research Center, Gif-sur-Yvette, France

**Eric MARÉCHAL**

*Director of research, CNRS,
Professeur Agrégé in Natural Sciences*
Laboratory of Plant Cell Physiology,
CNRS - CEA - INRA - Joseph Fourier University,
Institute of Life Sciences Research and Technologies, CEA, Grenoble, France
eric.marechal@cea.fr

**Jordi MESTRES**

Professor at the Pompeu Fabra University of Barcelona
Chemogenomics Laboratory,
Municipal Institute for Medical Investigation,
Pompeu Fabra University, Barcelona, Spain

**Didier ROGNAN**

Director of research, CNRS
Laboratory of Therapeutical Innovation,
CNRS - Louis Pasteur University of Strasbourg, Illkirch, France



Sylvaine ROY

Research engineer, CEA
Laboratory of Plant Cell Physiology,
CNRS - CEA - INRA - Joseph Fourier University,
Institute of Life Sciences Research and Technologies, CEA, Grenoble, France



Thierry SEVENET

Research director, CNRS
Institute for Natural Products Chemistry,
CNRS - Gif Research Center, Gif-sur-Yvette, France



André TARTAR

Professor at the University of Lille,
Co-founder and former Vice-President of the company Cerep
Unit of Biostructure and Medicine Discovery,
INSERM - Pasteur Institute of Lille, University of Lille, Lille, France



Samuel WIECZOREK

Research engineer, CEA
Laboratory of Large-Scale Biology,
Institute of Life Sciences Research and Technologies, CEA, Grenoble, France



Yung-Sing WONG

Research officer, CNRS
Department of Molecular Pharmacochimistry,
CNRS - Joseph Fourier University
Grenoble Institute of Molecular Chemistry, Grenoble, France

This work follows on from an *école thématique* organised by the CNRS and the CEA for students and researchers wishing to learn the new discipline born of automated pharmacological screening technologies: chemogenomics.



The authors would like to thank Yasmīna SAOUDI, Andreï POPOV and Cyrille BOTTÉ for having authorised the reproduction of their photographs in this work.