ADVANCES IN BIOCHEMICAL ENGINEERING/BIOTECHNOLOGY 121

Series Editor T. Scheper
Volume Editors C. Wittmann · R. Krull

Biosystems Engineering II

Linking Cellular Networks and Bioprocesses

Springer

121
Advances in Biochemical Engineering/Biotechnology

Series Editor: T. Scheper

Advances in Biochemical Engineering/Biotechnology

Series Editor: T. Scheper

Recently Published and Forthcoming Volumes

Biosystems Engineering II
Linking Cellular Networks and Bioprocesses
Volume Editors: Wittmann, C., Krull, R.
Vol. 121, 2010

Biosystems Engineering I
Creating Superior Biocatalysts
Volume Editors: Wittmann, C., Krull, R.
Vol. 120, 2010

Whole Cell Sensing Systems II
Volume Editors: Belkin, S., Gu, M.B.
Vol. 118, 2010

Whole Cell Sensing Systems I
Volume Editors: Belkin, S., Gu, M.B.
Vol. 117, 2010

Optical Sensor Systems in Biotechnology
Volume Editor: Rao, G.
Vol. 116, 2009

Disposable Bioreactors
Volume Editor: Eibl, R., Eibl, D.
Vol. 115, 2009

Engineering of Stem Cells
Volume Editor: Martin, U.
Vol. 114, 2009

Biotechnology in China I
From Bioreaction to Bioseparation and Bioremediation
Volume Editors: Zhong, J.J., Bai, F.-W., Zhang, W.
Vol. 113, 2009

Bioreactor Systems for Tissue Engineering
Volume Editors: Kasper, C., van Griensven, M., Poertner, R.
Vol. 112, 2008

Food Biotechnology
Volume Editors: Stahl, U., Donalies, U. E. B., Nevoigt, E.
Vol. 111, 2008

Protein – Protein Interaction
Volume Editors: Seitz, H., Werther, M.
Vol. 110, 2008

Biosensing for the 21st Century
Volume Editors: Renneberg, R., Lisdat, F.
Vol. 109, 2007

Biofuels
Volume Editor: Olsson, L.
Vol. 108, 2007

Green Gene Technology
Research in an Area of Social Conflict

Volume Editors: Fiechter, A., Sautter, C.
Vol. 107, 2007

White Biotechnology
Volume Editors: Ulber, R., Sell, D.
Vol. 105, 2007

Analytics of Protein-DNA Interactions
Volume Editor: Seitz, H.
Vol. 104, 2007

Tissue Engineering II
Basics of Tissue Engineering and Tissue Applications
Volume Editors: Lee, K., Kaplan, D.
Vol. 103, 2007

Tissue Engineering I
Scaffold Systems for Tissue Engineering
Volume Editors: Lee, K., Kaplan, D.
Vol. 102, 2006

Cell Culture Engineering
Volume Editor: Hu, W.-S.
Vol. 101, 2006

Biotechnology for the Future
Volume Editor: Nielsen, J.
Vol. 100, 2005

Biosystems Engineering II

Linking Cellular Networks and Bioprocesses

Volume Editors:
Christoph Wittmann · Rainer Krull

With contributions by

C. Cordes · A. Elsheikh · M.T. Fernández-Sandoval · E. Flaschel · G. Gosset · J.J. Heijnen · H. Horn · G. Huerta-Beristain · I. Kampen · M. Klann · R. Krull · A. Kwade · A. Lapin · A. Martinez · T.R. Neu · B. Nörtemann · S. Noack · M. Orencio-Trejo · M. Reuss · J. Utrilla · C. Vijayendran · W. Wiechert

Springer

Editors
Prof. Dr. Christoph Wittmann
Technische Universität Braunschweig
Biochemical Engineering Institute
Gaußstraße 17
38106 Braunschweig
Germany
c.wittmann@tu-braunschweig.de

Prof. Dr. Rainer Krull
Technische Universität Braunschweig
Biochemical Engineering Institute
Gaußstraße 17
38106 Braunschweig
Germany
r.krull@tu-braunschweig.de

ISSN 0724-6145 e-ISSN 1616-8542
ISBN 978-3-642-13865-2 e-ISBN 978-3-642-13866-9
DOI 10.1007/978-3-642-13866-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2010933955

Cover design: WMXDesign GmbH, Heidelberg, Germany

Printed on acid-free paper

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

Series Editor

Prof. Dr. T. Scheper

Institute of Technical Chemistry
University of Hannover
Callinstraße 3
30167 Hannover, Germany
scheper@iftc.uni-hannover.de

Volume Editors

Prof. Dr. Christoph Wittmann

Technische Universität Braunschweig
Biochemical Engineering Institute
Gaußstraße 17
38106 Braunschweig
Germany
c.wittmann@tu-braunschweig.de

Prof. Dr. Rainer Krull

Technische Universität Braunschweig
Biochemical Engineering Institute
Gaußstraße 17
38106 Braunschweig
Germany
r.krull@tu-braunschweig.de

Editorial Board

Honorary Editors

Advances in Biochemical Engineering/ Biotechnology Also Available Electronically

Advances in Biochemical Engineering/Biotechnology is included in Springer's eBook package *Chemistry and Materials Science*. If a library does not opt for the whole package the book series may be bought on a subscription basis. Also, all back volumes are available electronically.

For all customers who have a standing order to the print version of *Advances in Biochemical Engineering/Biotechnology*, we offer the electronic version via SpringerLink free of charge.

If you do not have access, you can still view the table of contents of each volume and the abstract of each article by going to the SpringerLink homepage, clicking on "Chemistry and Materials Science," under Subject Collection, then "Book Series," under Content Type and finally by selecting *Advances in Biochemical Bioengineering/Biotechnology*

You will find information about the

– Editorial Board
– Aims and Scope
– Instructions for Authors
– Sample Contribution

at springer.com using the search function by typing in *Advances in Biochemical Engineering/Biotechnology*.

Color figures are published in full color in the electronic version on SpringerLink.

Aims and Scope

Advances in Biochemical Engineering/Biotechnology reviews actual trends in modern biotechnology.

Its aim is to cover all aspects of this interdisciplinary technology where knowledge, methods and expertise are required for chemistry, biochemistry, microbiology, genetics, chemical engineering and computer science.

Special volumes are dedicated to selected topics which focus on new biotechnological products and new processes for their synthesis and purification. They give the state-of-the-art of a topic in a comprehensive way thus being a valuable source for the next 3-5 years. It also discusses new discoveries and applications.

In general, special volumes are edited by well known guest editors. The series editor and publisher will however always be pleased to receive suggestions and supplementary information. Manuscripts are accepted in English.

In references *Advances in Biochemical Engineering/Biotechnology* is abbreviated as *Adv. Biochem. Engin./Biotechnol.* and is cited as a journal.

Special volumes are edited by well known guest editors who invite reputed authors for the review articles in their volumes.

Impact Factor in 2009: 4.165; Section “Biotechnology and Applied Microbiology”: Rank 23 of 150

Attention all Users of the "Springer Handbook of Enzymes"

Information on this handbook can be found on the internet at springeronline.com

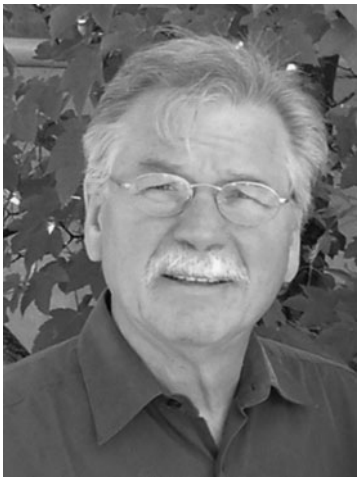
A complete list of all enzyme entries either as an alphabetical Name Index or as the EC-Number Index is available at the above mentioned URL. You can download and print them free of charge.

A complete list of all synonyms (more than 57,000 entries) used for the enzymes is available in print form (ISBN 978-3-642-14015-0) and electronic form (ISBN 978-3-642-14016-7).

Save 15%

We recommend a standing order for the series to ensure you automatically receive all volumes and all supplements and save 15% on the list price.

Laudatio



This special volume "Biosystems Engineering" is dedicated to Professor Dr.-Ing. Dietmar Christian Hempel on the occasion of his 65th birthday, whereby the different contributions display an excellent reflection of his research during the past 30 years, bridging engineering and life sciences.

Prof. Hempel, born in Königsberg, studied Construction and Process Engineering in Dortmund and Berlin (1963–1971). After his PhD work on "Heterogeneous catalytic fixed bed reactors" he became the head for "Reaction and Biochemical Engineering" at the Bayer AG in Leverkusen, where he started to apply chemical engineering principles to biological systems (1975–1980). This was obviously stimulating his later research as Professor of Technical Chemistry and Chemical Engineering at the University of Paderborn (1980–1994) and as Professor and Director of the Institute of Biochemical Engineering at the Technische Universität Braunschweig (1994–2009) where he kept his research interest on biological systems in technical applications – covering various aspects process engineering of biological and biochemical processes. As impressive outcome of his career, Dietmar Christian Hempel authored and co-authored about 300 publications and supervised 70 PhD students as a "Doktorvater". From 2001 to 2008 he was head and speaker of the still successfully continued interdisciplinary DFG-collaborate research centre SFB 578 "Development of biotechnological processes by integrating genetic and engineering methods – From gene to product". This collaborate research centre displays a landmark in the German biochemical engineering community and serves as important link between cellular biology and bioprocess engineering.

Preface

The special volume "Biosystems Engineering" reflects an emerging field of applied research that aims at a system-level understanding of biological systems toward their targeted design and improvement. To obtain system-wide insight, interdisciplinary approaches are applied, which integrate expertise from biologists, engineers, and computer scientists, who have developed powerful analytical methods, modelling concepts, and information technologies for this challenging task. With the increasing need for sustainable production of fuels and chemicals from renewable resources, a major focus in this area is on the optimization of biotechnological production strains and processes. Of specific importance hereby is the consideration of the microorganism as a part of the bioprocess in its entirety.

In this regard, the different contributions of this volume provide an outstanding review of novel tools, methods, and concepts in the field of biosystems engineering, a still young, but yet powerful direction of interdisciplinary research. Part I "Creating Superior biocatalysts" of this volume highlights how system-wide analysis, modelling and understanding of gene regulation, metabolic networks, and fluxes open a new era for rational design and optimization of superior production strains, one of the key pre-requisites for bio-based production of pharmaceuticals, chemicals, materials, and fuels. This includes the introduction of systems and synthetic metabolic engineering approaches as well as industrial application examples. Part II "Linking Cellular Networks and Bioprocesses" illustrates concepts and tools which are based on models and experiments to investigate metabolic networks in correlation with the cell environment. In addition to the introduction of fundamental strategies introducing thermodynamic principles as well as different modelling approaches to metabolic network simulation, different examples directly address the link of cells and bioreactors providing fascinating and valuable insights into the interaction between cellular metabolism and process environment.

Braunschweig, Summer 2010

Christoph Wittmann
Rainer Krull

Contents

Adv Biochem Engin/Biotechnol (2010) 121: 1–21
DOI: 10.1007/10_2009_60

Published online: 12 February 2010

Morphology of Filamentous Fungi: Linking Cellular Biology to Process Engineering Using *Aspergillus niger*

Dedicated to Prof. Dr.-Ing. Dietmar C. Hempel on the occasion of his 65th birthday.

Rainer Krull, Christiana Cordes, Harald Horn, Ingo Kampen, Arno Kwade, Thomas R. Neu, and Bernd Nörtemann

Abstract In various biotechnological processes, filamentous fungi, e.g. *Aspergillus niger*, are widely applied for the production of high value-added products due to their secretion efficiency. There is, however, a tangled relationship between the morphology of these microorganisms, the transport phenomena and the related productivity. The morphological characteristics vary between freely dispersed mycelia and distinct pellets of aggregated biomass. Hence, advantages and disadvantages for mycel or pellet cultivation have to be balanced out carefully. Due to this inadequate understanding of morphogenesis of filamentous microorganisms, fungal morphology, along with reproducibility of inocula of the same quality, is

R. Krull (✉) and B. Nörtemann
Institute of Biochemical Engineering, Technische Universität Braunschweig, Gaußstraße 17, 38106 Braunschweig, Germany
e-mail: r.krull@tu-bs.de; b.noertemann@tu-bs.de

C. Cordes
Applied Biosciences and Process Technology, University of Applied Science (FH), Bernburger Straße 55, 06354 Köthen, Germany
e-mail: christiana.cordes@bwp.hs-anhalt.de

H. Horn
Institute of Water Quality Control, Technische Universität München, Am Coulombwall, 85748 Garching, Germany
e-mail: horn@bv.tu-muenchen.de

I. Kampen and A. Kwade
Institute of Institute for Particle Technology, Technische Universität Braunschweig, Volkmaroder Straße 5, 38104 Braunschweig, Germany
e-mail: i.kampen@tu-bs.de; a.kwade@tu-bs.de

T.R. Neu
Department of River Ecology, Helmholtz-Zentrum für Umweltforschung – UFZ, Brueckstraße 3a, 39114 Magdeburg, Germany
e-mail: thomas.neu@ufz.de

often a bottleneck of productivity in industrial production. To obtain an optimisation of the production process it is of great importance to gain a better understanding of the molecular and cell biology of these microorganisms as well as the approaches in biochemical engineering and particle technique, in particular to characterise the interactions between the growth conditions, cell morphology, spore–hyphae-interactions and product formation. Advances in particle and image analysis techniques as well as micromechanical devices and their applications to fungal cultivations have made available quantitative morphological data on filamentous cells. This chapter provides the ambitious aspects of this line of action, focussing on the control and characterisation of the morphology, the transport gradients and the approaches to understand the metabolism of filamentous fungi. Based on these data, bottlenecks in the morphogenesis of *A. niger* within the complex production pathways from gene to product should be identified and this may improve the production yield.

Keywords *Aspergillus niger*, Fluid dynamics, Fungal morphology, Mechanical stress, Systems biotechnology

Contents

List of Abbreviations

AFM	Atomic force microscopy
CFD	Computational fluid dynamics
CLSM	Confocal laser scanning microscopy
$(dc/dr)_{max}$	Maximum oxygen concentration gradient
dh/dr	Gradient of the hyphal fraction h within the outer pellet periphery
GFP	Green fluorescent protein
h	Hyphal fraction
L_c	Concentration boundary layer
LES	Large eddy simulation
P/V	Volumetric power input
PIV	Particle image velocimetry
pspd	Position-sensitive photo-detector
r	Radial coordinate

RANS	Reynolds averaged Navier–Stokes equations
Re_P	Reynolds number at the pellet
RT-PCR	Reverse transcription polymerase chain reaction
SST	Shear stress transport turbulence model
STR	Stirred tank reactor
TKE	Turbulent kinetic energy

1 Introduction

In biotechnological production processes performed with filamentous fungi the monitoring and control of morphological development is difficult to obtain due to the highly complex relation between morphology and productivity [1]. Product formation by filamentous fungi, like *Aspergillus niger*, is closely linked to their morphology. Mycelial growth of this coagulating fungus has procedural disadvantages, for instance a high viscosity of the cultivation broth and therefore a low nutrient supply due to insufficient mixing. In comparison, cultivation broths with distinct pellets show Newtonian flow behaviour, but disadvantages related to a limited nutrient availability within the inner part of the biopellets can occur. Hence, in every biotechnological process, the optimal morphology varies due to specific product properties and cannot be generalised. The morphogenesis of *A. niger* cultivation can be controlled effectively by adjusting the pH value and the volumetric power input; see Fig. 1 [2, 3].

In the early phase of cultivation the aggregation of *A. niger* conidia is dominantly affected by the pH value, while the morphology of fungal pellets is mainly influenced by volumetric power input [4]. The comparison of the volumetric power input caused by agitation and aeration [5] revealed that the aeration has a higher impact to counteract the aggregation process, which leads to pellets with smaller diameter and an open structure, and also to a higher pellet concentration. Due to

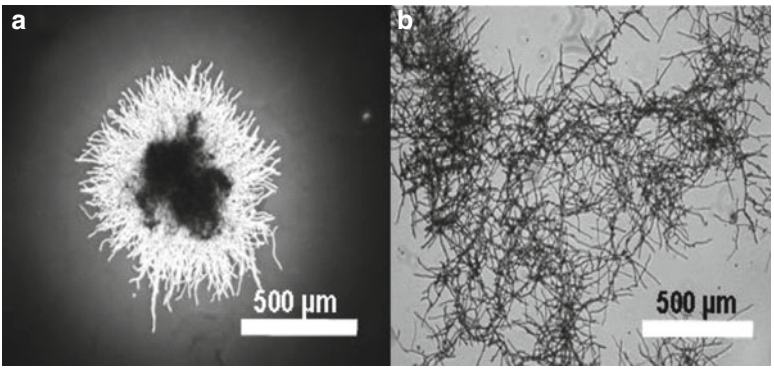


Fig. 1 Morphological characteristics of *A. niger* (volumetric power input $P/V = 100\ \mathrm{W\ m^{-3}}$), (**a**) biopellet at pH 5.5 after cultivation period of 32 h, (**b**) free dispersed mycelia at pH 3, after 24 h [2]

higher agitation induced power input denser pellets occur with a compact pellet surface [3].

Advantages in particle and image analysis techniques as well as in micromechanical devices and their applications to fungal cultivations have made available quantitative morphological data for the investigations of the micro- and macroscopic growth of filamentous cells. Measurements of the adhesive forces can be realised using atomic force microscopy (AFM). By this means, information about the morphological development at the beginning of the cultivation by adhesion forces between spores of *A. niger* could be measured. On the other hand, forces and tensions which lead to a hyphae breakage can be determined. The measurement of the breakage forces have been carried out using nanoindenter techniques. Thus, it is aimed to gain information about the cell wall mechanical properties. This is an important issue as the cell wall is mainly responsible for the resistance of the hyphae to breakage.

Based on intracellular reactions up to physico-chemical and fluid dynamic phenomena at a macroscopic level, which determine the fungal morphology, the morphogenesis of mycelial growth and pellet formation via distinct aggregation steps ought to be completely covered by population balancing and verified by different particle size analysis techniques.

The implication of product formation, transport and growth characteristics as a function of environmental conditions (e.g. cultivation conditions such as type and concentration of carbon source, temperature, pH value, fluid dynamics, transport processes, shear stress etc.) and, particularly when filamentous fungi are used as production strains, the integration of cell morphology and protein production processes in biopellets results in a comprehensive systems biotechnological approach. Macroscopic observed growth phenomena as well as growth abnormality and limitations should be predicted by gene expression data of morphogenesis markers. Protein synthesis and secretion as well as the gene expression of stress and morphology markers are analysed and quantified by quantitative reverse transcription polymerase chain reaction (RT-PCR) and enzyme activity tests level, respectively. Furthermore, investigations have been carried out with an approach in systems biology taking into account transcription, translation, and metabolic activity in order to determine and understand the crucial factors which affect the formation and secretion of high value-added products such as recombinant proteins. The data obtained from the transcriptome and proteome analyses under defined experimental conditions during batch or continuous cultivation, together with the information from the metabolome and metabolic flux analyses (including ^{13}C-labelling experiments), can complement the systems biotechnology studies on *A. niger*.

2 Microscopic Morphology

Mycelian growth of filamentous fungi can be differentiated into micro- and macroscopic morphology. First approaches to describe fungal morphology from microscopic images have been made by Metz and Kossen [6]. Characterisation of fungal

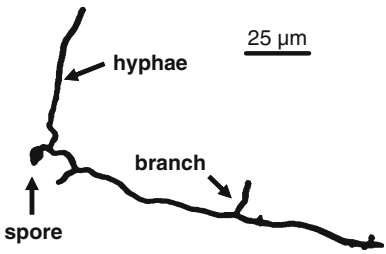


Fig. 2 A growing mycelium of *A. niger*, comprising several tips. The total hyphal length is obtained by summing up the length of all branches [11]

microscopic morphology in the early phase of cultivation can be described by the average and total hyphal length, which is obtained by the sum of all hyphal lengths in a mycelium, the number of tips and the branching of individual hyphae [6–8]. A fundamental relationship between tip growth and hyphal branching was the hyphal growth unit, which was defined as the ratio between the total hyphal length and the number of hyphal tips [9]. These deterministic processes interfere with stochastic components, e.g. the locus of branching, which assesses the directions of branching and growth [10]. All these parameters can be determined by digital image analysis (Fig. 2). Applied techniques consisted of documentation of morphogenesis of single mycelia in growth chambers and image analysis methods [11, 12].

Exponential growth of mycelia in the early phase of cultivation has been described with the specific hyphal length growth rate taken into account, together with the facts that the water content and the hyphal density as well as the hyphal diameter and the growth rate are constant, and no fragmentation occurred. Overall mycelial growth consists of an increase in length by polarised growth of each tip and of an increase of the number of tips caused by branching. Total hyphal length growth is reproduced with a constant tip growth rate of all tips. The branching process was found to be dependent on hyphal length and has been characterised with the local average branching constant [13]. Otherwise the growth of the hyphal length which follows the specific tip growth rate is proportional to the specific hyphal growth rate. Under these conditions the specific length growth rate can be calculated from the tip growth rate and the branching constant [14]. For coagulating filamentous fungi like *A. niger*, conidia aggregate in the early stage of morphologic development before germination and hyphal growth take place [4, 15] and subsequently influence the development of the pellet-type growth [16]. Grimm et al. characterised conidial inocula and seeding-cultures to assess the impact of the aggregation process by direct examination with an in-line particle size analyser (FBRM D600L, Lasentec, USA) [17]. The number of particles is influenced by conidial aggregation and particle morphology cannot be described by tip growth and branching alone. The authors found out that conidial aggregation consists of two distinct aggregation steps (Fig. 3).

The first step starts immediately after inoculation with conidia and forms conidial packages. It leads to a decrease of the total particle concentration in the first hour of cultivation and ends at a dynamic equilibrium. The particle

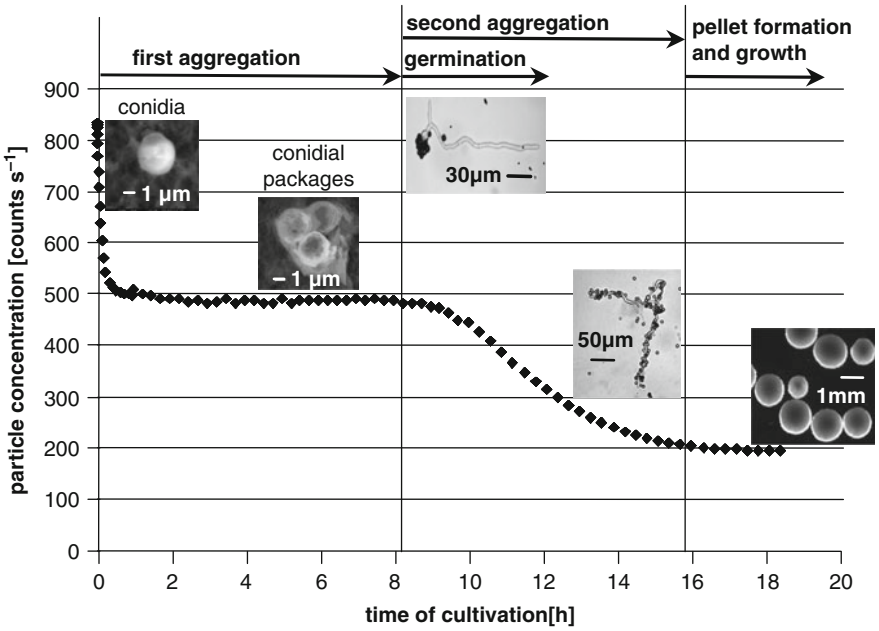


Fig. 3 Conidial aggregation of *A. niger* AB1.13 based on experimental results [17]

concentration remains constant through two inverse reactions: aggregation of conidia to conidial packages and disintegration of these packages. This process is dominantly affected by the pH value. The steady state is then disturbed by the germination of conidia. Germination and hyphal growth of germ tubes entails an increase in hyphal surface area, where conidia can attach to and trigger a second aggregation step. The hyphal growth rate was determined to be the driving force for this aggregation step [4]. This leads to a sharp decrease in particle concentration and depends on the pH value as well as on agitation and aeration induced power input [17, 18].

Lin et al. [19] simulated the aggregation mechanisms of *A. niger* AB1.13 by using the application software PARticle SIze eVALution (Computing in Technology, Germany) and validated the model with kinetic data of Grimm et al. [4, 20]. The derived population model is regarded as a platform for the precise description of the aggregation and growth processes. The population dynamics of the first aggregation step is a result of the formation and disappearance by aggregation and breakage of a class of particle with a certain size. Assumptions that the probability of collision of all particles is conterminous and breakage of all aggregates is independent of their sizes simplify the model of the first aggregation step. For simulating the germination, two parameters, the time of the maximal germination rate and the standard deviation of this, were adopted from published results of *Penicillium chrysogenum* [21] because these values could not be estimated for coagulating filamentous fungi. With the additional definition of the amount of

conidia per unit area, assuming random conidial attachment to hyphae, all parameters for the second aggregation are defined. Improvements are necessary, particularly regarding the description of the aggregate breakage. In respect of the second aggregation step, the population model should also consider the metabolism of the filamentous fungus. Orientation towards certain intracellular molecular markers, which reflect the activity and vitality of cells, is thus conceivable. It requires a thorough search and characterisation of the cell activity.

A morphologically structured model for *A. niger* was proposed by Bizukojc and Ledakowicz [22]. This model was based on a mathematical modelling framework formulated by Nielsen and Villadsen [23]. It took into consideration extracellular components that were detected from the cultivation broth. The model balances the fungal biomass in different hyphal zones of different physiological and functional states and included the effects of the most important nutrients and products, i.e. carbon sources, nitrogen source and citric acid. A direct linear correlation between a less metabolically active and more vacuolated area and the citric acid excretion has been established, confirming that this particular hyphal zone was responsible for acid excretion. Good agreement was found between the model and experimental data obtained under various process conditions. Papagianni [24] suggests the prediction of certain fungal morphologies in correlation with different environmental conditions and the calculation of the fractal mycel geometry.

3 Micromechanic Properties of Filamentous Fungi

Filamentous fungi react to several environmental conditions with a change in their morphology. When taking the aggregation steps into account (compare Fig. 3), it can be suggested that the change in morphology is dependent on the change in adhesion forces between spores or towards hyphae. AFM allows these adhesion forces to be measured directly [25–28]. However, the comparative difficulty in the handling of biological materials has resulted in such measurements being more frequently used for the investigation of non-biological systems. Besides the direct measurements of adhesion forces, other techniques are used for the characterisation of surface forces from *A. niger*. Ryoo and Choi derived a hydrophobic surface behaviour from contact angle measurements [29]. Fujita et al. considered van der Waals as the main acting forces from shear tests [30]. The influence of the pH value and the salt concentration can be characterised by zeta potential measurements. It could be shown, that the isoelectric point of *A. niger* spores is around pH 2. At higher pH values, the absolute zeta potential increases. This corresponds to the fact that the amount of aggregated cells decreases at higher pH values.

There are three types of long-range interactions between particles that are generally used to describe adhesion forces: van der Waals forces, electrical double layer forces and steric forces. The surface properties of particles in a suspension – in this case spores in growth media stirred in a bioreactor – can be described as the

sum of all single molecular interactions such as non-polar, polar or electrostatic interactions on the surface of the particle. In addition, hydrophobic and solvation forces may be important [27, 31]. Biological particles like spores or hyphae show complex surface assemblies that include proteins, sugars and cell wall components. Therefore a computational approach that can be used for inorganic particles is not possible [32]. This makes an experimental approach necessary.

The measurement of adhesion forces with an AFM was developed in the last 10 years to become a standard tool. Nowadays most AFM are able to execute a force measurement test and to analyse the data statistically. However, there are still several challenges to be overcome, especially for living cells.

A set-up for the measurement of adhesion forces is shown in Fig. 4 (left). The heart of an AFM is the cantilever. It works like a small and thin leaf-spring which is connected to a piezo drive to move the cantilever up and down with nanometer resolution. Bending of the cantilever is detected by a laser reflection which is mirrored to a position-sensitive photo-detector (pspd). The correlation between the bending of the cantilever and the pspd-signal is called sensitivity and has to be calibrated (see below). The detector signal combined with the spring constant of the cantilever can be used to calculate the force applied to a spore that has been fixed at the end of the cantilever.

For the measurement of spore/spore interactions it is also necessary to fix spores to a substrate (e.g. a wafer slide), which is attached to the xy-desk of the AFM to move the sample in the x and y directions and align two spores on top of each other. Then the z-piezo drive moves down and the spores achieve contact. Subsequently, moving the piezo drive upwards separates the spores. In doing so, the cantilever bends downwards, the degree being corresponding to the applied force. Adhesion forces can be calculated from the minimum force of the retraction curve and the value at the cantilever equilibrium position. At the point where the spores become separated, the adhesion force can be calculated. Figure 4 (right) shows a typical force/distance curve between two spores.

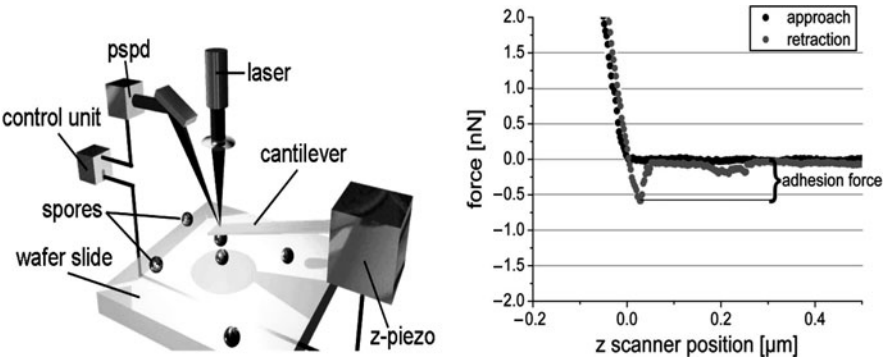


Fig. 4 Schematic of the assembly for the measurement of adhesion forces (*left*). Force way curve from the measurement of adhesion forces (*right*)

As mentioned above, a spore that is firmly attached on a substrate is necessary for the measurement. Otherwise the spore will be detached while retracting the cantilever. A hydrophobic surface has been shown to be suitable such as a glass or more simple a wafer that has been treated with a hydrophobic silane. The adhesion of the spores on the modified surface is strong enough for adhesion tests.

To fix a spore at the end of a cantilever, glue can be applied under a microscope equipped with a micromanipulator. For the handling of biological materials, the glue has to be chosen to avoid damage to cells by solvents or UV-light (used to set the glue). Good results could be obtained with glues which can be activated by blue light and then set in a short time period. In most cases, this method requires that the cell is removed from its media for a few minutes. An elegant alternative method for the attachment of bakers yeast was developed by Götzinger et al. [33]. The authors immobilised the sugar-binding protein Concanavalin A onto the cantilever and used it to attach the cell directly from the growth media by binding of the sugar rich cell-surface.

The next challenge in the measurement of adhesion forces is the determination of the sensitivity of the system. The sensitivity is the correlation of the cantilever deflection with the pspd-voltage-signal, and this must be determined for each experiment. Usually the tip of the cantilever is pressed on a glass slide (as an infinitely hard surface) and the pspd-signal is recorded. In this case the movement of the z-piezo drive is equal to the deflection of the cantilever. If a cell is attached to the cantilever these data will be influenced by the elastic deformation of the cell. A glass bead immobilised on the glass slide can be pressed with the cantilever right behind the spore, allowing the sensitivity to be measured accurately. In addition, the spring constant of the cantilever has to be determined – this can be done in several ways [34]. One possibility is the thermal noise method which is described by Hutter and Bechhoefer [35]. These measurements are usually carried out without the attached spore. The influence of the spore on the spring constant of the cantilever is assumed to be negligible due its small size in proportion to the cantilever.

After determination of the constants, the measurement can be performed. This requires careful handling to achieve the difficult task of aligning the two spores directly on top of each other. The spores must be exactly aligned, or shear forces will affect the result. However, looking from above or below, the cantilever blocks the line of sight. A very sharp-ended cantilever has been shown to be suitable to overcome this problem. Attaching the spore at the very end of the cantilever, the position of the spore on the cantilever is determined and the measurements can be performed. However, many measurements have to be made for the characterisation of the spore/spore adhesion, and the biological diversity makes a statistical approach for the interpretation of the data essential.

Adhesion measurements with spores of *A. niger* towards mica by AFM were done in water and in air [25, 26]. While the adhesion forces in air were strong (30–110 nN) the measurements in water showed only 2–25 nN. The measurements of the adhesion forces between spores and the influence of the pH value and the ionic strength are currently progressing.

4 Interaction Between Morphology, Mass Transfer and Reaction

After the initial phases of aggregation and germination, which have been described above, the hyphae start to form pellets. Typically filamentous fungi grow exponentially with constant specific growth rate until substrate limitation occurs [36]. The macroscopic growth of fungi to pellets is generally described with the pellet radius and the "critical" radius. The "critical" radius indicates the point where diffusion limitations occur in the outer shell of the pellet. Hyphae which extend into the bulk phase and deviate the pellet from an ideal-spheric form are characterised as "pellet hair". The average length of main hyphae of a free mycelium filament is in the range of approximately 100–400 μm. The radius of pellet can be in the range between 250 and 2,500 μm, depending on the environmental conditions of cultivation. Figure 5 gives an indication of the main processes which have to be considered for a growth model.

One of the first macro-morphologic growth models is the cube-root law based on observations of Emerson on the growth of *Nocardia crassa* [38]. The law describing pellet growth assumes that a substrate penetrates in the outer shell core of the constant length. The remaining biomass in the inner shells does not participate in any further reaction. Assuming a homogenous pellet density along the radial coordinate, the pellet biomass concentration is proportional to the cultivation time to the power of 3. Most pellet growth can be described in a first approximation by this cubic root law satisfactorily [39]. Marshall and Alexander [40] and Pirt [41] have shown this for different fungi and actinomycetes.

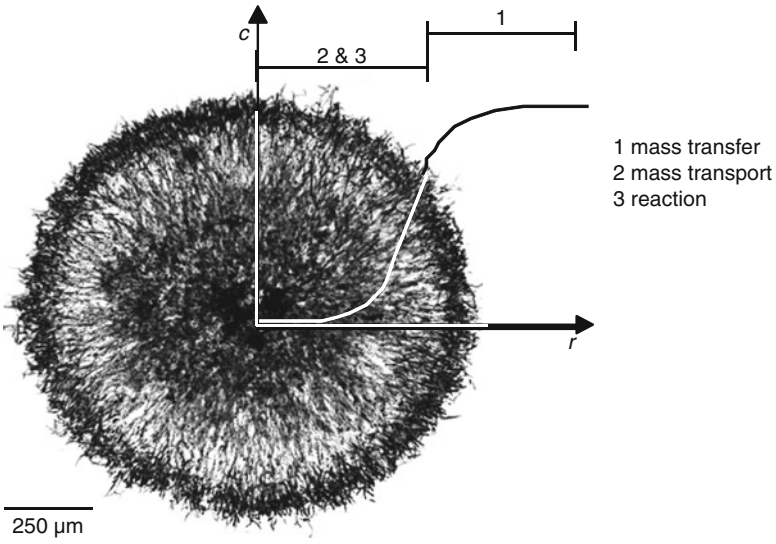


Fig. 5 Principle sketch of processes within a biopellet. Shown is a confocal laser scanning microscopy (CLSM) image of a cryomicrotome slice of *A. niger* AB1.13 [37]

Nevertheless, it has been shown that fungal pellet density varies over the radial coordinate [42] and changes with time of cultivation [43, 44]. Pellet morphology, which is a result of environmental conditions (compare Fig. 10) and pellet age has an important influence on the effective diffusion coefficient and the penetration depth of oxygen into the pellet [45–48]. These properties have been studied with microelectrode measurements. Oxygen concentration profiles have been presented for pellets of different filamentous fungi [47–52].

For the spatial resolution of pellet structure, biomass distribution confocal laser scanning microscopy (CLSM) in combination with digital image analysis has been introduced. Furthermore, a method to quantify the hyphal distributions in pellets by CLSM has been developed [49]. The CLSM-images have been used to define the morphological parameter (dh/dr), which is a gradient describing the hyphal fraction h within the outer pellet periphery (Fig. 6).

Detailed studies of hyphal or biomass distribution along the pellet radius reveal a huge range of pellet morphologies during cultivation [47]. It can be shown that both hairy and compact structures in the outer pellet periphery can lead to completely different oxygen consumption rates. This can still be valid for pellets with identical mean pellet densities. Table 1 shows results for two pellets with nearly the same radius and density but different surface morphology characterised by the gradient of the hyphal fraction h within the outer pellet periphery (dh/dr). Also, the mean biomass density is comparable; the maximum oxygen gradient differs by a factor of 2. This is reflected by dh/dr, which seems to be more reliable than the overall density.

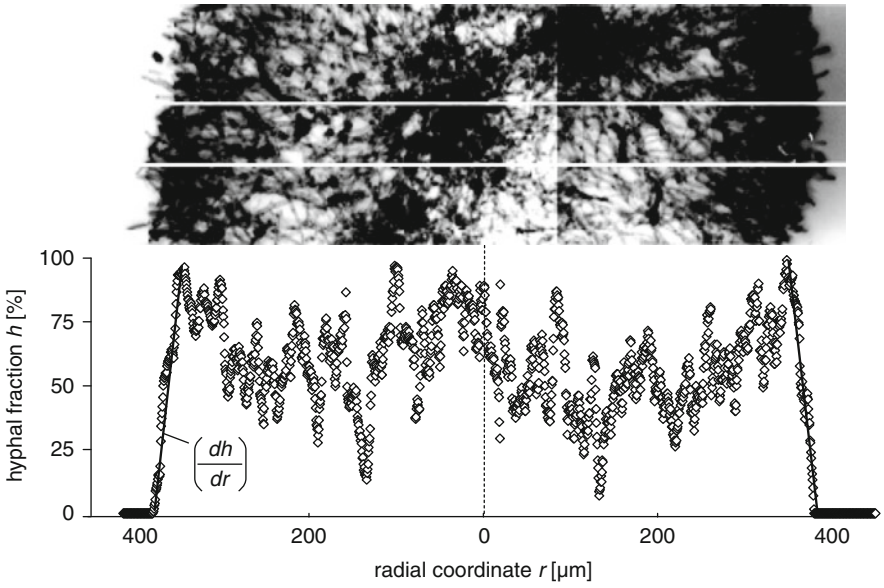


Fig. 6 CLSM-image showing a cryomicrotome slice of a pellet with the gradient of the hyphal fraction dh/dr [37]

Table 1 Characteristics of two pellets with comparable radius and biomass density but different surface morphology (dh/dr)

	Pellet radius (μm)	Pellet density (kg m^{-3})	Pellet age (h)	(dh/dr) (% μm^{-1})	(dc/dr)$_{max}$ (kg m^{-3} m^{-1})
Pellet 1	555	53	41	0.95 ± 0.03	121.9
Pellet 2	490	43	40.5	0.10 ± 0.009	66.5

(dc/dr)$_{max}$ is the maximum oxygen gradient, which is calculated from oxygen profiles [37]

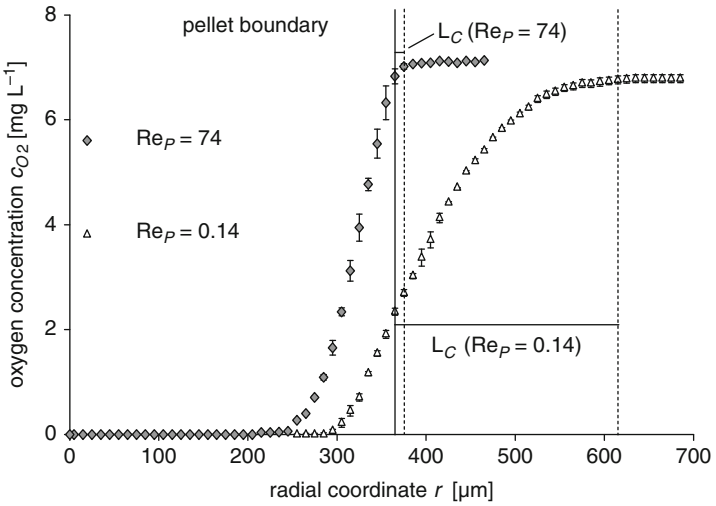


Fig. 7 Oxygen concentration profiles with concentration boundary layer L_c for two different Reynolds number at the pellet Re_P. (Pellet age 34 h, pellet radius 375 μm, (dh/dr) = 0.48% μm^{-1}, pellet density = 64.4 kg m^{-3}) [37]

Figure 7 shows two oxygen profiles measured within the same pellet at two different flow conditions. The used pellet was very compact and it can be assumed that the surface was nearly rigid. Typically the thickness of the concentration boundary layer L_c is dependent on the flow velocity.

With the indifferent surface structures, which can be manipulated by flow pressure, the classic concentration boundary theory is no longer valid. The moving bulk is able to penetrate the filamentous structures and can thereby increase the mass transport [47]. Furthermore, the flow pressure might compact the filamentous structure. This will also lead to a higher turnover based on a denser biomass.

5 Effect of Fluid Dynamics on Fungal Growth

A main task of the research in numerical simulation and experimental validation of the fluid dynamic effects in baffled stirred tank reactors (STRs) are the identification and quantification of mechanical stress on filamentous microorganisms.

Usually, Rushton turbines are used for the dispersion of oxygen and the homogenisation of the liquid phase regarding the nutrient supply. Factors that fluid-dynamically influence the environmental conditions of the cultivation are the geometry of the reactor, the dissipated energy through stirring and aeration, and the stirring devices. On the one hand, too slight stirring yields insufficient nutrient and oxygen supply, on the other hand, intense stirring causes injuries to the cell wall, cell disruption, and consequently inhibited growth and product formation [11, 52, 53]. The effect of agitation on fungal morphology has been discussed in several publications [51, 54–61]. The authors have concluded that the total hyphal length of a fungal mycel decreases with increasing specific power input (Fig. 8).

A possible approach to determine the reasons for the occurrence of this negative impact of mechanical stress on fungal growth is the numerical prediction of the flow field in an STR by using computational fluid dynamics (CFD) simulations. This technique affords modelling and calculation of the mean and turbulent velocity field, the pressure and the energy terms. When the influence of the fluid dynamic conditions can be characterised properly, simulations and experimental data from a well investigated shear sensitive non-biological clay-polymer-floc-system shall enable one to quantify and to optimise the cultivation conditions of filamentous fungi regarding mechanical stress [5].

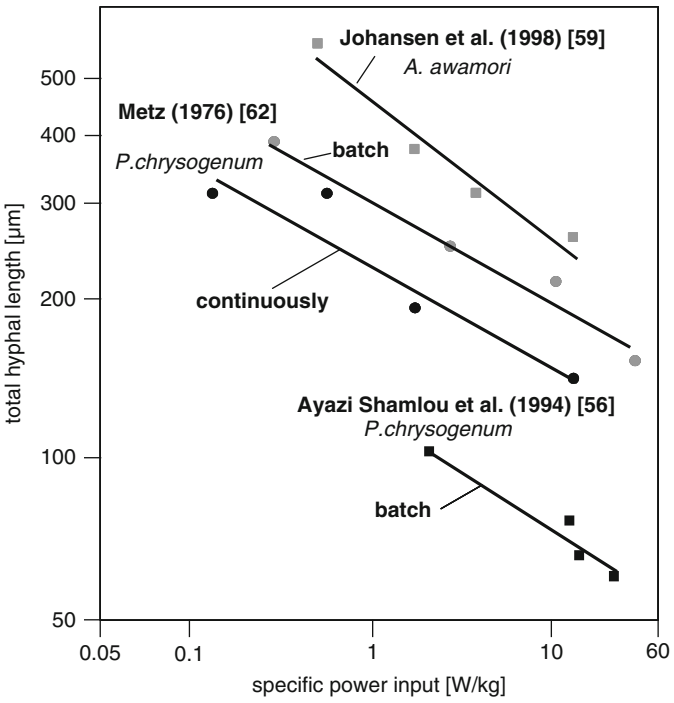


Fig. 8 Total hyphal length vs specific power input

It is assumed, that especially small-scaled eddies in the turbulent flow are the determining factor for mechanical stress on filamentous fungi [16]. The Kolmogorov micro-scale of length is determined by an exponent of the specific power input of approximately −0.25. The fragmentation of the hyphae results in short, compact and highly branched filaments for intensive agitation intensities [55, 56, 58, 62]. In STR, it is assumed to occur in a region of high stress close to the impeller. Changes in morphology have also been observed by the variation of the impeller type [63].

Hence, turbulence modelling in an STR particularly requires the involvement of the smallest eddies in hyphal fragmentation. The direct numerical simulation of turbulent flows requires that all scales are resolved by the computational grid. This is currently only possible for low Reynolds numbers and simple geometries. In principle, two strategies remain: large eddy simulation (LES) based on the decomposition of the flow field into large- and subgrid-scale structures, and Reynolds averaged Navier–Stokes (RANS) turbulence modelling. For LES the large scales are directly computed and the influence of the small scales is modelled. Certainly LES proves to be a well suited technique for the simulation of STR [64, 65]. In the case that velocity and pressure components are broken down into mean and fluctuation values, these terms are inserted into Navier–Stokes equations and time-averaged. As a result, the RANS-equations are obtained. These equations contain an additional term, called Reynolds stress tensor. As there are new unknown variables, a turbulence model for closing the equations is needed. There are several possibilities, e.g. the k-ε-turbulence model as a very well investigated and robust model [66]. However, this model is actually designed for isotropic turbulence [67–69]. In the k-ε-model the transport equations of the turbulent kinetic energy (TKE), which represents the normal stress fraction of turbulent stresses and the isotropic dissipation rate, are employed, which can be closed by empirical data. The k-ω-turbulence model deals with transport equations of the TKE and the characteristic frequency of dissipative eddies [70]. The benefit takes on good near-wall treatment. The shear stress transport turbulence modelling merges the benefits of the k-ε-turbulence model at wall-far areas and the k-ω-model at wall-near domains [71].

Appel et al. [72] achieved results by calculation on an unstructured, tetrahedral grid. After a grid convergence study the chosen grid for an STR ($V = 19.3$ L, three-stage Rushton stirrer) has 1.1×10^6 nodes and 6.2×10^6 tetrahedrons with a discretisation of 0.5 mm near the impeller baffles and 0.7 mm within the wake.

The experimental validation of the numerical simulation was carried out via particle image velocimetry (PIV) in a one-phase (liquid) operated STR with equal dimensions of the bioreactor for cultivation of *A. niger*. The STR was placed in a pentagonal, water filled container to avoid distortion of the laser light sheet on the convex surface of the reactor and to minimise the influence of the air/Perspex interface [73, 74]. PIV was used to investigate the area around the middle-positioned stirrer.

Figure 9 shows that the CFD simulation and the PIV measurement are good regarding consistency of distribution and intensity of TKE. The numeric grid is fine

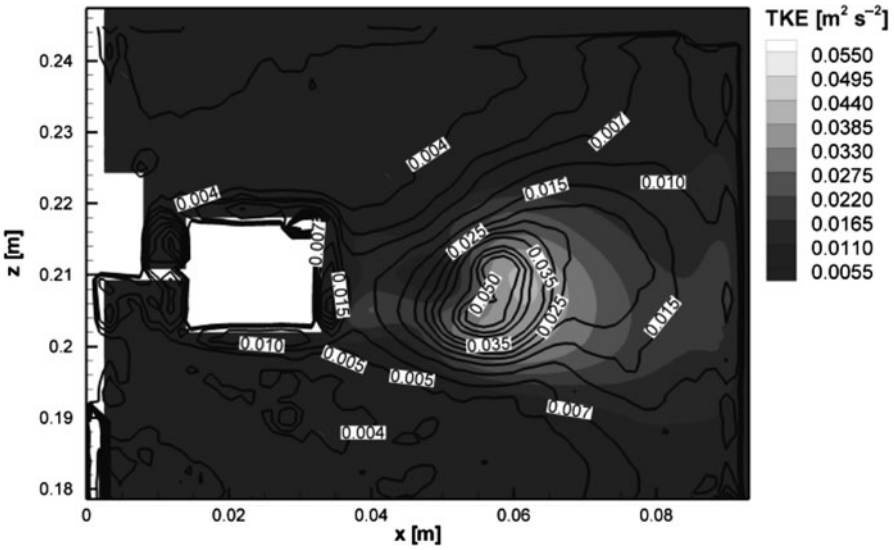


Fig. 9 Comparison of simulated data (CFD, *grey area* scheme) and PIV-data (*contour lines*): turbulent kinetic energy (TKE) of a stirrer blade plane, z: axial, x: radial direction of the STR [72]

enough, so the discretisation error is sufficiently small, and a not excessive blur of the simulated wake flow occurs. Concerning phase angle dependence, the simulation shows two tip vortices at the blade edges [71, 75].

6 Spatial Resolution of Morphology and Biological Function

Averaged morphological characteristics of fungal cultures such as pellet diameter, average biomass density of the aggregates or population balances based on particle size distribution are not sufficient to relate morphological shape to the resulting productivity and understand the underlying mechanisms [53, 76, 77]. Here, the inner structure of the fungal aggregate has to be taken into account.

As shown, the peripheral region of an aggregate determines the transport and reaction processes of the available substrates within the pellet and therefore affects productivity (compare Figs. 6 and 7). Variations in pellet structure occur, for example, due to different mechanical stress by means of different agitation speeds during cultivation. Image analysis accomplished by optical light microscopy, as shown in Fig. 10, demonstrates an increased densification of the peripheral hyphae at enhanced mechanical load which becomes more pronounced with increasing cultivation time [45].

As a consequence of this growing transport barrier, a decline of nutrient availability in the inner part of the aggregate occurs, which results in cell lysis and hollow pellets under conditions of high agitation speed. Such morphological

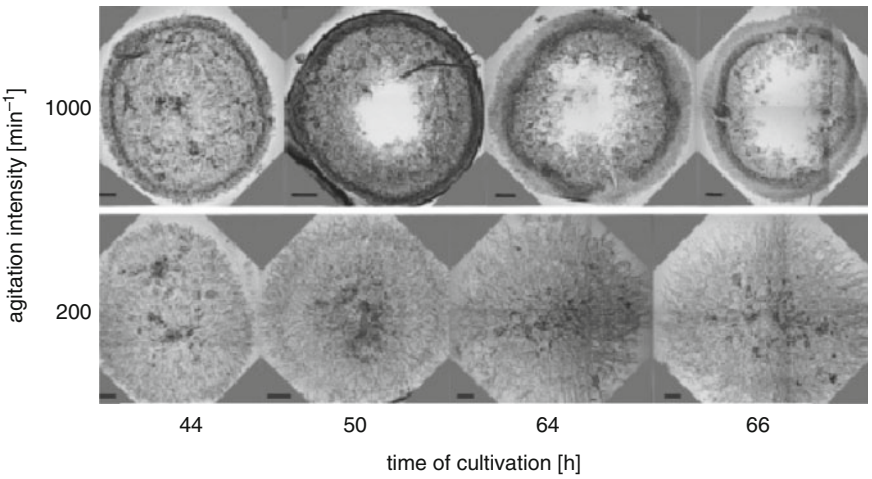


Fig. 10 Optical light microscopic images of pellet slices: different mechanical loading by means of agitation speed (*bar*: 200 μm) [45]

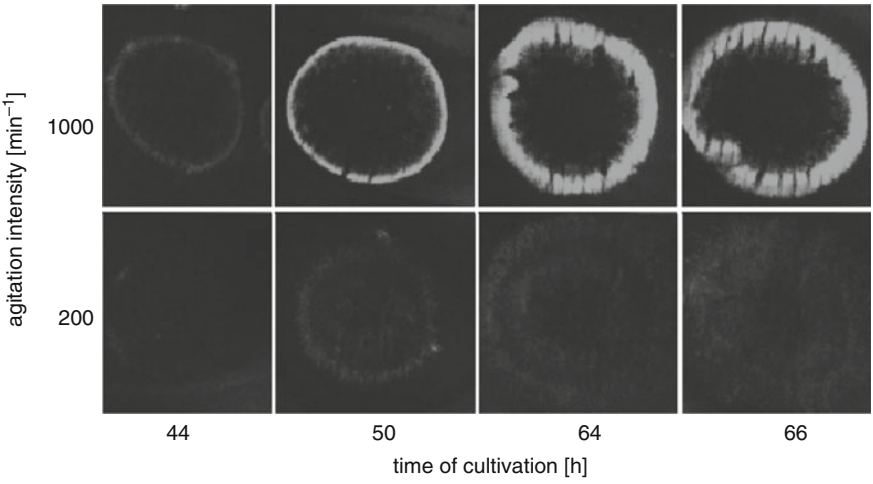


Fig. 11 Confocal laser scanning microscopic (CLSM) images of pellet slices; localisation of GFP-production at different mechanical loading by means of agitation speed [45]

changes of the inner pellet structure have a drastic influence on production performance.

The characteristics of recombinant protein formation in fungal aggregates of different morphological shape can be visualised by employing an *A. niger* strain that produces green fluorescent protein (GFP). Figure 11 shows the localisation of GFP-producing regions in equatorial pellet slices that were analysed with CLSM.

Obviously, only peripheral regions of the pellets contribute to protein production. This seems plausible since the peripheral regions feature a good nutrient supply and therefore develop good growth and productivity. According to that result, it seems desirable to generate high cell densities in these peripheral regions, which can be achieved for the highest mechanical load.

7 Conclusions and Future Perspectives

As shown in this chapter, various studies demonstrate the close link between operating parameters of the bioprocess, morphology of fungal aggregates and the spatially different metabolism of individual cells within these structures. Currently we are far from understanding the underlying metabolic and regulatory mechanisms. Newly arising experimental and computational technologies in systems biology and systems biotechnology, however, now provide a powerful toolbox as a step towards understanding this surely complex link between biological and engineering aspects of fungal cultures [78]. Boosted by sequencing of the genome of *A. niger* [79], omics technologies such as transcriptomics [80], proteomics [81, 82], metabolomics [83] or fluxomics [84] are available for detailed analysis and optimisation of *A. niger* at a systems level.

From early on, experimental research on *A. niger* was accompanied by strong efforts in modelling. Based on knowledge from intensive biochemical studies, condensed stoichiometric models focussing on specific parts of metabolism for *A. niger* have been developed and used for metabolite balancing studies [85–87]. With increasing knowledge, models could be extended to higher complexity and applied for optimisation of succinate production via flux balance analysis [88] or investigation of metabolic fluxes in protein producing cells for different carbon sources [89]. The latter study revealed the interesting finding that the flux through the pentose phosphate pathway under glucoamylase producing conditions was significantly higher than under non-producing conditions which might reflect an increased NADPH demand for production. A first study on metabolic pathway analysis using elementary flux modes provided a detailed insight into the compartmented metabolic network of *A. niger* [90]. It was shown that capacity, pathway usage, and relevant genetic targets for optimal fructofuranosidase production are strongly dependent on the network structure, the available nutrients and the metabolic state of the cell. A milestone in modelling of *A. niger* displays the recently created genome scale metabolic model [91]. The complexity of this eukaryotic microorganism becomes obvious from the fact that the metabolic network comprises 1,190 biochemical reactions and 1,045 metabolites in different compartments of the cell. Beyond being an excellent validated comprehensive bibliography on *A. niger* metabolism, the available genome scale model provides a sound basis for systems biology studies. As an example, combination of transcriptomics with genomic data provided a better understanding of the complex pH-regulation in *A. niger* [92]. Admittedly, all these systems oriented studies so far do not consider a

spatial resolution of biological function. This appears, however, crucial, especially for filamentous fungi, known for their complex morphology. The resulting gradients of carbon source, products and oxygen together with the characteristics of hyphal growth and metabolic activity mainly at hypheal tips unambiguously point at a complex spatial distribution of gene expression patterns or metabolic pathway fluxes. Here, fluorescence based methods applicable to localise specific proteins or the expression of specific genes as well as new mass spectrometry approaches able to resolve spatially metabolite profiles within cellular structures will be of great benefit.

At the same time, the link of the biological system to the environment of the bioreactor demands an extended resolution of its fluid dynamic properties. Here, measurement and validation of local flow fields in bioreactors is already possible and can be applied to replace the simple correlation based on the Kolmogorov micro-scale of length. Findings should couple models of the fungal morphogenesis in order to determine local mechanical forces which may affect the biomass. For this purpose, previous simulation methods and data of fluid dynamics within the bioreactor, should be generated with CFD and verified by experimental fluid dynamics (e.g. PIV), and ought to be linked to the shear stress of well examined non-biological clay-polymer-floc-systems. The resulting integral data of the mechanical forces have to be correlated with local CFD-data to obtain a fluid dynamic model of shear stress of filamentous microorganisms in STRs. In later stages, the dynamic processes of growth, aggregation and break-up during submerged cultivations might be connected by population balancing. In future, the description of morphology will also have to move from assuming average size of hyphae, aggregates or pellets to reproduction of tailoring size or radial density distributions. Methods of population balancing will help to model and characterise morphologic developments, aiming to connect fluid dynamic dependencies with intracellular biological reactions and productivity in more comprehensive models for rational, target-oriented design of cultivation processes with filamentous fungi.

Acknowledgements The authors gratefully acknowledge financial support provided by the German Research Foundation (DFG) through the Collaborative Research Center SFB 578 “From Gene to Product” at the Technische Universität Braunschweig, Germany. Special thanks are given to the colleagues involved in the different SFB-projects: Christina Appel, Kathrin Bohle, Markus Emmler, Alex Dalpiaz, Matthias Gehder, Yvonne Göcke, Luis H. Grimm, Timo Hagemann, Andrea Hille, Rochus Jonas, Anke Jungebloud, Christian J. Kähler (Institute of Fluid Mechanics and Aerodynamics, Universität der Bundeswehr München), Sven Kelly, Katina Kiep, Pey-Jin Lin, Xin Lu, Guido Melzer, Becky Sommer, Alexander Stintzing, Andreas Wargenau and Michael Wulkow (CiT GmbH, Rastede, Germany).

References

1. Cui YQ, Okkerse WJ, van der Lans RGJM, Luyben KChAM (1998) Biotechnol Bioeng 60:216
2. Hagemann T, Ringel A, Hempel DC, Krull R (2007) Biospektrum 2007. In: Proceedings of VAAM-Jahrestagung, Osnabrück, Germany

3. Lin PJ, Hagemann T, Stintzing A, Appel C, Hempel DC, Krull R (2008) In: Proceedings of European Bioperspectives 2008, Book of Abstracts, Hanover, Germany, p 50
4. Grimm LH, Kelly S, Völkerding II, Krull R, Hempel DC (2005) Biotechnol Bioeng 92:879
5. Stintzing A, Pilz R, Hempel DC, Krull R (2008) Chem Ing Tech 80:1837
6. Metz B, Kossen NWF (1977) Biotechnol Bioeng 14:781
7. Trinci APJ (1974) J Gen Microbiol 81:225
8. Schuhmann E, Bergter F (1976) Z Allg Mikrobiol 16:201
9. Caldwell IY, Trinci APJ (1973) Arch Microbiol 88:1
10. King R (1995) Chem Ing Tech 67:553
11. Spohr A, Dam-Mikkelsen C, Carlsen M, Nielsen J, Villadsen J (1998) Biotechnol Bioeng 58:541
12. Paul GC, Thomas CR (1998) In: Schügerl K (ed) Relation between morphology and process performances. Adv Biochem Eng Biotechnol, vol 60. Springer, Berlin, p 2
13. Grimm LH, Kelly S, Krull R, Hempel DC (2005) Appl Microbiol Biotechnol 49:375
14. Bergter F (1978) Z Allg Mikrobiol 18:143
15. Papagianni M (2004) Biotechnol Adv 22:189
16. Kelly S, Grimm LH, Bendig C, Hempel DC, Krull R (2006) Process Biochem 41:2113
17. Grimm LH, Kelly S, Hengstler J, Göbel A, Krull R, Hempel DC (2004) Biotechnol Bioeng 87:213
18. Amanullah A, Leonili E, Nienow AW, Thomas CR (2001) Bioprocess Biosyst Eng 24:101
19. Lin PJ, Grimm LH, Wulkow M, Hempel DC, Krull R (2008) Biotechnol Bioeng 99:341
20. Grimm LH (2006) In: Hempel DC (ed) ibvt-Schriftenreihe, vol 27. FIT-Verlag, Paderborn
21. Paul GC, Kent A, Thomas CR (1993) Biotechnol Bioeng 42:11
22. Bizukojc M, Ledakowicz S (2006) Process Biochem 41:1063
23. Nielsen J, Villadsen J (1994) Bioreaction engineering principles. Plenum, New York, London
24. Papagianni M (2006) Microb Cell Fact 5:1
25. Bowen WR, Lovitt RW (2000) Colloids Surf 173:205
26. Bowen WR, Lovitt RW (2000) J Colloid Interface Sci 228:428
27. Luckham PF (2004) Adv Colloid Interface Sci 111:29
28. Müller DJ, Dufrêne YF (2008) Nat Nanotechnol 3:261
29. Ryoo D, Choi CS (1999) Biotechnol Lett 21:97
30. Fujita M, Iwahori K (1994) J Ferment Bioeng 78:368
31. Israelachvili JN (1991) Intermolecular and surface forces. Academic, San Diego
32. Müller FW, Peukert W (2004) Int J Miner Process 74S:31
33. Götzinger M, Weigl B (2007) Colloids Surf B Biointerfaces 55:44
34. Butt HJ, Cappella B (2005) Surf Sci Rep 59:1
35. Hutter JL, Bechhoefer J (1993) Rev Sci Instrum 64:1868
36. Rinas U, El-Enshasy H, Emmler M, Hille A, Hempel DC, Horn H (2005) Chem Eng Sci 60:2729
37. Hille A (2008) In: Hempel DC (ed) ibvt-Schriftenreihe, vol 32. FIT-Verlag, Paderborn
38. Emerson S (1950) J Bacteriol 60:221
39. Prosser JI (1994) In: Gow NAR, Gadd GM (eds) The growing fungus. Chapman & Hall, London, p 319
40. Marshall KC, Alexander M (1960) J Bacteriol 80:412
41. Pirt SJ (1966) Proc R Soc B 166:369
42. Hamanaka T, Higashiyama K, Fujikawa S, Park EY (2001) Appl Microbiol Biotechnol 56:233
43. El-Enshasy H, Hellmuth K, Rinas U (1999) Appl Biochem Biotechnol 81:1
44. Hellendoorn L, Mulder H, Heuvel JCvd, Ottengraf SPP (1998) Biotechnol Bioeng 58:478
45. Bohle K (2009) In: Hempel DC (ed) ibvt-Schriftenreihe, vol 40. FIT-Verlag, Paderborn
46. Cui YQ, Lans RGJMvd, Luyben KCAM (1998) Biotechnol Bioeng 57:409
47. Hille A, Neu TR, Hempel DC, Horn H (2006) Chem Ing Tech 78:627
48. Hille A, Neu TR, Hempel DC, Horn H (2009) Biotechnol Bioeng 103:1202
49. Hille A, Neu TR, Hempel DC, Horn H (2005) Biotechnol Bioeng 92:614

50. Cronenberg CCH, Ottengraf SPP, Heuvel JCvd, Pottel F, Sziele D, Schügerl K, Bellgardt KH (1994) Bioproc Eng 10:209
51. Wittler R, Baumgartl H, Lübbers DW, Schügerl K (1986) Biotechnol Bioeng 28:1024
52. Huang MY, Bungay HR (1973) Biotechnol Bioeng 24:1193
53. Kelly S, Grimm LH, Hengstler J, Schultheis E, Krull R, Hempel DC (2004) Bioprocess Biosyst Eng 26:315
54. Papagianni M, Mattey M, Kristiansen B (1998) Biochem Eng J 2:197
55. Metz B, Bruijn Ewd, Suijdam JCv (1981) Biotechnol Bioeng 23:149
56. Suidjdam JCv, Metz B (1981) Biotechnol Bioeng 23:111
57. Ayazi Shamlou P, Makagiansar HY, Ison AP, Lilliy D, Thomas CR (1994) Chem Eng Sci 49:2621
58. Fujita M, Iwahori KST, Yanakawa K (1994) J Fermet Bioeng 78:368
59. Cui YQ, Lans RGJMvd, Luyben KCAM (1997) Biotechnol Bioeng 55:715
60. Johansen CL, Coolen L, Hunik JH (1998) Biotechnol Prog 14:233
61. Amanullah A, Christensen LH, Hansen K, Nienow AW, Thomas CR (2002) Biotechnol Bioeng 77:815
62. Amanullah A, Jüsten P, Davies A, Paul GC, Nienow AW, Thomas CR (2000) Biochem Eng J 5:109
63. Jüsten P, Paul GC, Nienow AW, Thomas CR (1996) Biotechnol Bioeng 52:672
64. Alkamo R, Micalea G, Grisafia F, Brucatoa A, Ciofalo M (2005) Chem Eng Sci 60:2303
65. Revstedt J, Fuchs L (2002) Chem Eng Technol 25:443
66. Chien KY (1982) AIAA J 20:33
67. Jones W, Launder B (1972) Int J Heat Mass Transf 15:301
68. Launder B, Sharma B (1974) Lett Heat Mass Transf 1:131
69. Launder B, Spalding D (1990) Comput Methods Appl Mech Eng 3:269
70. Wilcox D (1988) AIAA J 26:1414
71. Menter FR (1994) AIAA J 32:1598
72. Appel C, Hagemann T, Kähler CJ, Krull R (2007) Proc 10. Köthener Rührer- Kolloquium, Hochschule Anhalt (FH), 39–50, ISBN-13: 978-3-86011-015-7, Köthen
73. Prasad AK, Adrian RJ (1993) Exp Fluids 15:49
74. Prasad AK, Jensen K (1995) Appl Opt 34:7092
75. Appel C, Hagemann T, Lin PJ, Stintzing A, Kähler CJ, Krull R (2007) In: Proceedings of the International Conference on Multiphase Flow ICMF 2007, Leipzig
76. Rodriguez Porcel EM, Casas Lopez JL, Sanchez Perez JA, Fernandez Sevilla JM, Chisti Y (2005) Biochem Eng J 26:139
77. Casas Lopez JL, Sanchez Perez JA, Fernandez Sevilla JM, Rodriguez Porcel EM, Chisti Y (2005) J Biotechnol 116:61
78. Visser J (2009) Eurofung. Fungal Genet Biol 46:S1
79. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K, Andersen MR, Bendtsen JD, Benen JA, van den Berg M, Breestraat S, Caddick MX, Contreras R, Cornell M, Coutinho PM, Danchin EG, Debets AJ, Dekker P, van Dijck PW, van Dijk A, Dijkhuizen L, Driessen AJ, d'Enfert C, Geysens S, Goosen C, Groot GS, de Groot PW, Guillemette T, Henrissat B, Herweijer M, van den Hombergh JP, van den Hondel CA, van der Heijden RT, van der Kaaij RM, Klis FM, Kools HJ, Kubicek CP, van Kuyk PA, Lauber J, Lu X, van der Maarel MJ, Meulenberg R, Menke H, Mortimer MA, Nielsen J, Oliver SG, Olsthoorn M, Pal K, van Peij NN, Ram AF, Rinas U, Roubos JA, Sagt CM, Schmoll M, Sun J, Ussery D, Varga J, Vervecken W, van de Vondervoort PJ, Wedler H, Wösten HA, Zeng AP, van Ooyen AJ, Visser J, Stam H (2007) Nat Biotechnol 25:221
80. Andersen MR, Vongsangnak W, Panagiotou G, Salazar MP, Lehmann L, Nielsen J (2008) Proc Natl Acad Sci USA 105:4387
81. Tsang A, Butler G, Powlowski J, Panisko EA, Baker SE (2009) Fungal Genet Biol 46:153
82. Wright JC, Sugden D, Francis-McIntyre S, Riba-Garcia I, Gaskell SJ, Grigoriev IV, Baker SE, Beynon RJ, Hubbard SJ (2009) BMC Genomics 10:61

83. Frisvad JC, Rank C, Nielsen KF, Larsen TO (2009) Med Mycol 47:53
84. Pedersen H, Christensen B, Hjort C, Nielsen J (2000) Metab Eng 1:34
85. Prathumpai W, Gabelgaard JB, Wanchanthuek P, van de Vondervoort PJ, de Groot MJ, McIntyre M, Nielsen J (2003) Biotechnol Prog 19:1136
86. Alvarez-Vasquez F, Gonzalez-Alcon C, Torres NV (2000) Biotechnol Bioeng 70:82
87. Torres NV (1994) Biotechnol Bioeng 44:104
88. David H, Akesson M, Nielsen J (2003) Eur J Biochem 270:4243
89. Melzer G, Dalpiaz A, Kucklick M, Göcke Y, Jonas R, Dersch P, Franco-Lara E, Nörtemann B, Hempel DC (2007) J Biotechnol 132:405
90. Melzer G, Eslahpazir M, Franco-Lara E, Nörtemann B, Hempel DC, Wittmann C (2009) BMC Syst Biol 3:120
91. Andersen MR, Nielsen ML, Nielsen J (2008) Mol Syst Biol 4:178
92. Andersen MR, Lehmann L, Nielsen J (2009) Genome Biol 10:R47

Adv Biochem Engin/Biotechnol (2010) 121: 23–43
DOI: 10.1007/10_2009_53

Published online: 6 February 2010

Multi-Scale Spatio-Temporal Modeling: Lifelines of Microorganisms in Bioreactors and Tracking Molecules in Cells

Alexei Lapin, Michael Klann, and Matthias Reuss

Abstract Agent-based models are rigorous tools for simulating the interactions of individual entities, such as organisms or molecules within cells and assessing their effects on the dynamic behavior of the system as a whole. In context with bioprocess and biosystems engineering there are several interesting and important applications. This contribution aims at introducing this strategy with the aid of two examples characterized by striking distinctions in the scale of the individual entities and the mode of their interactions. In the first example a structured-segregated model is applied to travel along the lifelines of single cells in the environment of a three-dimensional turbulent field of a stirred bioreactor. The modeling approach is based on an Euler-Lagrange formulation of the system. The strategy permits one to account for the heterogeneity present in real reactors in both the fluid and cellular phases, respectively. The individual response of the cells to local variations in the extracellular concentrations is pictured by a dynamically structured model of the key reactions of the central metabolism. The approach permits analysis of the lifelines of individual cells in space and time.

The second application of the individual modeling approach deals with dynamic modeling of signal transduction pathways in individual cells. Usually signal transduction networks are portrayed as being wired together in a spatially defined manner. Living circuitry, however, is placed in highly malleable internal architecture. Creating a homogenous bag of molecules, a well-mixed system, the dynamic behavior of which is modeled with a set of ordinary differential equations is normally not valid. The dynamics of the MAP kinase and a steroid hormone pathway serve as examples to illustrate how single molecule tracking can be linked with the stochasticity of biochemical reactions, where diffusion and reaction occur in a probabilistic manner. The problem of hindered diffusion caused by macromolecular crowding is also taken into account.

M. Reuss (✉), A. Lapin, and M. Klann
Institute of Biochemical Engineering and Center Systems Biology, University of Stuttgart, Allmandring 31, 70569, Stuttgart, Germany
e-mail: reuss@ibvt.uni-stuttgart.de

Keywords Agent-based modeling, CFD, Stirred tank bioreactor, Glycolysis of *E. coli* populations, PTS, Signal transduction, Stochastic differential equations, Fokker-Planck, Macromolecular crowding, MAPK cascade, Steroid hormone pathway

Contents

1 Introduction

Biochemical engineers are concerned with biosystems and bioprocesses. Traditionally a key focus of their activities has been the mathematical modeling in both territories. Most current models for bioreactors and biosystems are expressed as systems of nonlinear differential equations. Despite of the many benefits of such models, as well as their simplicity, they give us a rather simplified picture of the reality because of the lack of any structured and segregated details. In both applications the approach is based on the assumption of well-mixed system and biomass (unstructured or structured) as well as concentrations reflecting the intracellular states are considered as a continuum. At the same time, a major challenge is to apply more detailed approaches. Life is segregated into structural and functionally discrete entities – individual cells in heterogeneous populations of unicellular organisms [1] and the large number of molecules within the cells, which are interconnected in networks additionally characterized by molecular mobility. As a matter of fact, there is an increasing demand for modeling that captures more of the relevant complexity than the aforementioned systems of ordinary differential equations can achieve.

This contribution is aimed to illustrate the framework of individual-based modeling, which can capture the contingent nature of local interactions between individual molecules or cells. The modeling strategy will be applied to two examples with noticeable differences in scale. The first example addresses the issue of modeling the dynamics of *Escherichia coli* populations in the three-dimensional turbulent field of a stirred tank bioreactor based on a structured-segregated approach. The approach permits one to account for the heterogeneity present in real reactors in both the abiotic and biotic phases and thus deals with problems of nonideal mixing.

The second example focuses on at the discrete event based stochastic simulation applied to the four-dimensional-spatial temporal dynamics of signal transduction processes in individual cells. The modeling strategy comprises the random walk of individual molecules (diffusion) and the stochastic characteristics of the interaction of the partners of interest.

It must be emphasized that the application of this individual-based modeling framework is in no way restricted to the examples chosen for this contribution. There are many more applications with particular relevance to medical applications of systems biology, such as endocytosis, random walk of molecules in cancer tumors, lipoprotein kinetics in blood plasma, network of interacting neurons, and other interesting biological problems.

2 Modeling the Dynamics of *E. coli* Populations in the Three-Dimensional Turbulent Field of a Stirred Tank Bioreactor

The physiological state of unicellular organisms in the highly dynamic environment of a bioreactor is the result of a complex interplay between the extracellular environment and the cellular machinery. The functionality of a biosystem for the purpose of bioproduction processes is therefore determined by the cooperative action of extracellular stimuli, the intracellular makeup and the dynamic response of the biological phase. The design of bioreactors in which living cells function as factories as well as the prediction of suitable operating conditions is further complicated because of the dynamic variations of the extracellular environment. Consequently, a quantitative description of these phenomena should rest upon the two interlinked aspects of structured bioprocess modeling: The first aspect concerns the complex interaction of the functional units within each cell, including the mathematical formulation of intracellular reaction rates and key regulatory responses of these networks to environmental changes. The second aspect involves the structure of the abiotic phases of the bioreactor for analysis of the impact of spatial and temporal variation in the intensities of mixing and mass transfer, leading to concentration gradients of various substrates and products. A modeling framework suited for capturing local variations in both intra- and extra-cellular concentrations should therefore rest on a link between metabolic network modeling or computational cell dynamics (CCD [2]) and computational fluid dynamics (CFD).

Previous attempts addressing this issue have been almost exclusively based on the Euler-Euler approach in which gas, liquid, and biophase are considered as a continuum [3–7]. It is only recently that more details have been included in modeling the dispersed phases. Most important are the class of models based on the Lagrangian-Euler approach, in which the liquid phase is treated as a continuum (Euler) and the dispersed phase is tracked with the aid of the Lagrangian representation. Important examples in the field of chemical engineering include the study of

gas–liquid flow in bubble column reactors [8, 9] and gas–liquid–solid three phase flow [10, 11].

In the case of bioreactors, however, modeling of the biophase most often follows the traditional approach, in which the microorganisms are lumped into a nonsegregated–non structured continuum. However, microorganisms are cellular in nature, and the continuum description is not rigorously correct. In fact, the continuum approach leads to a loss of realism if the individual history of the cells becomes the focus of attention, e.g., when considering cumulative starvation effects in cells during fed-batch fermentations or stability of plasmid containing microorganisms for production of recombinant proteins. In some instances, these problems can be tackled by combining the Euler approach for the fluid phases with population balance equations (PBEs) [12]. An inherent limitation of the PBE approach, however, is that the incorporation of a detailed intracellular reaction network leads to a computationally intractable model already for ideally mixed systems because a high-dimensional distribution function must be computed [13]. The complementary approach of combining CFD-simulations with population balances based on metabolically unstructured models of single cells faces similar practical limitations [12]. To overcome this problem, it is sometimes possible to resort to hybrid approaches combining multizonal models with CFD calculations, as reported by Bauer and Eigenberger [14, 15] for the case of gas–liquid bubble columns. Recently, Bezzo et al. [12] applied this strategy to xanthan gum production in stirred tanks. The authors combined an Eulerian description of the fluid flow with a multizonal model in which the reactor was divided into a limited number of spatial regions. These were assumed to be well-mixed and homogeneous and to be capable of material exchange with adjacent zones. Within each of these zones, a mass-structured population balance was formulated that was then combined with an unstructured kinetic model of substrate consumption and xanthan production.

Alternatively, one can account for the intracellular structure and dynamics of the cells while neglecting the spatial concentration gradients in the liquid phase. This corresponds to a Lagrange-type formulation of the equations for the corpuscular phase in an ideally mixed system [13, 16–18]. This description has the obvious limitation that bioprocesses are performed in real reactors where spatial variations in concentration usually cannot be neglected.

Detailed mathematical models capturing the variation in both the extracellular environment and the metabolism of the segregated biophase promise to aid significantly in describing the behavior of cell populations in bioreactors. This requires a combination of both approaches outlined above. For the first time this interaction between the intracellular state of the individual cells of the population and the turbulent flow field in the bioreactor has been tackled by Lapin et al. [19]. The chosen Euler-Lagrangian representation of the cell-ensemble approach permitted analysis of the lifeline of individual cells in space and time, as illustrated for the synchronization of autonomous glycolytic oscillations in yeast cells at the population level. With stirring conditions providing an environmental condition close to ideal mixing it was possible to predict the experimentally observed synchronization

of the individually autonomous oscillations at the population level. Simulation with reduced speed of agitation resulted in significant gradients of the extracellular attractor responsible for the synchronization (acetaldehyde). This leads to a dramatic loss of synchrony and eventually to almost complete desynchronization.

The work presented here picks up and summarizes the results of a second example, dealing with a problem of greater practical relevance, which, however, is also more complicated. The example, comprehensively worked out and described by Lapin et al. [20], concerns the population of the bacterium *E. coli*, which contains a phosphotransferase system (PTS) for the uptake of sugar.

The impact of extracellular gradients on biomass yield, byproduct formation and stress response of *E. coli* has been investigated in a couple of papers by the research group of Enfors [3, 21–25]. The results from measurement of glucose in a large scale bioreactor [3] during fed-batch cultures indicated that profound gradients exist which in turn give the cells an oscillatory pattern. In the case of simpler uptake systems, such as hexose transporters in yeast, a reasonable modeling approach can be based upon coupling unstructured kinetics fort he biophase with various models for mixing within the abiotic phase [3, 6, 26]. In the case of *E. coli*, however, the uptake system leads to a situation in which the local uptake rate of glucose not only depends upon the locally different concentration of glucose in the tank but also upon the intracellular state which in turn may depend on the individual history of the cell.

The agent based models presented in [19, 20], which incorporate intracellular reactions, is based on an Euler-Lagrange simulation. Here, modeling of the extracellular environment is still based on the continuous Euler approach, whereas the behavior of the biophase is characterized by a discrete cell-ensemble approach (Lagrange). This allows each single-computational-"cell," which still represents a large collective of real cells, to be endowed with its individual intracellular structure and state. In the modeling and simulation approach, mass transfer and reactions are assumed not to affect the turbulent flow field.

The three-dimensional turbulent flow for single- and two-phase systems has been simulated with the commercial software package PHOENICS employing a modified Chen-Kim k-ε turbulence model [5, 6, 27]. The population of microorganisms is distributed over the reactor, which is subdivided into finite volumes.

The random movement caused by the turbulent dispersion is superimposed on the convective flow represented by the velocity field $\vec{V}$. In the Lagrangian approach, the position of a notional particle is governed by the stochastic differential equation [28]:

$$\vec{x}(t+\Delta t) = \vec{x}(t) + \left(\vec{V} + \nabla D_T\right)dt + \left(2D_T\Delta t\right)^{1/2}\vec{\xi}, \tag{1}$$

where $\vec{x}(t+\Delta t)$ is the random position after a time step of Δt, ξ_i signifies a Gaussian random number with a mean value zero and covariance $\left\langle \xi_i \xi_j \right\rangle = \delta_{ij}$, and $D_\mathrm{T} = f(k,\varepsilon)$ stands for the local eddy diffusivity calculated from the CFD simulation as a function of the turbulent kinetic energy k and the turbulent energy

dissipation rate ε. Slip can be neglected because the fluid velocity can be shown to surpass the slip velocity of microbial cells by several orders of magnitude. Momentum transfer between the particles and the fluid phase also does not require explicit consideration because the suspension can be treated as a quasi-single phase for particles smaller then the mesh spacing, as shown previously [9].

Thus, it can be assumed that the position of the notional particle predicted from (1) represents the behavior of the microbial cell along its trajectory in the turbulent flow field.

Prediction of the intracellular state of a single cell along the trajectory is performed by incorporating the system of intracellular balance equations into the model

$$\frac{\mathrm{d}\vec{c}_{\mathrm{in},m}}{\mathrm{d}t} = \mathbf{A}_m \vec{r}_m\left(\vec{c}_{\mathrm{in},m}(t), \vec{c}_{\mathrm{ex}}(\vec{x},t)\right). \tag{2}$$

Here, $\vec{c}_{\mathrm{in},m}$ denotes the vector of intracellular metabolite concentrations in the individual cell m and $\vec{c}_{\mathrm{ex}}$ is the concentration vector of extracellular compounds at the position of this cell $(\vec{x})$. $\mathbf{A}_m$ stands for the stoichiometric matrix of the metabolic network in cell m, and the term $\vec{r}_m$ represents its vector of intracellular reactions rates, which, in general, are nonlinear functions of $\vec{c}_{\mathrm{in},m}$ and $\vec{c}_{\mathrm{ex}}$.

Inclusion of $\vec{c}_{\mathrm{ex}}$ in the intracellular balance equations is required for the description of transport processes across the cell membrane, i.e., substrate uptake and product excretion. These occur in the cellular reaction rates $\vec{r}_m$ and are also considered as a source term, $\vec{S}$(sink/source), in the Euler simulation of the extracellular state. Assuming that $T$of a total of R intracellular reaction rates represent transport rates across the cell membrane, then the system of extracellular state can be written as

$$\frac{\partial \vec{c}_{\mathrm{ex}}(\vec{x})}{\partial t} + (\vec{v}\,\nabla)\vec{c}_{\mathrm{ex}}(\vec{x}) = \mathrm{div}(D_{\mathrm{T}}\mathrm{grad}\,\vec{c}_{\mathrm{ex}}(\vec{x})) + \vec{S}(\vec{x}) \tag{3}$$

with

$$\vec{S}(\vec{x}) = \frac{\varphi V_{\mathrm{R}}}{N_{\mathrm{C}}} \sum_{m=1}^{N_C} \mathbf{A}_{T,m}\vec{r}_{T,m}(\vec{c}_{in}(t), \vec{c}_{\mathrm{ex}}(\vec{x},t))\delta(\vec{x}-\vec{x}_m) \tag{4}$$

where $\vec{S}(\vec{x})$ equals the vector of net transport rates of all metabolites across the cellular membrane of all cells present at position $\vec{x}$. This is obtained by multiplying every addend of the sum over all N_{C} cells by Dirac's delta function, which signifies whether cell m (with position $\vec{x}_m$) is present at $\vec{x}$ or not. The subvector $\vec{r}_{T,m}$ of $\vec{r}_m$ contains only those T reaction rates accounting for transport across the cell membrane. $\mathbf{A}_{T,m}$ represents the associated submatrix of $\mathbf{A}_m$ consisting of columns that correspond to the transport reactions included in $\vec{r}_{T,m}$. The term preceding the sum reflects the influence of the total cell volume on the intensity of the exchange: φ

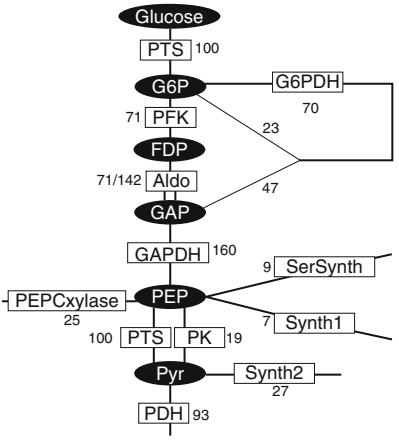


Fig. 1 Reduced metabolic network model for the sugar uptake system, glycolysis and pentose phosphate shunt. Reduction of the original model [29] is based on the hierarchy of flux control coefficients. The numbers alongside the enzymes depict the metabolic fluxes related to glucose uptake rate 100

denotes the ratio of the whole cellular volume $N_C V_m$ and the reactor volume V_R, where V_m corresponds to the volume of a (simulated) single cell. Equation (3) accounts for convection and turbulent diffusion in the liquid phase and describes the coupling between the extracellular environment and the intracellular metabolism of the single cells.

Next we set-up the metabolically structured model, i.e., the system of intracellular balance equations introduced with (2) is substantiated. The selected biological example involves the sugar uptake system of the bacterium *E. coli*, a group translocation (PTS), where transport is associated with the phosphorylation of glucose to glucose-6-phosphate (G6P). The phosphate is donated by phosphoenolpyruvate (PEP), which is converted to pyruvate (PYR) (Fig. 1). Thus actual sugar uptake rate not only depends upon the local extracellular glucose concentration but also upon the concentrations of intracellular metabolites, which in turn are governed by the dynamics of the carbon and energy metabolism, essentially glycolysis and pentose phosphate shunt. The initial point of the model development is the dynamic model of Chassagnole et al. [29] that deals with the metabolic network of the central metabolism of *E. coli* wild-type strain W3110. It comprises modules for the PTS and the Emden–Meyerhoff–Parnas pathway providing PEP as well as PYR. The model also considers the pentosephosphate pathway which is linked to the glycolysis via fructose 6-phosphate (F6P) and glyceraldehyde 3-phosphate (GAP). The original model has been developed based on the measurement of metabolites in a continuous culture that has been perturbed by a glucose pulse. The model comprises 25 state variables and 30 kinetic rate expressions have been assigned. This model, of course, is inappropriate for the problem in hand as the computational effort (q.v. (1)) cannot be handled. It is therefore inevitable to perform a systematic model reduction to decrease the number of state variables.

A well proven concept for model reduction in metabolic engineering is based on the time hierarchy of the metabolism. The kernel of this method is a modal analysis, which considers the eigenvalues and eigenvectors of the Jacobian associated with the dynamic model [30]. The application of this time scale separation resulted in

assumptions of quasi steady state conditions for 11 eigenvectors possessing the highest values of $\mathrm{Re}(\lambda_i)$. The result of this reduction, which shows reasonable agreement between the dynamic response of the original and reduced model, yielded, however, a differential-algebraic system. Because the algebraic equations did not allow an explicit analytical solution it is necessary to resort to an advanced and efficient solver for differential-algebraic systems. Thus, in context with the modeling concept developed in this chapter the decreased number of state variables would be compensated by the increased effort for the numerical solution.

As a promising alternative to the modal analysis we employed a sensitivity analysis based on the flux control coefficients. These coefficients relate the fractional change of the steady state fluxes to the infinitesimal changes in the total enzyme concentrations [30]. From the hierarchy of these flux control coefficients-predicted from the original model – the reactions with the highest values in relation to the flux control coefficient of the glucose uptake were selected. The resulting network is depicted in Fig. 1. Because of low flux control coefficients, the reactions for the phosphoglucoisomerase, the triose phosphate isomerase, the phosphoglycerate kinase, the phosphoglycomutase and the enolase could be neglected. By the same reasoning the glucose-6-phosphate dehydrogenase was selected as the rate determining enzyme for the pentose phosphate shunt. It is important to emphasize that the entire set of kinetic parameters identified from the measured intracellular metabolites [29] are the same as in the original model. The remaining set of balance equations, the kinetic rate expressions and the kinetic parameters are listed in original paper [20].

In context with the modeling task of linking the spatial variations of extracellular glucose with the dynamics of the individual cells it is important to emphasize that the system of balance equations for the intracellular state has been reduced to the feasible number of 5.

The simulations have been performed for a stirred bioreactor with 900 L operating volume, equipped with three six-bladed Rushton impellers (tank diameter 0.83 m, height 1.76 m, impeller diameter 0.33 m). The speed of agitation is 400 rpm. The number of control volumes is given by $65 \times 65 \times 128$.

The results from the simulation with 150,000 *E.coli* cells are displayed in Fig. 2. The combined model predicts distinct gradients in the extracellular glucose concentration (Fig. 2a). Glucose concentrations are highest at the top of the tank where the feed of concentrated glucose solution (concentration 600 kg m^{-3}) is introduced. In the center of the bottom of the tank the glucose concentrations are close to zero, indicating strong limitations of sugar supply in this particular part of the tank. At first glance one should worry about serious starvation effects in case *E. coli* cells frequently enter this region. The pronounced gradients in the tank reflect the poor axial pumping and thus mixing capabilities of Rushton turbines.

In Fig. 2b the distribution of the ratio of the concentrations of the two intracellular key metabolites PEP and PYR are shown. The simple kinetic expression for the uptake of the sugar [20] indicate that this ratio makes a pivotal impact on the corresponding reaction rate. As evident from Fig. 2b the behavior of the ratio of PEP to PYR is reversed to the gradient of glucose – high values are observed at the

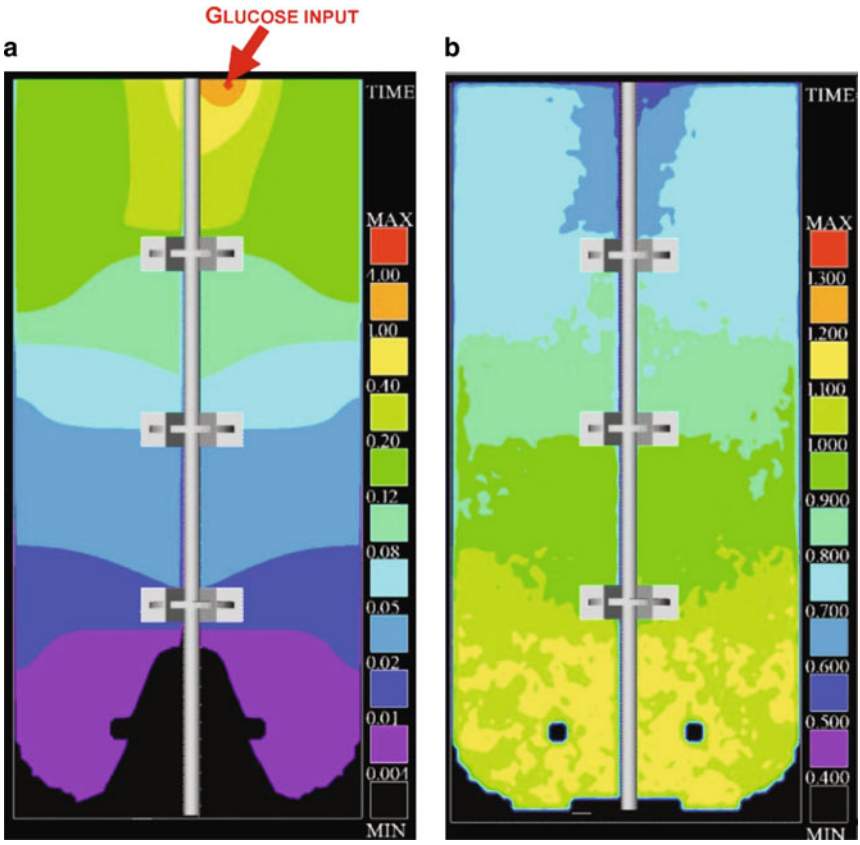


Fig. 2 Concentration fields during a fed-batch process in a 900 L bioreactor equipped with three Rushton turbines. (**a**) Extracellular glucose, (**b**) distribution of the ratio of the intracellular concentration of phosphoenolpyruvate (PEP) and pyruvate (PYR)

top, low values at the bottom of the tank. The predicted results are in agreement with expectations. Due to the high glucose concentrations at the top of the tank, the glucose uptake rate is also high. Because the transport is associated with the phosphorylation, PEP immediately decreases and is converted to PEP. Once the cells enter the regions with lower glucose concentrations the delayed flux through glycolysis leads to a refill of the pool of PEP and readjustment of the concentration of PEP.

For further interpretation of these simulation results we refer to the experimental observations of Hewitt et al. [31–33]. These authors have been concerned about studying the influence of scale of cultivation and, as such, different intensity of mixing, on the viability of *E. coli* populations during fed-batch fermentations with constant feeding rates. For these purposes, multiparameter flow cytometry has been used. With the introduction of specific fluorescent dyes, valuable quantitative information on cell physiology and particularly viability could be obtained. The analysis revealed that a temporally varying environment with respect to

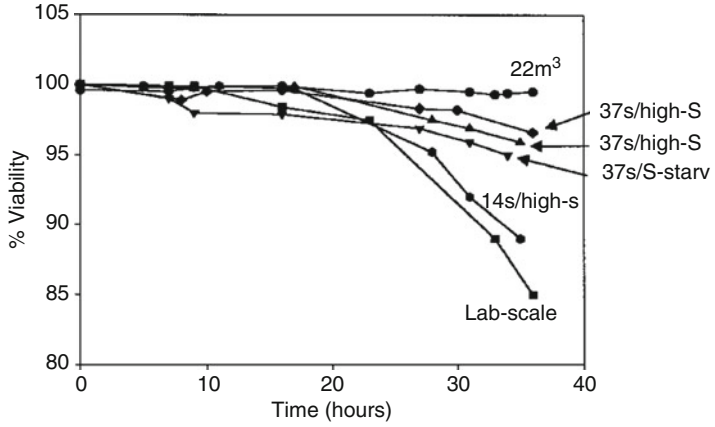


Fig. 3 *E. coli* viability (measured with flow cytometry) in a lab-scale, a 22 m^3 scale and in a scale-down reactor with 14–37 s passages through a glucose starvation or glucose excess zone (reprinted from [32] during fed-batch operation with constant feeding rate)

glucose concentration has a profound effect on the viability of the cells. The comparison between a 22 m^3 and a 5 L scale demonstrates distinct differences in the cell viability (Fig. 3). Obviously the small-scale, well-mixed fermentation gave the lowest cell viability. The relatively poor mixed conditions in the large scale fermenter were found to lead to high cell viability. The reasoning regarding the positive influence of fluctuations in the microenvironment on cell viability were further supported with corresponding observations in an experimental set-up for scaled-down simulations. In these scaled-down experiments a small well-mixed stirred tank reactor (STR) was coupled to a plug flow reactor (PFR) via recycling flow. Highly concentrated glucose solution was fed into the STR and into the inlet of the plug flow. This set-up allows the successful simulation of poor mixing in large scale reactors. The same authors started to discuss possible reasons for these results in terms of environmental stress associated with the ever-increasing glucose limitation in the well-mixed case under conditions of constant feeding. It was furthermore argued that with the large scale and, depending on the residence time in the PFR also with the scale-down simulations, cells periodically encounter regions of relatively higher glucose concentrations.

It appears obvious to interpret our own simulation results further in the context of the aforementioned experimental observations. It appears beneficial to use the additional information about the distribution of intracellular state variables (PEP, PYR in Fig. 2b) to sustain the hypotheses regarding the impact of interlinked regulatory stress response pathways which was put forward by Hewitt et al. [33]. It is known that under conditions of glucose depletion, rapid increase of the intracellular "second messengers" cAMP and ppGpp is observed. cAMP is involved in the regulation processes related to the phenomenon of catabolite repression. The signal is built up by the enzyme adenylate cyclase which in turn is activated by one of the phosphorylated proteins in the PTS at conditions of sugar depletion and high ratios

of PEP/PYR. At natural conditions the signal is responsible for regulation phenomena leading to the expression of alternative sugar sources. The dynamic response of the PTS under conditions of substrate depletion is also linked to the chemotaxis allowing the bacteria to swim towards more favorable conditions e.g., higher glucose concentrations. In summary this dynamic response leading to the build up of the alarmone cAMP may be characterized by an offensive response of *E.coli* because constructive activities are mobilized to overcome the limitations in the carbon and energy limitations. In contrast to this offensive response the second alarmone ppGpp leads to regulation phenomena summarized with the term "stringent response," which in a more defensive way reduces the energy and carbon demand by downregulating anabolic activities of the cell [34]. This phenomenon eventually leads to loss of viability via stress response (starvation). One of the key players in this stress response is the sigma factor (σ^S). Interestingly, the two alarmones cAMP and ppGpp compete as transcription factors in the expression of the sigma factor [35].

Based on these molecular details regarding the link between metabolism and its regulation and the results of the simulations in Fig. 6 it is tempting to speculate on the following scenario within the large tank. Those cells traveling through the region of high sugar concentrations at the top of the tank respond with a corresponding high sugar uptake rate and a drop in the phosphorylation potential (PEP/PYR). At the same time as the cells are moving towards the bottom of the tank where they are exposed to extremely low sugar concentration, the ratio of PEP/PYR is very high. The dynamic response of the PTS system under these conditions should result in a fast increase of cAMP. Even under conditions of lasting sugar limitations the peak of cAMP would counteract the stringent response signal of ppGpp and thus would prevent the stress response. In the case of the small well-mixed reactor the sugar concentration would always remain at a very low level, the ppGpp could prevail and eventually pronounced stress response is initiated which also impacts cell viability by inducing the programmed death of the bacteria [36].

3 Stochastic Simulations of Four-Dimensional-Spatial Temporal Dynamics of Signal Transduction Processes

Traveling along lifelines of individual cells and populations in bioreactors is only one but nevertheless an important example of the application of the agent based or individual modeling in bioprocess and biosystems engineering. The issue of detailed quantitative modeling of spatiotemporal effects at the scale of individual cells or multicellular systems comprises a manifold of important problems. Important examples comprise chemotaxis and quorum sensing in bacteria, signal transduction pathways, endocytosis, phagocytosis as well as movement of drug molecules in complex tissues such as solid tumors, to name a few.

Many cellular signaling events occur in small subcellular volumes and involve low-abundance molecule species. This context introduces a major difference and additional complication compared to the bioreactor modeling illustrated above.

Reactions involving a low number of molecule species occur in a probabilistic manner. Thus, in addition to the random walk simulating the diffusive motion of the individual molecules the stochasticity of the reaction has to be taken into account. As such, the assembly approach presented in the aforementioned example is extended to a coupled reaction–diffusion process in which the individual agents change their characteristic properties through interactions.

Spatial aspects of cellular signaling have already been the subject of various experimental and theoretical investigations [37, 38]. Modeling the interaction of diffusion and reaction is most often based on the continuum approach, thus investigating the corresponding system of partial differential equations. The pioneering work of Kholodenko [38, 39] addressed for the first time the issue of spatially heterogeneous and time varying cellular signal transduction cascades. They developed computational models of the mitogenic signaling network to analyze the complex structure of the spatial distribution of the activated compound ERK. At a distance larger than several microns from the plasma membrane the phosphorylation signal is attenuated practically to basal levels, provided that the phosphatase activity in the cytosol is sufficiently high. These and similar investigations by Howe [40] are, however, restricted to a macroscopic, deterministic continuum approach, which neglects the random walk of individual molecules and the stochastic characteristics of the interaction of the partners of interest. These stochastic properties can only be modeled using Monte-Carlos simulations.

The most important approach for stochastic simulation of coupled chemical reactions trace back to the famous work of Gillespie [41], solving master equations which describe the evolution of the so-called grand probability for the number of molecules of the different species. This equation, however, is only valid for spatially homogenous mixtures, in other words, ideally well-mixed systems. There are a couple of attempts to couple diffusion problems to this master equation [42, 43]. A critical assessment of these approaches, in which the system is divided into small subvolumes, demonstrates two important drawbacks. First, the sub-volume is assumed to be well-mixed and the diffusion processes are restricted to the boundaries of the grids. Second, the molecules are represented as point particles. Therefore, it is not possible to reproduce crowded conditions because volume exclusion from both reactive and nonreactive crowded molecules cannot be represented explicitly.

On the other hand microscopic models including all molecules of a cell as well as all interactions – leading to a complete molecular dynamics (MD) simulation – are computationally expensive and therefore unable to cover the dimensions of a complete cell [44]. Only mesoscopic models are until now able to span all necessary ranges to model signaling processes with sufficient detail. The underlying Smoluchowski model for diffusion according to Brownian motion and reactions in a reaction–diffusion system are implemented by different groups. Particles [45, 46] or so called agents [47] perform a random walk representing the diffusion process. Reactions will take place if the distance between reacting species falls below a predefined reaction-distance. While MCell [46] restricts reactions to fixed positions, all these models have the drawback that there is no physical property defining the fixed model-depending reaction-distance.

The method presented in this chapter aims at studying the interconnected effects of signal transduction networks by a new multi-scale approach. The probability of collision and reaction between the interesting species is modeled by the solution of the Fokker–Planck equation providing the probability that two molecules will collide and react in the next interval Δt. The trajectories for the random walk of protein molecules are modeled by stochastic differential equations (Lagrangian approach). This allows incorporating the effects of macromolecular architecture of the cells and thus to investigate the hindered diffusion due to crowding caused by the cytoskeleton. The criterion for change on the model scale is a threshold value for the distance between the reaction partners. The new modeling approach will be exemplified for the RAS-MAPK pathway and a steroid hormone pathway.

3.1 The Multi-Scale Modeling Approach

3.1.1 Random Walk of the Molecules

The random walk simulations are based on the numerical solution of the stochastic differential equation for single molecule tracking:

$$\vec{x}(t + \mathrm{d}t) = \vec{x}(t) + (2D\mathrm{d}t)^{1/2}\vec{\xi} \tag{5}$$

with the three-dimensional position vector $\vec{x}$, the diffusion coefficient D and the Gaussian random number ξ_i with mean zero and covariance $\langle \xi_i \xi_j \rangle = \delta_{ij}$

3.1.2 Fokker–Planck Equation

Our interest is to study the coupling of this random walk of molecules with the probability of collisions with potential reaction partners and eventually the degree of the diffusion-limited reaction itself. For this purpose the random movements of both molecules (e.g., phosphorylated protein and phosphatase or steroid and steroid-receptor) are simulated according to the aforementioned strategy. If paths of interacting proteins come within a distance ε of each other, a change in the scale of modeling is performed by switching to the analytical solution of the Fokker–Planck equation for the probability density function (p.d.f.) P of diffusive movement given by:

$$\frac{\partial P}{\partial t} = (D_A + D_B)\Delta P \tag{6}$$

We only need to keep track of the magnitude of the difference of position [48] and hence one molecule can be considered to be in a fixed position whereas the second one diffuses through the three-dimensional space. The boundary conditions for (6) are given by the following considerations: The particles can not overlap: if

the particles reach the distance $r = R_A + R_B$ ether the reaction will occur (leading to a flux of the particles into each other) or the particles will be reflected [49]. This leads to the partially reflecting boundary condition

$$r \to \infty: \quad P = 0 \qquad r = R_A + R_B: \quad (D_A + D_B)\frac{\partial P}{\partial r} = k'P \quad \text{(reaction)}$$

with R_A, R_B radii of the two molecules, r radial coordinate and k' surface reaction constant. The relation with the macroscopic reaction constant for well-mixed systems can be derived easily:

$$k' = \frac{k_{\text{macro}}}{4\pi(R_A + R_B)D - k_{\text{macro}}}$$

(Dimensionless parameters and variables simplify the calculations: $k_{\text{macro}} = k_{\text{macro}}(R_A + R_B)^2/D$; $r = r/(R_A + R_B)$; $t = tD/(R_A + R_B)^2$). Inert obstacles can be modeled with $k' = 0$ leading to pure reflection.

For the initial condition that the particles are separated by $\vec{r}_0$ at $t = 0$, the initial p.d.f. is $P = \delta(\vec{r} - \vec{r}_0)$, and the analytical solution of (6) reads:

$$\begin{aligned} P &= \int_0^\infty e^{-\lambda^2 t} \sum_{n=0}^{\infty} (n + 1/2) P_n(\cos\ \theta) R_{n\lambda}(r) R_{n\lambda}(r_0) \mathrm{d}\lambda \\ R_{n\lambda} &= \frac{L_{yn\lambda} j_n(\lambda r) - L_{jn\lambda} y_n(\lambda r)}{\sqrt{L_{jn\lambda}^2 + L_{yn\lambda}^2}} \\ L_{yn\lambda} &= \lambda y'_n(\lambda) - k y_n(\lambda) \\ L_{jn\lambda} &= \lambda j'_n(\lambda) - k j_n(\lambda) \end{aligned} \tag{7}$$

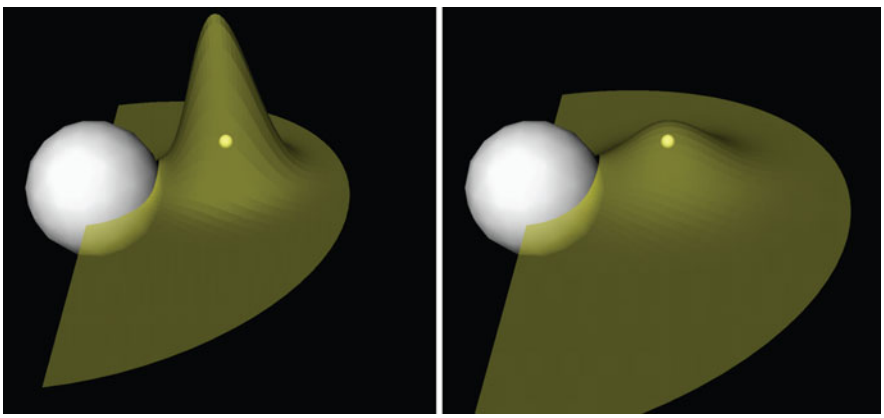


Fig. 4 Time development of the probability density function for the possible position of a particle after time Δt and $2\Delta t$ (analytical solution of the Fokker–Planck equation). The initial position is marked with the small sphere, the large sphere represents the distance R_A+R_B. The function broadens due to diffusion and the overall probability is reduced by the reaction probability

where P_n is the Legendre polynomial and j_n and y_n are spherical Bessel functions. Equation (7) describes the time evolution of the p.d.f. of the relative position of the molecules. Figure 4 shows a graphical presentation of this solution for two snapshots in time.

Eventually the reaction probability has to be estimated. The calculation is based on the simple fact that the sum of the probabilities that a molecule is still present within the p.d.f. P and the probability that a molecule disappeared through reaction must be 1.

Position probability + reaction probability = 1, or

$$R_P = 1 - \int\limits_{r \geq R_A + R_B} P(\vec{r}, t)\mathrm{d}\vec{r}$$
$$R_P = \frac{k'}{k' + 1 + x_0} \times \left[\mathrm{erfc}\left(\frac{x_0}{\sqrt{\tau 4}}\right) - \exp(x_0 + \tau)\mathrm{erfc}\left(\frac{x_0 + 2\tau}{\sqrt{\tau 4}}\right)\right] \tag{8}$$

again with dimensionless variables: $x_0 = (r_0 - 1)(1 + k')$, $\tau = (1 + k')^2 \Delta t$

A uniform random number $\zeta (0 \leq \zeta \leq 1)$ is then used as an indicator of a successful reaction, such that:

(1) *if* $\zeta \leq R_P \longrightarrow$ *reaction* (2) *if* $\zeta \geq R_P \longrightarrow$ *no reaction*

In the case of (2), the molecule continues moving on according to $P(\vec{r}, t)$. As soon as it reaches the critical distance ε between the two molecules, it is further tracking according to (5).

3.1.3 Effect of Macromolecular Crowding

Molecular crowding and the cytoskeleton have to be taken into account to get a more realistic consideration of the cellular architecture [38]. The cytoskeleton is simulated by inserting randomly distributed cylinders into the cell. Thus the free volume is reduced by 30%. To model the interaction with the cytoskeleton cylinders we rejected steps that would end inside a cylinder. The effective diffusion coefficient is about proportional to the free volume in this case [50]. The effective concentration of reaction partners is increased because the free volume is reduced, and thus molecular crowding increases the reaction speed.

3.1.4 Advantages of the Multi-Scale Modeling Approach

By switching to the solution of the Fokker–Planck equation we ensure that the effect of reactions as well as obstacles is properly considered in the diffusion path of the diffusing molecules. Fixed reaction distances like in Smoldyn [45] simply cut

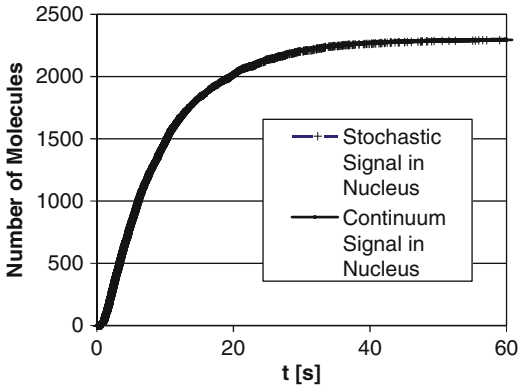


Fig. 5 Comparison of stochastic and continuum signal

off the Gaussian probability density function for the diffusion steps, ignoring the unknown positions of the particles between t and $t + \Delta t$. The probability density function according to the Fokker–Planck equation is written to a lookup table so there is only a minor increase in computation time. The comparison under spatial homogeneous continuum conditions shows excellent agreement (see Fig. 5).

3.2 Applications

3.2.1 Impact of Spatial Separation of Kinases and Phosphatases on the Output Signal of the MAPK Cascade

The MAPK cascades contain three interconnected cycles of MAPK, a MAPK kinase (MAPKK) and a MAPKK kinase (MAPKKK). In the most well characterized MAPK/Erk cascade, the system consists of ERK, MEK and Raf. Upon stimulation and Ras activation, the cytosolic Raf is recruited to the cell membrane, where it binds to and phosphorylates MEK. The phosphorylated MEK drifts into the cell interior where it phosphorylates ERK. ERK then travels through the cytoplasm into the nucleus, where it triggers the expression of certain genes. During its random walk ERK can be attacked by various phosphatases and after successful reaction would lose its activation state [39].

The example is, first of all, used to examine the reliability of the approach and the various numerical methods required for the simulations. For this purpose we use a simulation of the temporal-spatial distribution of 4,000 ERK molecules starting at time zero from the cell membrane. According to the fluctuation theory the variance is proportional to $1/\sqrt{N}$, where N is the number of particles. With $N = 4{,}000$ one should expect an agreement between continuum and discrete simulation within 1.6%. For the continuum approach we solve the partial differential equation for

diffusion and first order reaction (the number of phosphatase molecules is not changed in the process).

$$\begin{aligned}
&\frac{\partial c(r,t)}{\partial t} = D\frac{1}{r^2}\frac{\partial}{\partial r}r^2\frac{\partial c(r,t)}{\partial r} - Kc(r,t)\\
&\qquad B.C.: \quad c(r,t)|_{r=R_0} = 0 \qquad\qquad (R_0 \text{ is the radius of the nucleus})\\
&\qquad D\frac{\partial c(r,t)}{\partial r}\Big|_{r=R} = \phi_0\delta(t) \qquad\qquad (\text{initial flux from the cell membrane})\\
&\qquad I.C.: \quad c(r,t=0) = 0
\end{aligned} \tag{9}$$

Analytical solution:

$$c(r,t) = \sum_{n=1}^{\infty}\frac{2\phi_0}{rD}\frac{e^{-\left(\lambda_n{}^2/(1-R_0)^2+Da\right)t}\sin\left(\lambda_n(r-R_0)/(1-R_0)\right)}{(1-R_0)(\lambda_n\sin(\lambda_n)-R_0\cos(\lambda_n))}$$

$$\text{with Damkohler Number} \quad Da = \frac{KR_{\text{cell}}^2}{D}$$

$$\text{and eigenvalues} \quad \lambda_n\cos\lambda_n = (1-R_0)\sin\lambda_n$$

From this solution the flux into the nucleus and the concentration inside the nucleus can be derived:

$$\phi_{\text{nucleus}}(t) = D\frac{\partial c(r,t)}{\partial r}\Big|_{r=R_0} \qquad c_{\text{nucleus}}(t) = 4\pi R_0^2\int_{t'=0}^{t}\phi_{\text{nucleus}}(t')dt'$$

The result of the comparison (Fig. 5) is more than satisfactory as one cannot see any difference in the dynamic response between the continuum and stochastic simulation of the problem.

As expected, the differences are much more pronounced if the number of molecules is decreased.

Figure 6 shows simulation result for a scenario with only 100 ERK molecules. There are differences between the discrete and continuum simulations and, additionally, the dynamic response in case of the stochastic simulation differs from run to run.

For adequate representation of the simulation results, a visualization framework has been implemented [51]. It allows an interactive exploration of the data from the simulation. Two visualization styles have been developed (see Fig. 7): a microscopic and a more schematic representation. A virtual microscope creates images, which look like the results from a confocal fluorescence microscope. With this representation the simulation can easily be compared with microscope images from experiments. The schematic visualization is more abstract and visualizes all

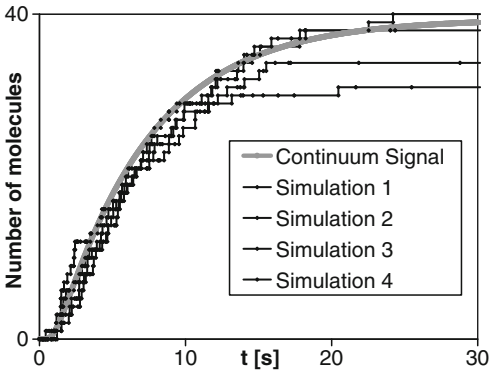


Fig. 6 Comparison of stochastic and continuum signal for 100 ERK molecules (the remaining ERK molecules are dephosphorylated and do not reach the nucleus)

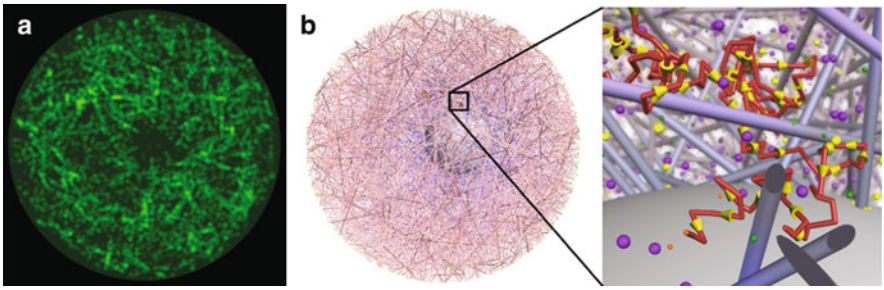


Fig. 7 Visualization of the virtual cell. (a) fluorescence microscope. (b) detailed image showing the cytoskeleton, nucleus, crowding molecules (*violet, green and yellow*) and signaling molecules (*orange*), including the path of a signaling molecule (*red*, *arrow cones* in *yellow* show the direction of the path)

simulated components. Crowding and signaling molecules are represented by spheres with the respective radius, and the filaments of the cytoskeleton are represented by cylinders. Molecule paths can be highlighted to follow a molecule of interest.

3.2.2 A Steroid Hormone Pathway: A Case Study of a Bimolecular Reaction

The steroid pathway with its ligand activated steroid hormone receptor (androgen, androgen-receptor) differs from the membrane anchored receptor–ligand interactions and mobilization of phosphorylated proteins illustrated for the MAP kinase in that the steroid hormone is able to penetrate the cell membrane and then binds to the receptor (see Fig. 8). The receptor–ligand complex then travels to and is imported into the nucleus [52].

From the point of view of kinetics, the important difference is that the problem is characterized by a real bimolecular reaction A+B=C. It is the nonlinearity of the

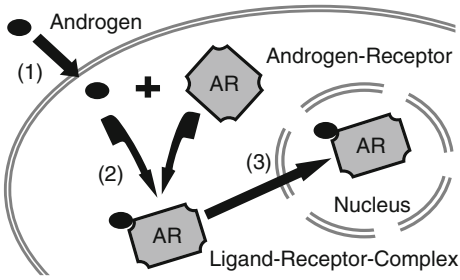


Fig. 8 Steroid hormone pathway for androgen: androgen-hormone diffuses into cell (1), binds to androgen-receptor (2) and the complex is imported into the nucleus (3)

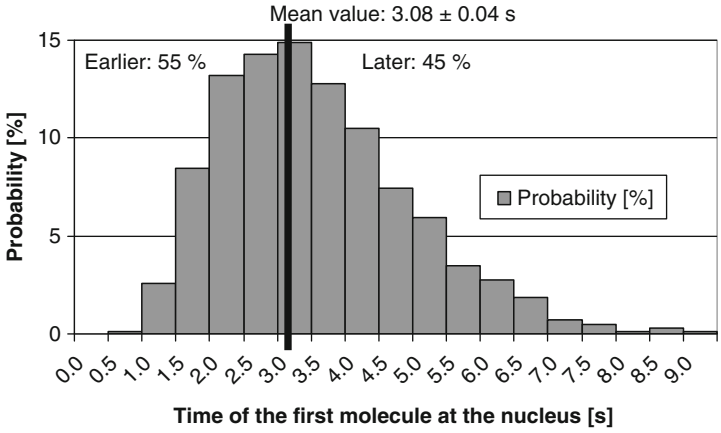


Fig. 9 Probability distribution for the first passage time of the first androgen–receptor-complex reaching the nucleus

reaction which should lead to much more pronounced differences between discrete and continuum simulations.

It is not possible to derive an analytical solution for the continuum in case of a bimolecular reaction so we solved (10) with the bimolecular reaction term $K c_1(r,t)c_2(r,t)$ numerically. In the stochastic simulation the first passage time of the androgen–receptor-complex was recorded in 1,008 trials with 500 androgen molecules initially located at the cell membrane and 500 receptors randomly distributed in the cell. The statistical analysis reveals that 55% of the complexes reach the nucleus earlier than the sample mean of 3.08 s with a standard error of 0.04 s (see Fig. 9). In the continuum approach the first particle arrives 4% later (3.20 s); already 57% of the trials arrived within that time. This shows, that local density fluctuations have an effect on the average reaction of bimolecular reactions (and this example has not even been optimized for a maximum nonlinearity effect; it was designed to reflect natural conditions).

4 Conclusion

The basic idea central to the agent based or individual modeling approach presented in this chapter are entities as objects (cells or molecules) traveling along paths which are computed from stochastic differential equations. In the case of the bioreactor model the random walk of individual cells is calculated from CFD-simulations, in which turbulent dispersion is superimposed to the three-dimensional convective movement in the turbulent flow field. The example presented deals with the impact of sugar transport into bacterial cells containing a PTS. The method allows the population behavior to be described at the outcome of the interaction between the intracellular state of its individual cells and the turbulent flow field in the bioreactor. The chosen Euler-Lagrange representation of the cell-ensemble approach permits analysis of the lifelines of individual cells in space and time. The approach presented integrates CFD with a segregated description of a cell population in a stirred tank thereby accounting for a detailed intracellular structure of the single cells. The biological example tackled with this approach is of great practical relevance. The simulation results point to serious differences in the dynamics of the intracellular states at different scale of operation with significant impact on the viability of the cells.

The second application of this approach is the random walk of molecules in individual cells. The two examples chosen comprise the MAP kinase and a hormonal stimulation. In contrast to the first example, in which the communication between the objects is moderated by the extracellular environment, thus neglecting a direct interaction, the signal transduction examples involve a direct molecular interaction via biochemical reactions. To overcome the problem of a step size dependent influence upon the collision frequency of molecules the reaction probability is estimated from the theoretical solution of the Fokker–Planck equation. This switch in the model approach is a special kind of multi-scale modeling characterized by the transition from the physical into the probabilistic space.

Acknowledgements The authors acknowledge support of the Deutsche Forschungsgemeinschaft (DFG) within the collaborative research center “Sonderforschungsbereich 412” and the Ministry of Science, Research and Arts Baden-Württemberg within the Center Systems Biology University Stuttgart.

References

1. Frederickson AG, McGee RD, Tsuchiya HM (1970) Adv Appl Microbiol 23:419
2. Kitano H (2006) Nat Rev Mol Cell Biol 7:163
3. Larsson G et al (1996) Bioproc Eng 14:281
4. Noorman H et al (1993) Proceedings of the 3rd international conference on bioreactor and bioprocess fluid dynamics, Cambridge 241:258
5. Reuss M, Schmalzriedt S, Jenne M (2000) Bioreaction engineering: modeling and control. Springer, Berlin 207:246

6. Schmalzriedt S et al (2003) Adv Biochem Eng 18:68
7. Träghardh C (1988) The second international conference on bioreactor fluid dynamics, King, Applied Sciences Publishing, Cambridge 117:134
8. Lapin A et al (2002) Chem Eng Sci 57:1419
9. Sokolochin A et al (1997) Chem Eng Sci 52:61
10. Bourloutski E, Sommerfeld M (2002) Proceedings of the 10th workshop on two-phase flow predictions, Merseburg 113:223
11. Van Sint Annaland M, Deen NG, Kuipers JAM (2005) Chem Eng Sci 60:6188
12. Bezzo F, Macchietto S, Pantelides CC (2003) AIChE J 49:2133
13. Henson MA, Müller D, Reuss M (2002) Biochem J 368:433
14. Bauer M, Eigenberger G (1999) Chem Eng Sci 54:5109
15. Bauer M, Eigenberger G (2001) Chem Eng Sci 56:1067
16. Ataii MM, Shuler ML (1985) Biotechnol Bioeng 27:1026
17. Domach MM, Shuler ML (1984) Biotechnol Bioeng 26:877
18. Kim BG, Shuler ML (1990) Biotechnol Bioeng 36:581
19. Lapin A, Müller D, Reuss M (2004) Indus Eng Chem Res 43:4647
20. Lapin A, Schmid J, Reuss M (2006) Chem Eng Sci 61:4783–4797
21. Bylund F, Collet E, Enfors S-O, Larsson G (1998) Bioprocess Eng 18:171
22. Bylund F et al (1999) Bioprocess Eng 20:377–389
23. Schweder T et al (1999) Biotechnol Bioeng 65:151
24. Teich A et al (1999) Biotechnol Prog 15:123
25. Xu B et al (1999) Appl Microbiol Biotechnol 51:564
26. Bajpai RK, Reuss M (1982) Can J Chem Eng 60:384
27. Jenne M, Reuss M (1999) Chem Eng Sci 54:3921
28. Fox RO (2003) Computational models for turbulent reacting flows. University Press, Cambridge
29. Chassagnole C et al (2002) Biotechnol Bioeng 79:53
30. Heinrich R, Schuster S (1996) The regulation of cellular systems. Chapman & Hall, New York
31. Hewitt CJ, Nebe-Von Caron G (2001) Cytometry 44:179
32. Hewitt CJ et al (1999) Biotechnol Bioeng 63:705
33. Hewitt CJ et al (2000) Biotechnol Bioeng 70:381
34. Lengeler JW, Drews G, Schlegel HG (1999) Biology of the prokaryotes. Thieme, Stuttgart
35. Lewis K (2000) Microbiol Mol Biol Rev 64:503
36. Loewen PC et al (1998) Can J Microbiol 44:707
37. Takahashi K, Arjunan SNV, Tomita M (2005) FEBS Lett 579:1783
38. Kholodenko BN (2006) Nat Rev 7:165
39. Kholodenko BN (2003) J Exp Biol 206:2073
40. Howe CL (2005) Theor Biol Med Model 2:43
41. Gillespie DT (1976) J Comp Phys 22:165
42. Stundzia AB, Lumsden CJ (1996) J Comput Phys 127:196
43. Ander M, Beltrao P, Di Ventura B, Ferkinghoff-Borg J, Foglierini M, Kaplan A, Lemerle C, Tomás-Oliveira I, Serrano L (2004) Sys Biol 1:129
44. Tolle DP, Le Novère N (2006) Curr Bioinform 1:3
45. Andrews SS, Bray D (2004) Phys Biol 1:137
46. Stiles JR, Bartol TM (2000) In: de Schutter E (ed) Computational neuroscience: realistic modeling for experimentalists. CRC, Boca Raton, FL, p 87
47. Pogson M, Holcombe M, Smallwood R, Qwarnstrom E (2006) BioSystems 85:37
48. Batada NN, Shepp LA, Siegmund DO (2004) Proc Natl Acad Sci USA 101:6445
49. Rice SA (1985) In: Bamford CH, Tripper CFH, Compton RG (eds) Diffusion-limited reactions. Amsterdam, Elsevier
50. Trinh S, Arce P, Locke BR (2000) Transp Porous Media 38:214
51. Falk M, Klann M, Reuss M, Ertl T (2009) Proceedings of IEEE pacific visualization symposium 2009, Beijing 169:176
52. Tyagi RK, Lavrovsky Y, Ahn SC, Song CS, Chatterjee B, Roy AK (2000) Mol Endocrinol 14:1162

Adv Biochem Engin/Biotechnol (2010) 121: 45–70
DOI: 10.1007/10_2009_56

Published online: 26 February 2010

Impact of Profiling Technologies in the Understanding of Recombinant Protein Production

Chandran Vijayendran and Erwin Flaschel

Abstract Since expression profiling methods have been available in a high throughput fashion, the implication of these technologies in the field of biotechnology has increased dramatically. Microarray technology is one such unique and efficient methodology for simultaneous exploration of expression levels of numerous genes. Likewise, two-dimensional gel electrophoresis or multidimensional liquid chromatography coupled with mass spectrometry are extensively utilised for studying expression levels of numerous proteins. In the field of biotechnology these highly parallel analytical methods have paved the way to study and understand various biological phenomena depending on expression patterns. The next phenomenological level is represented by the metabolome and the (metabolic) fluxome. However, this chapter reviews gene and protein profiling and their impact on understanding recombinant protein production. We focus on the computational methods utilised for the analyses of data obtained from these profiling technologies as well as prominent results focusing on recombinant protein expression with *Escherichia coli*. Owing to the knowledge accumulated with respect to cellular signals triggered during recombinant protein production, this field is on the way to design strategies for developing improved processes. Both gene and protein profiling have exhibited a handful of functional categories to concentrate on in order to identify target genes and proteins, respectively, involved in the signalling network with major impact on recombinant protein production.

C. Vijayendran
International NRW Graduate School in Bioinformatics and Genome Research, CeBiTec, Bielefeld University, 33594 Bielefeld, Germany
e-mail: cvijayen@cebitec.uni-bielefeld.de

E. Flaschel (✉)
Faculty of Technology, Chair of Fermentation Engineering, Bielefeld University, 33594 Bielefeld, Germany
e-mail: efl@fermtech.techfak.uni-bielefeld.de

Keywords Bacteria, *Escherichia coli*, Proteomics, Recombinant protein production, Transcriptomics

Contents

1 Introduction

In the field of biology and biotechnology, vital questions are answered by focusing on information obtained from all functional levels of the cell represented by the transcriptome, the proteome and the metabolome while analysing genes, proteins and metabolites, respectively. The metabolome, however, often gives only valuable information if rates of individual reaction steps can be assessed leading to the fluxome. For studies of protein expression, gene activities (transcriptomics) and protein abundances (proteomics) are most frequently analysed. Therefore, this review focuses on these two major profiling technologies.

Gene expression analysis as a high throughput methodology is commonly based on microarrays [1] – apart from serial analysis of gene expression (SAGE) [2]. The set of all mRNA species from a given cell population in a certain environment is referred to as the transcriptome. Microarray experiments commonly quantify the transcriptome of a sample in comparison with a reference sample, thus providing a functional measure of relative gene expression. In general, mRNA from a given culture is used to generate a labelled sample. Together with a differently labelled reference mRNA sample both are hybridised (complementary base paring) together onto a large number of DNA sequences immobilised on a solid support in an

ordered array. The arrays are read by detecting the spot intensities with respect to both labels, normally represented by fluorescent stains. As a result, microarray experiments accelerate the process of identifying genes of many entire genomes which are up- or down-regulated under certain cultivation conditions.

The field of proteomics has emerged in parallel as a rapidly advancing technology-driven field [3–5]. Mass spectrometry of protein-derived peptides in combination with separation tools such as two-dimensional gel electrophoresis or multidimensional liquid chromatography are commonly utilised as major methods for protein profiling analyses [6].

One of the major contributions of global expression profiling in the field of biotechnology consists in the identification of vital genes responsible for recombinant protein production [7–9]. This review will briefly focus on the computational methods available to handle the data generated for these profiling analyses prior to discussing the findings and the impact of these profiling technologies in the field of recombinant protein production. More specifically, it will focus on the application of microarray technology for mRNA profiling and two-dimensional gel electrophoresis for protein profiling for a better understanding of recombinant protein production in the field of biotechnology.

2 Experimental Procedures

In this chapter a short introduction into the main wet-lab experimental methodologies of profiling will be given.

2.1 Microarray-Based Gene Expression Profiling

The most commonly used microarray platforms can be classified into two groups, complementary DNA- (cDNA-) and oligonucleotide-microarrays. Arrays of cDNA are based on the first microarray technique which has been developed [1]. These are produced by the robotic application of cDNA or genomic clones onto a glass surface in an arrayed format. Oligonucleotide arrays consist of short 20–25mers synthesised in situ, either by photolithography (high-density oligonucleotide arrays, achieved through step by step attachment of a single nucleotide using a series of photolithographic masks) or by ink-jet technology. cDNA arrays are hybridised with equal amounts of two samples each labelled with a different fluorophor (CY3 and CY5). The use of different fluorescent dyes allows RNAs from two different populations to be mixed and hybridised to the same array. Usually high-density oligonucleotide microarrays yield a single fluorescence intensity measurement for each spot on the array, whereas cDNA microarrays produce a ratio between two signals being measured concurrently for each spot. Being scanned for two different wavelengths, the intensity of the same spot in both the channels is compared.

This results in a measurement of the ratio of RNA levels for each element represented on the microarray. Before obtaining microarray measurements for a particular set of experiments or platforms, the reported measurements have to be normalised because of inequalities in the construction or detection of the two fluorescent libraries, background noise, and systematic biases in the measurements [10]. Apart from normalisation, data transformation [11] and data filtering [12] are major pre-processing steps involved before analysis or classification. Detailed approaches for microarray data analysis are reviewed elsewhere [13, 14]. To date, computational analyses of RNA-expression data sets have centred on two approaches, namely supervised and unsupervised ones. Algorithms used for class prediction [15] (to identify a list of candidates which assist to accurately predict new samples to their appropriate class from the expression pattern) within the data set fall into the category of supervised, whereas algorithms handling data sets for class discovery [15] (to identify different classes among multiple and biologically homogeneous samples) are assigned to the category of unsupervised classification methods.

2.2 Protein Expression Profiling Based on Two-Dimensional Gel Electrophoresis

Two-dimensional electrophoresis is a powerful and widely used method for the analysis of protein pools extracted from biological samples. This technique separates proteins according to two separate properties in two steps called the first (isoelectric focusing) and the second dimension (SDS-PAGE). In the first dimension the proteins are separated according to their isoelectric points and in the second dimension according to their molecular mass. Following the visualisation of the two-dimensional gels, the protein spots are excised from the gels to undergo tryptic digestion. The digested protein samples are analysed by mass spectrometry for annotating the peptide mass fingerprints against relevant databases. Gels which are scanned and digitised are subjected to image smoothing, spot detection, spot quantification, image alignment, spot matching, spot annotation, and molecular mass as well as pI calculation by various image analysis software. For each protein spot, the annotated information along with the peak area and normalised quantity values are obtained for further statistical analyses.

3 Evaluation of Profiling Experiments

This chapter introduces methods for treating complex data from mRNA and protein profiling in order to extract meaningful results. These methods may be divided into two distinct classes according to the classification to be accomplished by means of supervised or unsupervised strategies.

3.1 Unsupervised Classification of Expression Profiles

Unsupervised learning is one of the most popular statistical techniques to be applied in microarray data analysis, and it involves the aggregation of diverse collection of data into clusters based on different features in a data set. In unsupervised classification, algorithms are used to cluster sets of entities (genes/proteins) to reveal similarities of expression across multiple samples. For example, the exploration of new physiological behaviour during recombinant protein production may be obtained by clustering genes based on coexpression. By compiling and clustering various strain/condition samples at once based on their expression profiles, samples can be grouped into classes on the basis of the similarity in their expression profiles. Likewise, by clustering entities among all samples, other elements could be identified sharing a similar pattern of expression (coexpression).

These unsupervised clustering approaches allow for the identification of formation and structure of complex data sets without requiring any prior hypotheses and assumptions to be made. Working methodologies of clustering algorithms and practical strategies for using clustering algorithms are discussed elsewhere in detail [16]. Most frequently, used clustering methods in unsupervised learning are hierarchical clustering, *k*-means clustering, self-organising maps and principal components analysis (PCA). These clustering methods use dissimilarity measures to create groups of features with similar patterns. Commonly used dissimilarity measures are the Euclidean distance, the Pearson correlation coefficient, and the rank correlation coefficient.

Euclidean distance is based on visual space in which each entity is treated as a point (vector) in multidimensional space. Each axis is a separate biological sample and the coordinate on each axis is the amount of, e.g. RNA expression in that sample. This allows entities to be clustered according to their degree of similarity in terms of expression ratios [17]. Thus, entities that exhibit similar expression levels are found in close proximity. An expression vector is then calculated to describe the position of each entity. Using these expression vectors, distance metrics are calculated between each pair of entities in the dataset to provide the similarity among them. Distance metrics are then used for various clustering purposes [17].

In the case of the Pearson correlation coefficient, the measurement between two entities is treated as a vector of measurements assuming that the measurements of, e.g. mRNA expression among the samples are normally distributed [18].

3.1.1 Hierarchical Clustering

Hierarchical clustering algorithms classify the clusters of entities (genes/proteins) with similar patterns of expression. The two most closely related entities in terms of smallest distance metrics based on expression measurements are grouped into a single cluster. These calculations proceed iteratively until all the entities have been clustered. The visual representation of the result is in the form of a dendrogram

which is represented as a tree and entities as leaves of that tree. Numerous expression patterns within the data set can be observed quickly without any prior hypotheses and assumptions and hence this is a popular technique for analysing global profiling expression data [17].

3.1.2 *k*-Means Clustering

In *k*-means clustering, entities are organised in a number of *k* clusters, in which *k* is given in advance by the user. The elements are initially assigned randomly to a cluster, and then the mean vector for all the elements is computed in each cluster and subsequently they are each reassigned to a cluster whose centre is closest to the element. This process is repeated *n* times from a different initial cluster till the optimum cluster is found many times. The main parameters that control *k*-means clustering are the number of clusters specified by the user and the number of trials to be performed [17].

3.1.3 Self-Organising Maps

Self-organising maps create clusters of entities in multidimensional space, in which neighbouring clusters are similar. Each entity is defined by coordinates according to expression levels. Thus, each sample is considered to have a separate dimension in space. With the number of clusters provided by the user, the map is arranged in an arbitrary shape with the centres of each cluster, normally called centroids. The optimised self-organising map is obtained by iterating the process until no further centroid movements are detected [17, 19, 20].

3.1.4 Principal Components Analysis

PCA is a technique for multivariate data analysis which reduces the dimensionality and complexity of the dataset without losing the ability to calculate accurate distance metrics. It transforms the expression data into a more manageable form in which a number of clusters may be more easily discriminated. In PCA, the data vectors are written as a linear sum over principal components. Principal components are a set of vectors in multidimensional space that decreasingly capture the variation seen in the dataset. The number of principal components is equal to the number of dimensions of the data vectors. Thus, they represent the axes of a transformed coordinate system, in which the first principal component determines the basic orientation. The principal components are found by calculating the eigenvectors of the covariance matrix of the data. The corresponding eigenvalues determine how much of the variance present in the data is explained by each principal component. The first principal component captures more variation than the second, and so on [17, 21].

3.2 Supervised Classification of Expression Profiles

Supervised learning enables class prediction or discrimination of an independent test dataset with the help of a training dataset. The training dataset is obtained by identifying informative entities which exhibit differential expression in a defined group. New sample datasets can be assigned or related to a group based on a cut-off value obtained from a training dataset. Supervised learning incorporates the knowledge of class label information from trained datasets. In applying this information to independent datasets, distinctions of interest can be performed. For instance, a subset of differentially expressed genes could be selected which can significantly distinguish between two conditions and build a model based on these candidate genes to classify that particular condition from other varied conditions. The most commonly used supervised methods that accurately predict or distinguish a pattern from the given dataset are the nearest neighbour approach, artificial neural networks (ANN) [22] and support vector machines (SVM) [23]. In the correlation-based classification methods, numerous statistical procedures are applied to the expression profile datasets in order to obtain a discriminatory weight for each element. Based on this, the entities are ranked and the supervised classifiers are constructed based on the top ranked entities. ANN- and SVM-based algorithms are capable of learning complex patterns from expression datasets. Once the dataset is trained, the parameters of ANN or SVM can provide vital information about the relative importance of each element in the learning of the classes [22, 23].

4 Profiling Analysis During Recombinant Protein Production

During the last two decades or so, recombinant proteins have become an important and diverse class of biotechnological products. Bacterial expression systems are still the most attractive means for their production owing to low cost, high productivity and versatility. Bacteria grow rapidly and at high-density on inexpensive substrates. Other advantages are found in their often well-characterised genetics and the availability of a large number of cloning vectors and mutant strains. Among the bacterial host strains, the Gram-negative bacterium *Escherichia coli* is still the most commonly used organism for recombinant protein production because it is comparatively well known, its cultivation is established in numerous laboratories and its genome has been sequenced several times. However, the lack of most of the post-translational protein modifications, differences in codon usage and the production of inactive proteins due to the formation of inclusion bodies offer clear limitations and significant challenges for the use of such expression systems. Problems faced during the production of recombinant proteins by means of bacterial systems are due to (1) the presence of multicopy expression vectors, (2) the over-expression of desired genes leading to metabolic stress, accumulation of endotoxins, and (3) inclusion body formation or protein misfolding in general.

From earlier studies it had been reported that several factors such as growth rate, growth phase, growth method and medium ingredients determine recombinant protein production. Different genetic and metabolic strategies have been developed to obtain superior host and vector systems for recombinant protein production. Among these widely explored strategies the coexpression or knock-out of certain key genes are found, genes which play a significant role in recovering from physiological stress. Therefore, heat-shock proteins (HSP) with protease- and/or chaperone activity are considered as prime candidates for differential expression to achieve improved yields of recombinant proteins. Recent advancements in gene expression and proteome analyses have offered a systematic approach for the identification of various genes and proteins which are either over- or under-expressed in many physiologically stressful conditions [7, 24–28].

Earlier experiments in the field of microbiology and biotechnology have exploited the advancement in genomics, allowing many functional genomic-based studies towards understanding global metabolic changes caused by differences in genotype and differing cultivation conditions [29–34]. Microarray technology has been established as the efficient methodology for simultaneous exploration of the expression levels of numerous genes. Microarray experiments commonly quantify only the transcriptome of a sample of cells as compared to a reference sample, but they do not give direct access to the concentration of the individual main players of the cellular machinery – the proteins. This is the reason why the proteome has to be known as well.

With respect to proteomics, profiling approaches have been employed to examine the protein expression level changes under various conditions [5, 33, 35]. Proteomics profiling has largely relied on data from conventional two-dimensional polyacrylamide gel electrophoresis which is still the major method for global proteome analysis [6], though it presents quite a number of inconveniences. Nevertheless, this method of analysis can help to identify targets for improving the expression of recombinant proteins, such as genes to delete from the host cell chromosome [36] or to co-express with a product [7, 9, 37]. These global profiling technologies have been utilised extensively for the identification of protein targets which play a vital role in recombinant protein production. Upon the induction of recombinant proteins, the products of many genes may show up transcribed from

Table 1 Regulation of genes during recombinant protein production according to the functional categories to which they belong

Functional category of genes	Regulation	References
Heat-shock and stringent response	Up-regulated	[9, 21, 33–37, 44–46]
Phage-related	Up-regulated	[7, 21, 34, 47]
Elongation factors	Down-regulated	[9, 50]
Ribosomal protein	Down-regulated	[7, 9, 21, 36, 45, 51, 52]
Amino acid biosynthesis and tRNA-related	Down-regulated	[8, 9, 21, 23, 36]
Transposon-related	Up-regulated	[21, 34]
Nucleotide biosynthesis	Down-regulated	[7, 8]
Tricarboxylic acid (TCA) cycle	Up-regulated	[23, 33, 56]

genes like those of (1) heat-shock and stringent response, (2) phage-related ones, (3) elongation factors, (4) ribosomal proteins, (5) amino acid biosynthesis- and tRNA-related ones, (6) transposon-related ones, (7) nucleotide biosynthesis and (8) tricarboxylic acid (TCA) cycle. Many of these genes were significantly differentially regulated. Table 1 gathers this information from the literature under review and shows which functional category is influenced in its regulation pattern and if this influence is of positive (up-regulated, over-expressed) or negative (down-regulated, under-expressed) nature. These classes will be discussed in more detail in the following paragraphs.

4.1 Heat-Shock and Stringent Response Genes

Various profiling studies have identified genes that were sensitive to overproduction of the recombinant protein. Among these, HSP were significantly over-represented in most of the studies [9, 26, 38–42]. HSP are a group of proteins whose expression is induced when a cell undergoes various types of environmental stresses like heat, cold and oxygen deprivation. A group of HSP are chaperones that monitor and assist in protein folding in the cell [43], while other HSP are proteases, which degrade unfolded or damaged proteins [44].

A special case is represented by plasmid-based systems, the expression of which is induced by a shift to higher temperature, normally because of runaway replication [45, 46]. Such thermo-inducible systems have been developed in order to circumvent the addition of large quantities of inducer chemicals (like IPTG) as required for the induction of classical promoters during large scale protein production [46–48]. To examine the differences in the metabolism of the cell during recombinant protein production at elevated temperatures, a recent study was carried out to analyse the transcriptional changes during thermal induction only and for the case in which the inducer IPTG was used in addition [41]. As a result, it was shown that about 200 genes were differentially expressed. Among these, heat-shock-related genes and amino acid-tRNA genes were significantly regulated [41]. Since amino acid-tRNA genes play a vital role in protein translation, differential regulation of these genes can be attributed to the improved recombinant protein production during thermal induction. Another gene expression profiling study during recombinant protein production has shown differential expression of heat-shock genes and decreased expression levels of genes related to transcription, translation and energy metabolism [26]. In an earlier study it was observed that heat-shock genes were over-expressed during shake-flask experiments due to recombinant protein overproduction [42]. Similarly, in a typical proteome profiling experiment during the overproduction of a serine-rich protein, it was observed that HSP were over-expressed [9]. Likewise, it was shown in various individual studies that HSP were over-expressed in parallel with the recombinant proteins [39, 49, 50]. By combining translational and transcriptional levels during the recombinant protein production, it was shown that there was a drastic increase in σ^{32} transcription

factor-dependent genes [38]. The sigma factor 32 (σ^{32}) is required for the normal expression of heat-shock genes and for the heat-shock response [51]. Target gene over-expression may induce metabolic stress which is counteracted by reorganisation of the genetic regulatory network.

4.2 Phage-Related Genes

During proteomic profiling of an antibody fragment-producing recombinant *E. coli*, it was observed that the synthesis of the phage-shock protein A (PspA), a stress protein, strongly correlated with the synthesis of the recombinant product. By coexpressing PspA, it was shown that the yield of the recombinant product was improved significantly [7]. Higher PspA expression was also reported for the production of a recombinant human growth hormone [52]. PspA is a stress protein and acts as a negative transcriptional regulator of the *psp* operon [53]. PspA is known to be induced under a variety of conditions, including filamentous phage infection, blocked protein secretion, and the addition of proton motive force uncouplers [54]. While analysing the transcriptome profiles in recombinant protein producing *E. coli* during high cell density fed-batch cultivations, it was shown that numerous phage-related genes were differentially regulated [26]. In an earlier study, bacteriophage associated genes (*ftsH*, *recA*, *alpA*, *uvrB*) were found to be significantly up-regulated after the induction of recombinant proteins [39]. These results taken together indicate that, during recombinant protein production, various signalling networks related to phage defence mechanisms are activated, maybe due to the metabolic burden the additional protein synthesis may impose on the cell. Nevertheless, various cascade signalling processes occurring during the stress due to this metabolic burden are still not completely understood.

4.3 Elongation Factor Genes

Proteome profiles of *E. coli* in response to the overproduction of human leptin (a serine-rich protein) showed elevated levels of proteins involved in heat-shock, whereas genes of elongation factors were significantly down-regulated [9]. Previous reports have shown that during recombinant protein production the levels of proteins involved in translation and ribosomal subunit components were lowered [14, 55]. These reports are consistent with the down-regulation of elongation factors during recombinant protein production, because they are responsible for the protein translation activity of the cell [9]. These findings show that the overproduction of recombinant proteins negatively influences the capacity of cellular translation processes and that one potential reason for this is the down-regulation of elongation factors.

4.4 Ribosomal Protein Genes

While examining the gene expression levels during thermal induction of recombinant proteins, it was shown that 39 ribosome genes were significantly down-regulated [41]. Proteomic analyses in high cell density fermentations showed that the expression levels of ribosomal proteins were significantly lowered as well [7]. Previous work has confirmed that the synthesis rates of ribosomal proteins [50, 56] and the associated rRNA levels [57] were decreased during recombinant protein production. Based on the variations in proteome profiles in response to the overproduction of human leptin, it was found that the levels of the 30S ribosomal protein were lowered [9]. In a complete study of the regulation of all 55 ribosomal protein genes, it was noticed that rRNA levels were lower in the case of recombinant protein production [26] as well. Many of the ribosome genes were profoundly down-regulated during various recombinant protein production processes. This reveals a counteracting situation in which enhanced transcriptional activity of a target gene leads to lowering of the concentration of the translationary machinery of the cell.

4.5 Amino Acid Biosynthesis and tRNA-Related Genes

During the production of recombinant proteins in high cell density fed-batch cultures, it was shown that 17 of the 24 genes encoding aminoacyl-tRNA synthetases – catalysing the addition of amino acids to tRNA – were differentially regulated and, of these, 14 genes were down-regulated [26]. Likewise, 70 of the 80 tRNA genes were significantly differentially regulated, and 67 of these were down-regulated during recombinant protein production [26]. Global transcriptome analyses of recombinant protein production processes have shown differences in the levels of amino acid-tRNA genes in the case of thermal induction [41]. In this situation the expression levels of amino acid-tRNA genes were found to be elevated, which might account for improved productivity of the desired protein in this study [41]. However, the genes encoding aminoacyl-tRNA synthetases were significantly down-regulated [41]. Since tRNA play a critical role in the translation of mRNA to proteins, over-expression of tRNA genes can be attributed to be the cause for increased productivity of the desired protein in the case of thermal induction. This feature may counteract other negative effects caused by this particular induction strategy. Proteome profiling of *E. coli* in response to the overproduction of human leptin has shown that some enzymes involved in amino acid biosynthesis were present at lower concentrations [9]. More specifically, the levels of enzymes involved in the biosynthesis of the serine family of amino acids were significantly down-regulated. By coexpression of the *cysK* (cysteine synthase A) gene, the cell growth rate and leptin productivity could be improved [9]. During transcriptome profiling of recombinant *E. coli* producing human insulin-like growth factor I fusion protein in high cell density fed-batch culture, it was shown that the expression of the genes associated with the biosynthesis of nucleotides and amino acids

were significantly down-regulated [8]. A combined study of transcriptome and proteome profiling of *E. coli* during high cell density cultivation revealed that most of the amino acid biosynthesis genes were down-regulated when the cell density increased [28], a behaviour which may explain the reduced selectivity of protein production often observed in the case of high cell density cultivations.

4.6 Transposon-Related Genes

Transcriptome analyses of *E. coli* have shown that genes associated with transposons and IS-elements were significantly up-regulated due to recombinant protein induction during high cell density fed-batch fermentation [26]. Gene expression analyses were carried out in order to evaluate the stress response to over-expression of five recombinant proteins [39]. These studies revealed dramatic changes in the transcription rates of transposon-related genes after induction of recombinant protein expression. Specifically, the transcript level of IS5 transposase was significantly up-regulated [39]. It is known that the activation of insertion elements indicates an increase in transposition events which should lead to increased genetic diversity [58]. IS5 transposase is also known to be active in nutritionally deprived, resting cells [59, 60]. The strategy behind this is that *E. coli* tries to adapt to stressful environments by increasing its mutation rate. The increased transcript levels of IS5 transposase may represent an additional strategy to achieve high mutational frequencies.

4.7 Nucleotide Biosynthesis Genes

Transcriptome profiling of recombinant *E. coli* producing a human insulin-like growth factor I fusion protein during high cell density fed-batch cultivation has shown that the genes involved in the biosynthetic pathway of nucleotides were significantly down-regulated [8]. This was confirmed by a proteome analysis of *E. coli* producing a recombinant antibody fragment in high cell density fermentation, since down-regulation of ribosomal as well as nucleotide biosynthesis proteins was observed [7]. Not only nucleotide biosynthetic proteins involved in the purine and pyrimidine biosynthesis were found to be down-regulated, but also proteins associated with the synthesis of nucleoside triphosphates as well as proteins with regulatory function with respect to these proteins [7].

4.8 Tricarboxylic Acid Cycle Genes

A proteomics analysis of *E. coli* in high cell density fed-batch fermentation during over-expression of plasmid-encoded 6-phosphogluconolactonase showed that proteins participating in the TCA cycle were up-regulated [61]. Another study of

E. coli during high cell density cultivation, in which the transcriptome was analysed in addition to the proteome, showed that the expression of the genes involved in the TCA cycle were up-regulated and that the levels of TCA cycle enzymes were increased as well about twofold [28]. Further confirmation is available from analyses of the transcriptional and translational levels of various genes during recombinant protein production showing that the proteins involved in the TCA cycle were significantly up-regulated [38]. Since the TCA cycle is of primary importance for most of the general biosynthetic pathways it can be anticipated that a high TCA cycle activity would be required for a high synthesis rate of recombinant proteins.

4.9 Association Network Analysis

By combining the data (gene/protein regulation) generated from various studies mentioned in this chapter, we generated an association network for these genes involved in the recombinant protein production and analysed the network (Fig. 1). It has been demonstrated that functionally related genes are preferentially linked in coexpression networks [62]. By integrating various other information (fusion evidence, neighbourhood evidence, co-occurrence evidence, experimental evidence, textmining evidence, database evidence) along with the coexpression data information from the STRING database [63], we were able to build an association network for the genes which are differentially regulated during recombinant protein production (Fig. 1). In many cases there were meaningful relationships between network substructure, gene function and network association (Fig. 1b–e). This kind of network analysis provides the possibility to use it as an analytical tool to unravel the relationships among genes that govern the cellular functions [64], for example, in the substructure network Fig. 1b, which consists of 14 genes among which ten genes co-occur with the term "recombinant protein production" in the PubMed database (node border colour in green). Among these ten genes, six genes (*dnaK*, *dnaJ*, *grpE*, *clpB*, *htpG* and *groL*) code for proteins which are well-characterised molecular chaperones known to play a vital role during recombinant protein production [65, 66]. The substructure network genes are preferentially linked, clearly denoting their functional similarities. This shows the relevance and potential of association network analysis. Hence, the genes in these substructure networks (Fig. 1b–e) are of great importance to understanding the gene regulatory network during recombinant protein production.

5 Conclusion

Both gene and protein profiling have demonstrated their potential in the identification of important molecular genetic targets which could be modified in order to optimise the industrial-scale production of recombinant proteins. However, it has to

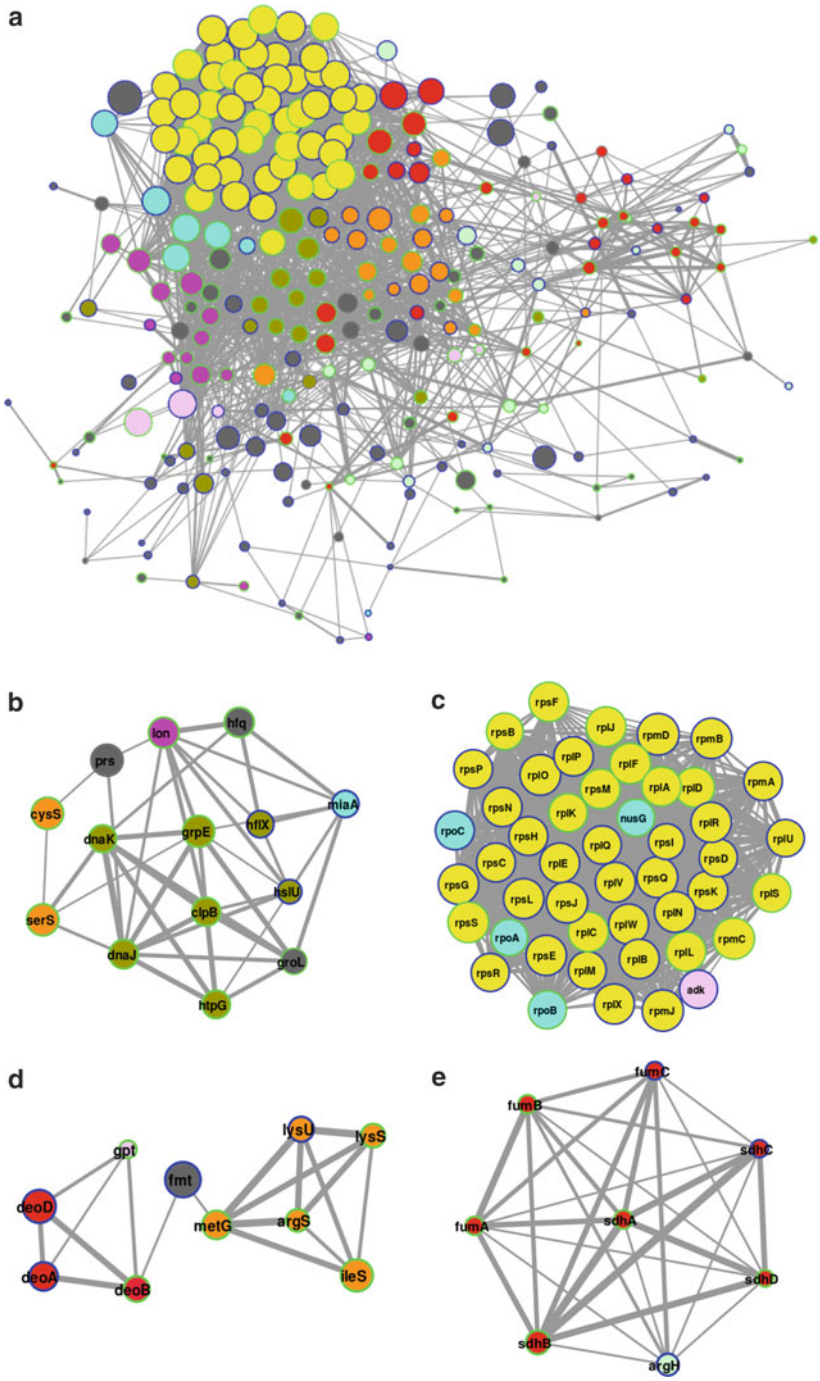


Fig. 1 (**a**) The association network of genes known to be differentially expressed during recombinant protein production. Nodes (genes) are coloured according to their functional category (amino acid related genes = *light green*; cellular processes = *dark green*; energy metabolism = *red*; nucleotide

be considered seriously that most of the experiments reporting profiling analyses for recombinant protein production have been conducted under rather restricted conditions with respect to the media, promoters and plasmids used as well as to the recombinant proteins expressed. Thus, the actual results accumulated cannot fully and in general answer the broad range of questions associated with which group of genes may be directly involved in the signalling network modulating recombinant protein production. The genes/proteins which are differentially regulated under such a condition are summarised in Table 2 for up-regulated genes and Table 3 for down-regulated ones. These data are gathered from the literature cited in this review. In order to give additional information, a column called "Literature count" has been added to both tables. This column represents the number of publications in the PubMed database which include the specific gene name and the term "recombinant protein production" – e.g. "groEL AND recombinant protein production" is searched against the PubMed database. This means that those genes for which there is a "0" in this column may be novel candidates to look at since their regulation pattern is significantly altered by expression of a recombinant protein, but there have been very limited studies of these genes. Genes and proteins belonging to various functional categories highlighted in this review (heat-shock, stringent response, phage-related, elongation factors, ribosomal proteins, amino acid biosynthesis, tRNA-related, transposon-related, nucleotide biosynthesis, as well as TCA cycle genes and proteins) and gathered in Table 1 should be targeted by various approaches to learn how to design new strategies for improving recombinant protein production. The whole network and effective relationship between the above-mentioned functional categories of genes and proteins and their effective role in recombinant protein production is still far from being completely understood. However the genes which are preferentially linked and densely connected functionally associated genes in the association network substructure are of great importance to unravel the gene regulatory network involved during recombinant protein production (Fig. 1b–e). Technologies like ChIP-on-Chip which combine chromatin immunoprecipitation (ChIP) with genomic microarray analysis (chip) can be used to investigate interactions between DNA and target proteins (in vivo) which are known to be significantly expressed during recombinant protein production. By utilising this technology, the binding sites of DNA-binding proteins can be identified in an efficient manner which may unravel the whole network and

Fig. 1 (continued) related = *pink*; regulatory related = *violet*; ribosome related = *yellow*; transcription related = *cyan*; translation related = *orange*, others = *grey*). Node size is based on the network neighbourhood connectivity. Node border colour denotes the presence (*green*) or absence (*blue*) of publications in which the corresponding gene name co-occurs with the term "recombinant protein production" in the PubMed database. The network edges represent the presence (one or many) of predicted functional association denoting (1) fusion evidence, (2) neighbourhood evidence, (3) co-occurrence evidence, (4) experimental evidence, (5) textmining evidence, (6) database evidence and (7) co-expression evidence from the STRING database [63]. Edge size represents the number of predicted functional association (thicker edge size represents a higher number of functional associations) from the STRING database [63] (see above). (**b–e**) Substructure extracted from STRING network (**a**) with MCODE algorithm [82], showing preferentially linked and densely connected functionally associated genes

Table 2 Genes which are up-regulated (over-expressed) during recombinant protein production identified by mRNA or protein profiling. The last column represents the number of publications in which the corresponding gene name co-occurs with the term "recombinant protein production" in the PubMed database

Gene	B-number	Gene name	Literature count
aceF	b0115	Dihydrolipoyllysine-residue acetyltransferase	2
acnB	b0118	Aconitate hydratase 2	0
adiA	b4117	Biodegradative arginine decarboxylase	1
Adk	b0474	Adenylate kinase	0
AhpF	b0606	Alkyl hydroperoxide reductase subunit F	0
AldH	b1300	Putative aldehyde dehydrogenase.	3
araE	b2841	Arabinose-proton symporter	0
argF	b0273	Ornithine carbamoyltransferase chain F	0
argH	b3960	Argininosuccinate lyase	0
argI	b4254	Ornithine carbamoyltransferase chain I	0
argS	b1876	Arginyl-tRNA synthetase	1
Asd	b3433	Aspartate-semialdehyde dehydrogenase	8
AsnS	b0930	Asparaginyl-tRNA synthetase	0
b0257	b0257	CP4-6 prophage; partial transposase of insertion element IS91 1A	0
b1145	b1145	Putative lambdoid prophage e14 repressor protein C2	0
b1362	b1362	Putative Rz endopeptidase from lambdoid prophage Rac	0
b1374	b1374	Putative DNA-invertase from lambdoid prophage Rac	0
b1579	b1579	Putative lambdoid prophage Qin defective integrase	0
b2442	b2442	Putative prophage CPZ-55 integrase	0
b4285	b4285	Transposase insM for insertion sequence element IS600	0
bioC	b0777	Biotin synthesis protein bioC	1
btuB	b3966	Vitamin B12 transporter btuB precursor	0
cadA	b4131	Lysine decarboxylase, inducible	3
cadB	b4132	Probable cadaverine/lysine antiporter	0
cheA	b1888	Chemotaxis protein cheA	1
citD	b0617	Citrate lyase acyl carrier protein	0
citE	b0616	Citrate lyase beta chain	3
citF	b0615	Citrate lyase alpha chain	0
clpB	b2592	Chaperone clpB	9
csgB	b1041	Minor curlin subunit precursor	1
cvpA	b2313	Colicin V production protein	0
cysS	b0526	Cysteinyl-tRNA synthetase	3
DapA	b2478	Di hydrodipicolinate synthase	2
dapE	b2472	Succinyl-diaminopimelate desuccinylase	0
deoD	b4384	Purine nucleoside phosphorylase deoD-type	0
dinF	b4044	DNA-damage-inducible protein F	0
dsbB	b1185	Disulfide bond formation protein B	4
endA	b2945	Endonuclease-1 precursor	2
exbB	b3006	Biopolymer transport exbB protein	2
FabI	b1288	Enoyl-[acyl-carrier-protein] reductase [NADH]	2
fadL	b2344	Long-chain fatty acid transport protein precursor	2
Fes	b0585	Enterochelin esterase	4
flgE	b1076	Flagellar hook protein flgE	1
fliD	b1 924	Flagellar hook-associated protein 2	0
FliY	b1920	Cystine-binding periplasmic protein precursor	0

(*continued*)

Table 2 (continued)

Gene	B-number	Gene name	Literature count
fruL	b0079	Very hypothetical fruR/shl operon leader peptide	0
fumA	b1612	Fumarate hydratase class I, aerobic	1
fumB	b4122	Fumarate hydratase class I, anaerobic	1
fumC	b1611	Fumarate hydratase class II	0
fxsA	b4140	Protein fxsA	0
GatY	b2096	Tagatose-1,6-bisphosphate aldolase gatY	1
glcB	b2976	Malate synthase G	0
GldA	b3945	Glycerol dehydrogenase	1
gltA	b0720	Citrate synthase	6
gppA	b3779	Guanosine-5′-triphosphate,3′-diphosphate pyrophosphatase	0
gpt	b0238	Xanthine phosphoribosyltransferase	12
groEL	b4143	60 kDa chaperonin	68
groES	b4142	10 kDa chaperonin	44
GrpE	b2614	Protein grpE	21
Gst	b1635	Glutathione S-transferase	276
hipA	b1507	Protein hipA	0
hipB	b1508	HTH-type transcriptional regulator hipB	0
HisJ	b2309	Histidine-binding periplasmic protein precursor	0
hscA	b2526	Chaperone protein hscA	1
hslU	b3931	ATP-dependent hsl protease ATP-binding subunit hslU	0
htgA	b0012	Very hypothetical heat shock protein htgA	0
HtpG	b0473	Chaperone protein htpG	4
htpX	b1 829	Probable protease htpX	0
htrC	b3989	Heat shock protein C	0
htrE	b0139	Outer membrane usher protein htrE precursor	0
hyaB	b0973	Hydrogenase-1 large chain	0
hycC	b2723	Formate hydrogenlyase subunit 3	0
I bpA	b3687	Small heat shock protein ibpA	8
IbpB	b3686	Small heat shock protein ibpB	5
ilvA	b3772	Threonine dehydratase biosynthetic	2
IlvI	b0077	Acetolactate synthase isozyme III large subunit	0
insA	N/A	N/A	10
intA	b2622	Prophage CP4-57 integrase	5
intB	b4271	Prophage P4 integrase	0
intC	b2349	Putative prophage CPS-53 integrase	0
intD	b0537	Prophage DLP12 integrase	0
intF	b0281	Putative prophage CP4-6 integrase	0
kgtP	b2587	Alpha-ketoglutarate permease	0
lacY	b0343	Lactose permease	5
lacZ	b0344	Beta-galactosidase	331
lamB	b4036	Maltoporin precursor	37
lit	b1139	Bacteriophage T4 late gene expression-blocking protein	3
LpcA	b0222	Phosphoheptose isomerase	0
lysA	b2838	Diaminopimelate decarboxylase	1
MdaB	b3028	Modulator of drug activity B	0
mdh	b3236	Malate dehydrogenase	7
metC	b3008	Cystathionine beta-lyase	0
modF	b0760	Putative molybdenum transport ATP-binding protein modF	0
mutM	b3635	Formamidopyrimidine-DNA glycosylase	0

(*continued*)

Table 2 (continued)

Gene	B-number	Gene name	Literature count
nagC	b0676	N-acetylglucosamine repressor	0
nfrA	b0568	Bacteriophage N4 adsorption protein A precursor	0
nfrB	b0569	Bacteriophage N4 adsorption protein B	1
nhaR	b0020	Transcriptional activator protein nhaR	1
nlp	b3188	Sugar fermentation stimulation protein B	0
nmpC	b0553	Outer membrane porin protein nmpC precursor	0
nohA	b1548	Prophage Qin DNA packaging protein NU1 homolog	0
nrd F	b2676	Ribonucleoside-diphosphate reductase 2 beta subunit	0
NusG	b3982	Transcription antitermination protein nusG	1
ogrK	b2082	Prophage P2 OGR protein	0
OmpF	b0929	Outer membrane protein F precursor	13
OmpR	b3405	Transcriptional regulatory protein ompR	7
ompX	b0814	Outer membrane protein X precursor	2
OppA	b1243	Periplasmic oligopeptide-binding protein precursor	2
pckA	b3403	Phosphoenolpyruvate carboxykinase [ATP]	4
pdxJ	b2564	Pyridoxal phosphate biosynthetic protein pdxJ	1
PepD	b0237	Aminoacyl-histidine dipeptidase	0
pheM	b1715	Phenylalanyl-tRNA synthetase operon leader peptide	0
PhoB	b0399	Phosphate regulon transcriptional regulatory protein phoB	0
plsB	b4041	Glycerol-3-phosphate acyltransferase	1
ppdD	b0108	Prepilin peptidase dependent protein D precursor	0
proP	b4111	Proline/betaine transporter	0
PrsA	b1207	Ribose-phosphate pyrophosphokinase	3
PstS	b3728	Phosphate-binding periplasmic protein precursor	5
ptrB	b1845	Protease 2	0
PurB	b1131	Adenylosuccinate lyase	0
pyrB	b4245	Aspartate carbamoyltransferase catalytic chain	1
pyrC	b1062	Dihydroorotase	0
pyrI	b4244	Aspartate carbamoyltransferase regulatory chain	1
pyrl	b4246	PyrBI operon leader peptide	0
racC	b1351	Protein racC	0
RbsB	b3751	D-ribose-binding periplasmic protein precursor	0
rhsE	b1456	Protein rhsE	0
Rrf	b0172	Ribosome recycling factor	3
sdaC	b2796	Serine transporter	0
sdhA	b0723	Succinate dehydrogenase flavoprotein subunit	2
sdhB	b0724	Succinate dehydrogenase iron-sulfur protein	2
sdhC	b0721	Succinate dehydrogenase cytochrome b556 subunit	0
sdhD	b0722	Succinate dehydrogenase hydrophobic membrane anchor protein	1
sgcR	b4300	Putative sgc region transcriptional regulator	0
sieB	b1353	Superinfection exclusion protein B	0
SodB	b1656	Superoxide dismutase [Fe]	2
speC	b2965	Ornithine decarboxylase, constitutive	2
ssb	b4059	Single-stranded DNA-binding protein	9
SspA	b3229	Stringent starvation protein A	4
t150	b3558	Putative transposase insK for insertion sequence element IS1 50	0
tdcE	b3114	Keto-acid formate acetyltransferase	0
TesB	b0452	Acyl-CoA thioesterase II	4
tig	b0436	Trigger factor	4

(*continued*)

Table 2 (continued)

Gene	B-number	Gene name	Literature count
TpiA	b3919	Triosephosphate isomerase	0
Tpx	b1324	Thiol peroxidase	4
tra5	N/A	N/A	0
tra5	N/A	N/A	0
treC	b4239	Trehalose-6-phosphate hydrolase	0
trkG	b1363	Trk system potassium uptake protein trkG	0
trs5	N/A	N/A	0
tsx	b0411	Nucleoside-specific channel-forming protein tsx precursor	0
ugpB	b3453	Glycerol-3-phosphate-binding periplasmic protein precursor	0
uraA	b2497	Uracil permease	0
Usg	b2319	USG-1 protein	0
uxaC	b3092	Uronate isomerase	0
yagU	b0287	Inner membrane protein yagU	0
yah B	b0316	Putative HTH-type transcriptional regulator yahB	0
yaiU	b0374	flagellar protein	0
YajQ	b0426	UPF0234 protein yajQ	0
ybbD	b0500	Hypothetical protein ybbD	0
bcS	b0555	Probable lysozyme from lambdoid prophage DLP12	2
ybcT	b0556	Putative Rz endopeptidase from lambdoid prophage DLP12	0
ybcU	b0557	Lipoprotein bor homolog from lambdoid prophage DLP12 precursor	0
YbdQ	b0607	Universal stress protein G	0
ybeD	b0631	UPF0250 protein ybeD	0
YcaC	b0897	Protein ycaC	0
YchF	b1203	GTP-dependent nucleic acid-binding protein engD	0
yciS	b1279	Inner membrane protein yciS	0
YdjA	b1765	Protein ydjA	0
yejG	b2181	Hypothetical protein yejG	0
YgfZ	b2898	Unknown protein from 2D-page	0
yhgE	b3402	Hypothetical protein yhgE	0
YhgI	b3414	Protein gntY	0
yi21	N/A	N/A	0
yi22	N/A	N/A	0
yi22	N/A	N/A	0
yi41	b4278	Transposase insG for insertion sequence element IS4	0
yi5A	b3557	Insertion element IS1 50 hypothetical 197 kDa protein	0
yi91b	b4283	Transposase insN for insertion sequence element IS91 1 B	0
yiaK	b3575	Hypothetical oxidoreductase yiaK	0
yidE	b3685	Putative transport protein yidE	0
yjdE	b4115	Arginine/agmatine antiporter	0
yjeH	b4141	Inner membrane protein yjeH	0
ylcB	b0572	Cation efflux system protein cusC precursor	0
yojH	b2210	Malate:quinone oxidoreductase	0
yrfG	b3399	Hypothetical protein yrfG	0
yrfH	b3400	Heat shock protein 15	0
yrfI	b3401	33 kDa chaperonin	0
ytfE	b4209	Regulator of cell morphogenesis and NO signaling	0

Table 3 Genes which are down-regulated (under-expressed) during recombinant protein production identified by mRNA or protein profiling. The last column represents the number of publications in which the corresponding gene name co-occurs with the term "recombinant protein production" in the PubMed database

Gene	B-number	Gene name	Literature count
alaS	b2697	Alanyl-tRNA synthetase	6
argQ	b2691	arginine tRNA 2 (duplicate of argV,Y,Z)	0
argU	b0536	arginine tRNA 4	7
argV	b2694	arginine tRNA 2 (duplicate of argQ,Y,Z)	0
argZ	b2692	arginine tRNA 2 (duplicate of argV,Y,Q)	0
artP	b0864	Arginine transport ATP-binding protein artP	0
asnU	b1986	asparagine tRNA	0
aspS	b1866	Aspartyl-tRNA synthetase	2
aspT	b3760	aspartate tRNA 1 (duplicate of aspV,U)	0
aspV	b0216	aspartate tRNA 1 (duplicate of aspT,U)	0
atpC	b3731	ATP synthase epsilon chain	0
atpD	b3732	ATP synthase beta chain	1
atpF	b3736	ATP synthase B chain	0
clpA	b0882	ATP-dependent Clp protease ATP-binding subunit clpA	0
clpP	b0437	ATP-dependent Clp protease proteolytic subunit	5
clpX	b0438	ATP-dependent Clp protease ATP-binding subunit clpX	1
deoA	b4382	Thymidine phosphorylase	0
deoB	b4383	Phosphopentomutase	1
dnaJ	b0015	Chaperone protein dnaJ	30
dnaK	b0014	Chaperone protein dnaK	46
dps	b0812	DNA protection during starvation protein	6
edd	b1851	Phosphogluconate dehydratase	0
emrR	b2684	Transcriptional repressor mprA	1
eno	b2779	Enolase	2
fba	b2925	Fructose-bisphosphate aldolase class 2	3
fecB	b4290	Iron	1
feoA	b3408	Ferrous iron transport protein A	0
fmt	b3288	Meth ionyl-tRNA formyltransferase	0
ftsJ	b3179	Ribosomal RNA large subunit methyltransferase J	0
ftsZ	b0095	Cell division protein ftsZ	8
gapA	b1779	Glyceraldehyde-3-phosphate dehydrogenase A	3
glpD	b3426	Aerobic glycerol-3-phosphate dehydrogenase	2
glpF	b3927	Glycerol uptake facilitator protein	3
gltT	b3969	glutamate tRNA 2 (duplicate of gltU,V,W)	0
gltU	b3757	glutamate tRNA 2 (duplicate of gltT,V,W)	0
gltV	b4008	glutamate tRNA 2 (duplicate of gltT,U,W)	0
gltW	b2590	glutamate tRNA 2 (duplicate of gltT,U,V)	0
gltX	b2400	Glutamyl-tRNA synthetase	0
glyQ	b3560	Glycyl-tRNA synthetase alpha chain	0
glyS	b3559	Glycyl-tRNA synthetase beta chain	0
hemM	b1209	Outer-membrane lipoprotein lolB precursor	0
hflB	b3178	Cell division protein ftsH	1
hflX	b4173	GTP-binding protein hflX	0
hfq	b4172	Protein hfq	2
hisS	b2514	H istidyl-tRNA synthetase	0
hrpB	b0148	ATP-dependent RNA helicase hrpB	0

(*continued*)

Table 3 (continued)

Gene	B-number	Gene name	Literature count
hslV	b3932	ATP-dependent protease hslV	0
htrA	b0161	Protease do precursor	8
ileS	b0026	Isoleucyl-tRNA synthetase	1
leuS	b0642	Leucyl-tRNA synthetase	0
leuT	b3798	leucine tRNA 1 (duplicate of leuQ,P,V)	0
lon	b0439	ATP-dependent protease La	14
lpdA	b0116	Dihydrolipoyl dehydrogenase	4
lysS	b2890	Lysyl-tRNA synthetase	2
lysU	b4129	Lysyl-tRNA synthetase, heat inducible	0
metG	b2114	Methionyl-tRNA synthetase	1
metL	b3940	Bifunctional aspartokinase/homoserine dehydrogenase II	0
metW	b2815	initiator methionine tRNA-f1 (duplicate of metZ,V)	0
metY	b3171	initiator methionine tRNA-f2	0
metZ	b2814	initiator methionine tRNA-f1 (duplicate of metW,V)	15
miaA	b4171	tRNA delta(2)-isopentenylpyrophosphate transferase	0
mopA	b4143	60 kDa chaperonin	25
mopB	b4142	10 kDa chaperonin	0
narK	b1223	Nitrite extrusion protein 1	0
nohB	b0560	Prophage QSR' DNA packaging protein NU1 homolog	0
nrdB	b2235	Ribonucleoside-diphosphate reductase 1 beta subunit	1
nuoC	b2286	NADH-quinone oxidoreductase chain C/D	0
ompT	b0565	Protease 7 precursor	15
osmY	b4376	Osmotically-inducible protein Y precursor	2
pheS	b1714	Phenylalanyl-tRNA synthetase alpha chain	1
pheT	b1713	Phenylalanyl-tRNA synthetase beta chain	0
ppiB	b0525	Peptidyl-prolyl cis-trans isomerase B	0
pspA	b1304	Phage shock protein A	16
rbsC	b3750	Ribose transport system permease protein rbsC	1
recR	b0472	Recombination protein recR	0
rfaD	b3619	AD P-L-g lycero-D-man no-heptose-6-epi merase	0
rfbD	b2040	dTD P-4-dehydrorham nose reductase	1
rhsD	b0497	Protein rhsD precursor	0
rplA	b3984	50S ribosomal protein L1	4
rplB	b3317	50S ribosomal protein L2	0
rplC	b3320	50S ribosomal protein L3	2
rplD	b3319	50S ribosomal protein L4	1
rplE	b3308	50S ribosomal protein L5	0
rplF	b3305	50S ribosomal protein L6	2
rplJ	b3985	50S ribosomal protein L1 0	2
rplK	b3983	50S ribosomal protein L1 1	1
rplL	b3986	50S ribosomal protein L7/L12	1
rplM	b3231	50S ribosomal protein L13	0
rplN	b3310	50S ribosomal protein L14	0
rplO	b3301	50S ribosomal protein L15	0
rplP	b3313	50S ribosomal protein L16	0
rplQ	b3294	50S ribosomal protein L1 7	0
rplR	b3304	50S ribosomal protein L1 8	0
rplS	b2606	50S ribosomal protein L1 9	1
rplT	b1716	50S ribosomal protein L20	0

(*continued*)

Table 3 (continued)

Gene	B-number	Gene name	Literature count
rplU	b3186	50S ribosomal protein L21	0
rplV	b3315	50S ribosomal protein L22	0
rplW	b3318	50S ribosomal protein L23	0
rplX	b3309	50S ribosomal protein L24	0
rplY	b2185	50S ribosomal protein L25	1
rpmA	b3185	50S ribosomal protein L27	0
rpmB	b3637	50S ribosomal protein L28	0
rpmC	b3312	50S ribosomal protein L29	1
rpmD	b3302	50S ribosomal protein L30	0
rpmE	b3936	50S ribosomal protein L31	0
rpmF	b1089	50S ribosomal protein L32	0
rpmG	b3636	50S ribosomal protein L33	0
rpmH	b3703	50S ribosomal protein L34	0
rpmI	b1717	50S ribosomal protein L35	19
rpmJ	b3299	50S ribosomal protein L36	0
rpoA	b3295	DNA-directed RNA polymerase alpha chain	1
rpoB	b3987	DNA-directed RNA polymerase beta chain	3
rpoC	b3988	DNA-directed RNA polymerase beta' chain	0
rpoD	b3067	RNA polymerase sigma factor rpoD	2
rpoE	b2573	RNA polymerase sigma-E factor	1
rpoH	b3461	RNA polymerase sigma-32 factor	6
rpoN	b3202	RNA polymerase sigma-54 factor	5
rpoS	b2741	RNA polymerase sigma factor rpoS	12
rpoZ	b3649	DNA-directed RNA polymerase omega chain	0
rpsA	b0911	30S ribosomal protein S1	4
rpsB	b0169	30S ribosomal protein S2	1
rpsC	b3314	30S ribosomal protein S3	0
rpsD	b3296	30S ribosomal protein S4	0
rpsE	b3303	30S ribosomal protein S5	0
rpsF	b4200	30S ribosomal protein S6	1
rpsG	b3341	30S ribosomal protein S7	0
rpsH	b3306	30S ribosomal protein S8	0
rpsI	b3230	30S ribosomal protein S9	0
rpsJ	b3321	30S ribosomal protein S10	0
rpsK	b3297	30S ribosomal protein S11	0
rpsL	b3342	30S ribosomal protein S12	0
rpsM	b3298	30S ribosomal protein S13	1
rpsN	b3307	30S ribosomal protein S14	0
rpsO	b3165	30S ribosomal protein S15	0
rpsP	b2609	30S ribosomal protein S16	0
rpsQ	b3311	30S ribosomal protein S17	0
rpsR	b4202	30S ribosomal protein S18	0
rpsS	b3316	30S ribosomal protein S19	1
rpsU	b3065	30S ribosomal protein S21	0
rrfA	b3855	5S rRNA	0
rrfB	b3971	5S rRNA	0
rrfC	b3759	5S rRNA	0
rrfD	b3274	5S rRNA	0
rrfE	b4010	5S rRNA	0

(*continued*)

Table 3 (continued)

Gene	B-number	Gene name	Literature count
rrfF	b3272	5S rRNA	0
rrfG	b2588	5S rRNA	0
rrfH	b0205	5S rRNA	0
rrlA	b3854	23S rRNA	0
rrlC	b3758	23S rRNA	0
rrlD	b3275	23S rRNA	0
rseA	b2572	Sigma-E factor negative regulatory protein	0
serS	b0893	Seryl-tRNA synthetase	1
speD	b0120	S-adenosylmethionine decarboxylase proenzyme	0
talC	b3946	Fructose-6-phosphate aldolase 2	1
thrS	b1719	Threonyl-tRNA synthetase	0
tnaA	b3708	Tryptophanase	6
tnaL	b3707	Tryptophanase leader peptide	0
upp	b2498	Uracil phosphoribosyltransferase	1
ybeY	b0659	UPF0054 protein ybeY	0
ybjC	b0850	Hypothetical protein ybjC	0
ycaR	b0917	Protein ycaR	0
yccA	b0970	Inner membrane protein yccA	0
yccV	b0966	Hypothetical protein yccV	0
ycfN	b1106	Thiamine kinase	0
yeaF	b1782	MltA-interacting protein precursor	0
yedU	b1967	Chaperone protein hchA	0
yfiD	b2579	Autonomous glycyl radical cofactor	0
yhdN	b3293	Hypothetical protein yhdN	0
yhiE	b3512	Transcriptional regulator gadE	0
yljA	b0881	ATP-dependent Clp protease adaptor protein clpS	0
yqjE	b3099	Inner membrane protein yqjE	0

effective relationship among the various functional categories mentioned in this chapter.

Considerable improvement may come from the utilisation of bacteria with minimal genomes by reducing the complexity and at the same time stabilising the genotype as accomplished for, e.g. *E. coli* [67] and *Bacillus subtilis* [68]. It has also always to be kept in mind that wildtype organisms have developed to cope with natural environments, the conditions of which are fundamentally different from those imposed in industrial cultivation processes. The knock-out of metabolic burden and superfluous regulatory pathways may lead to a more solid foundation for the optimisation of biological processes. The strategy of creating organisms with minimal genomes could pave the way.

Integrating metabolome and especially fluxome data along with proteomics and transcriptomics is necessary in order to understand the interplay of media composition and recombinant protein yield. These studies have most often been carried out under well defined conditions with respect to media composition in order to be able to close the mass balance for the main elemental sources. Thus, the influence of single nutrients can be studied on different molecular levels like for transcription and metabolic fluxes, e.g. [69, 70]. Often, flux analyses are performed without

combination with other molecular levels since it integrates the outcome of all lower-level influences, e.g. [71–73]. Such analyses have been applied to study the impact of recombinant proteins on the metabolism of the producing bacterium *E. coli* [74–76] or other microorganisms [77, 78]. However, the metabolic state of microorganisms is a much better indicator for processes directed to the production of low molecular mass metabolites instead of recombinant proteins. Problems with protein expression are diverse and involve the whole cellular machinery. In consequence, primary information is to be expected from data representing the transcriptome and the proteome. Since even this information is often not precise enough, novel techniques may lead to a more complete picture in the future. Such novel methods may be found in applying, e.g. novel massively parallel sequencing techniques to detect mRNA during the process of translation by ribosome profiling [79]. Thus, the work of the ribosomes would be visible in details which have not been accessible up to now.

These endeavours embracing classical as well as novel approaches should lead to a more complete understanding of the molecular phenomena on the cellular level which can be translated into models describing the behaviour of microorganisms under different cultivation conditions – finally leading to combining of systems bio (techno)logy with bioprocess engineering [80], integrating experimental evidence from transcriptome, proteome and fluxome in order to approximate the real cellular behaviour. Since recombinant protein expression is thought to represent a heavy metabolic burden for the synthesising cellular machinery, the metabolic stress is addressed as a primary variable. Thus, it has to be kept in mind that the specific growth rate may also be a central variable of influence on protein yield and quality. This main variable can be controlled to a large extent by means of different process engineering strategies. It is here where cellular and engineering information will have to be integrated in order to obtain optimal process performance for recombinant protein production. The complexity of the task is not the only reason that the whole living nature has to be considered as host systems for the economic expression of recombinant proteins [81].

References

1. Schena M, Shalon D, Davis RW, Brown PO (1995) Science 270:467
2. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Science 270:484
3. Patterson SD, Aebersold RH (2003) Nat Genet 33 (Suppl):311
4. Dutt MJ, Lee KH (2000) Curr Opin Biotechnol 11:176
5. Pandey A, Mann M (2000) Nature 405:837
6. Rabilloud T (2002) Proteomics 2:3
7. Aldor IS, Krawitz DC, Forrest W, Chen C, Nishihara JC, Joly JC, Champion KM (2005) Appl Environ Microbiol 71:1717
8. Choi JH, Lee SJ, Lee SY (2003) Appl Environ Microbiol 69:4737
9. Han MJ, Jeong KJ, Yoo JS, Lee SY (2003) Appl Environ Microbiol 69:5772
10. Quackenbush J (2002) Nat Genet 32 (Suppl):496
11. Rocke DM, Durbin B (2003) Bioinformatics 19:966

12. Pounds S, Cheng C (2005) J Comput Biol 12:482
13. Butte A (2002) Nat Rev Drug Discov 1:951
14. Allison DB, Cui X, Page GP, Sabripour M (2006) Nat Rev Genet 7:55
15. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA et al (1999) Science 286:531
16. D'Haeseleer P (2005) Nat Biotechnol 23:1499
17. Quackenbush J (2001) Nat Rev Genet 2:418
18. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Proc Natl Acad Sci USA 95:14863
19. Nikkila J, Toronen P, Kaski S, Venna J, Castren E, Wong G (2002) Neural Netw 15:953
20. Toronen P, Kolehmainen M, Wong G, Castren E (1999) FEBS Lett 451:142
21. Raychaudhuri S, Stuart JM, Altman RB (2000) Pac Symp Biocomput: 455
22. Khan J, Wei JS, Ringner M, Saal LH, Ladanyi M, Westermann F, Berthold F, Schwab M, Antonescu CR, Peterson C et al (2001) Nat Med 7:673
23. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares M Jr, Haussler D (2000) Proc Natl Acad Sci USA 97:262
24. Durrschmid K, Marzban G, Durrschmid E, Striedner G, Clementschitsch F, Cserjan-Puschmann M, Bayer K (2003) Electrophoresis 24:303
25. Franchini AG, Egli T (2006) Microbiology 152:2111
26. Haddadin FT, Harcum SW (2005) Biotechnol Bioeng 90:127
27. Raman B, Nandakumar MP, Muthuvijayan V, Marten MR (2005) Biotechnol Bioeng 92:384
28. Yoon SH, Han MJ, Lee SY, Jeong KJ, Yoo JS (2003) Biotechnol Bioeng 81:753
29. Lockhart DJ, Winzeler EA (2000) Nature 405:827
30. Richmond CS, Glasner JD, Mau R, Jin H, Blattner FR (1999) Nucleic Acids Res 27:3821
31. Tao H, Bausch C, Richmond C, Blattner FR, Conway T (1999) J Bacteriol 181:6425
32. Wei Y, Lee JM, Richmond C, Blattner FR, Rafalski JA, LaRossa RA (2001) J Bacteriol 183:545
33. Vijayendran C, Polen T, Wendisch VF, Friehs K, Niehaus K, Flaschel E (2007) J Biotechnol 128:747
34. Vijayendran C, Barsch A, Friehs K, Niehaus K, Becker A, Flaschel E (2008) Genome Biol 9: R72
35. Anderson NL, Matheson AD, Steiner S (2000) Curr Opin Biotechnol 11:408
36. Chen C, Snedecor B, Nishihara JC, Joly JC, McFarland N, Andersen DC, Battersby JE, Champion KM (2004) Biotechnol Bioeng 85:463
37. Han MJ, Lee SY (2003) Proteomics 3:2317
38. Jurgen B, Lin HY, Riemschneider S, Scharf C, Neubauer P, Schmid R, Hecker M, Schweder T (2000) Biotechnol Bioeng 70:217
39. Gill RT, Valdes JJ, Bentley WE (2000) Metab Eng 2:178
40. Gill RT, DeLisa MP, Valdes JJ, Bentley WE (2001) Biotechnol Bioeng 72:85
41. Harcum SW, Haddadin FT (2006) J Ind Microbiol Biotechnol 33:801
42. Lesley SA, Graziano J, Cho CY, Knuth MW, Klock HE (2002) Protein Eng 15:153
43. Henry MD, Yancey SD, Kushner SR (1992) J Bacteriol 174:743
44. Baneyx F, Mujacic M (2004) Nat Biotechnol 22:1399
45. Yabuta M, Onai-Miura S, Ohsuye K (1995) J Biotechnol 39:67
46. Harcum SW, Bentley WE (1999) Appl Biochem Biotechnol 80:23
47. Makrides SC (1996) Microbiol Rev 60:512
48. Schmidt M, Babu KR, Khanna N, Marten S, Rinas U (1999) J Biotechnol 68:71
49. Dong H, Nilsson L, Kurland CG (1995) J Bacteriol 177:1497
50. Rinas U (1996) Biotechnol Prog 12:196
51. Grossman AD, Straus DB, Walter WA, Gross CA (1987) Genes Dev 1:179
52. Champion KM, Nishihara JC, Aldor IS, Moreno GT, Andersen D, Stults KL, Vanderlaan M (2003) Proteomics 3:1365
53. Weiner L, Brissette JL, Model P (1991) Genes Dev 5:1912
54. Model P, Jovanovic G, Dworkin J (1997) Mol Microbiol 24:255

55. Birnbaum S, Bailey JE (1991) Biotechnol Bioeng 37:736
56. Schweder T, Kruger E, Xu B, Jurgen B, Blomsten G, Enfors SO, Hecker M (1999) Biotechnol Bioeng 65:151
57. Sanden AM, Prytz I, Tubulekas I, Forberg C, Le H, Hektor A, Neubauer P, Pragai Z, Harwood C, Ward A et al (2003) Biotechnol Bioeng 81:158
58. Taddei F, Vulic M, Radman M, Matic I (1997) EXS 83:271
59. Naas T, Blot M, Fitch WM, Arber W (1994) Genetics 136:721
60. Naas T, Blot M, Fitch WM, Arber W (1995) Mol Biol Evol 12:198
61. Wang Y, Wu SL, Hancock WS, Trala R, Kessler M, Taylor AH, Patel PS, Aon JC (2005) Biotechnol Prog 21:1401
62. Wolfe CJ, Kohane IS, Butte AJ (2005) BMC Bioinformatics 6:227
63. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M et al (2009) Nucleic Acids Res 37:D412
64. Barabasi AL, Oltvai ZN (2004) Nat Rev Genet 5:101
65. Thomas JG, Baneyx F (2000) Mol Microbiol 36:1360
66. Lund PA (2001) Adv Microb Physiol 44:93
67. Sharma SS, Blattner FR, Harcum SW (2007) Metab Eng 9:133
68. Morimoto T, Kadoya R, Endo K, Tohata M, Sawada K, Liu S, Ozawa T, Kodama T, Kakeshita H, Kageyama Y et al (2008) DNA Res 15:73
69. Oh MK, Liao JC (2000) Biotechnol Prog 16:278
70. Oh MK, Rohlin L, Kao KC, Liao JC (2002) J Biol Chem 277:13175
71. Selvarasu S, Ow DS, Lee SY, Lee MM, Oh SK, Karimi IA, Lee DY (2009) Biotechnol Bioeng 102:923
72. Guebel DV, Canovas M, Torres NV (2009) Biotechnol Bioeng 102:910
73. Hardiman T, Lemuth K, Keller MA, Reuss M, Siemann-Herzberg M (2007) J Biotechnol 132:359
74. Hoffmann F, Weber J, Rinas U (2002) Biotechnol Bioeng 80:313
75. Weber J, Hoffmann F, Rinas U (2002) Biotechnol Bioeng 80:320
76. Wittmann C, Weber J, Betiku E, Kromer J, Bohm D, Rinas U (2007) J Biotechnol 132:375
77. Furch T, Wittmann C, Wang W, Franco-Lara E, Jahn D, Deckwer WD (2007) J Biotechnol 132:385
78. Melzer G, Dalpiaz A, Grote A, Kucklick M, Gocke Y, Jonas R, Dersch P, Franco-Lara E, Nortemann B, Hempel DC (2007) J Biotechnol 132:405
79. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS (2009) Science 324:218
80. Teixeira AP, Carinhas N, Dias JM, Cruz P, Alves PM, Carrondo MJ, Oliveira R (2007) J Biotechnol 132:418
81. Demain AL, Vaishnav P (2009) Biotech Adv 27(3):297
82. Bader GD, Hogue CW (2003) BMC Bioinformatics 4:2

Adv Biochem Engin/Biotechnol (2010) 121: 71–107
DOI: 10.1007/10_2009_61

Published online: 17 February 2010

Engineering the *Escherichia coli* Fermentative Metabolism

M. Orencio-Trejo, J. Utrilla, M.T. Fernández-Sandoval, G. Huerta-Beristain, G. Gosset, and A. Martinez

Abstract Fermentative metabolism constitutes a fundamental cellular capacity for industrial biocatalysis. *Escherichia coli* is an important microorganism in the field of metabolic engineering for its well-known molecular characteristics and its rapid growth. It can adapt to different growth conditions and is able to grow in the presence or absence of oxygen. Through the use of metabolic pathway engineering and bioprocessing techniques, it is possible to explore the fundamental cellular properties and to exploit its capacity to be applied as industrial biocatalysts to produce a wide array of chemicals. The objective of this chapter is to review the metabolic engineering efforts carried out with *E. coli* by manipulating the central carbon metabolism and fermentative pathways to obtain strains that produce metabolites with high titers, such as ethanol, alanine, lactate and succinate.

Keywords Alanine, Central carbon metabolism, *Escherichia coli*, Ethanol, Fermentative metabolism, Glucose, Lactate, Metabolic engineering, Succinate, Xylose

Contents

A. Martinez (✉), M. Orencio-Trejo, J. Utrilla, M.T. Fernández-Sandoval, G. Huerta-Beristain, and G. Gosset
Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Apdo. Postal 510-3, Cuernavaca, Morelos 62250, México
e-mail: alfredo@ibt.unam.mx

1 Introduction

Metabolic networks in *Escherichia coli* consist of hundreds of metabolites that are interconnected through a large number of biochemical and regulatory reactions [1]. In principle, metabolites could flow through various reactions; however, under some conditions several reactions are not used by the cell [2]. Fermentative metabolism constitutes a fundamental cellular capacity for industrial biocatalysis. Endogenous organic compounds used by cells as terminal electron acceptors – under oxygen deprivation – are converted into biochemical products that represent valuable molecules to humanity [3]. *E. coli* is an important microorganism in the field of metabolic engineering for its well-known molecular characteristics and its rapid growth. It has several alternative pathways and sophisticated sensing mechanisms for cell growth and survival under fermentative conditions [4].

Growth can be carried out by fermentative processes, in which redox balance is achieved internally, or by respiratory processes, requiring an exogenous electron acceptor. Compounds such as fumarate, trimethylamine *N*-oxide, dimethylsulfoxide, nitrite, nitrate, or O_2 can serve as the acceptor. Reduction of the electron acceptor is catalyzed by a membrane-associated complex, which results in efflux of H^+, thus generating a proton-motive force. This protonic potential enables the influx of H^+ across the plasma membrane to drive energy-requiring processes, such as the synthesis of ATP, the transport of nutrients, or cell motility. Whenever an exogenous electron acceptor is available, the cell restrains its fermentation process in favor of respiration, thereby diverting most of the carbon source for biosynthesis. Not only does the cell favor respiration over fermentation, given a choice of electron acceptors, it also preferentially uses the compound that has the highest redox potential [5].

In *E. coli*, the fermentation of glucose may be regarded as a process in which the four extra reducing equivalents (or $4H^+ + 4e^-$) released, as the metabolic price for making two ATPs by substrate-level phosphorylation, are disposed of by the further metabolism of two pyruvate molecules. Glucose is converted to a mixture of metabolic products consisting primarily of acetate and formate, as well as smaller amounts of lactate, succinate and ethanol (Fig. 1) [6, 7]. The proportions of the excreted products depend on the growth conditions, thereby indicating that the cell possesses considerable flexibility in maintaining its redox and energy balance [7].

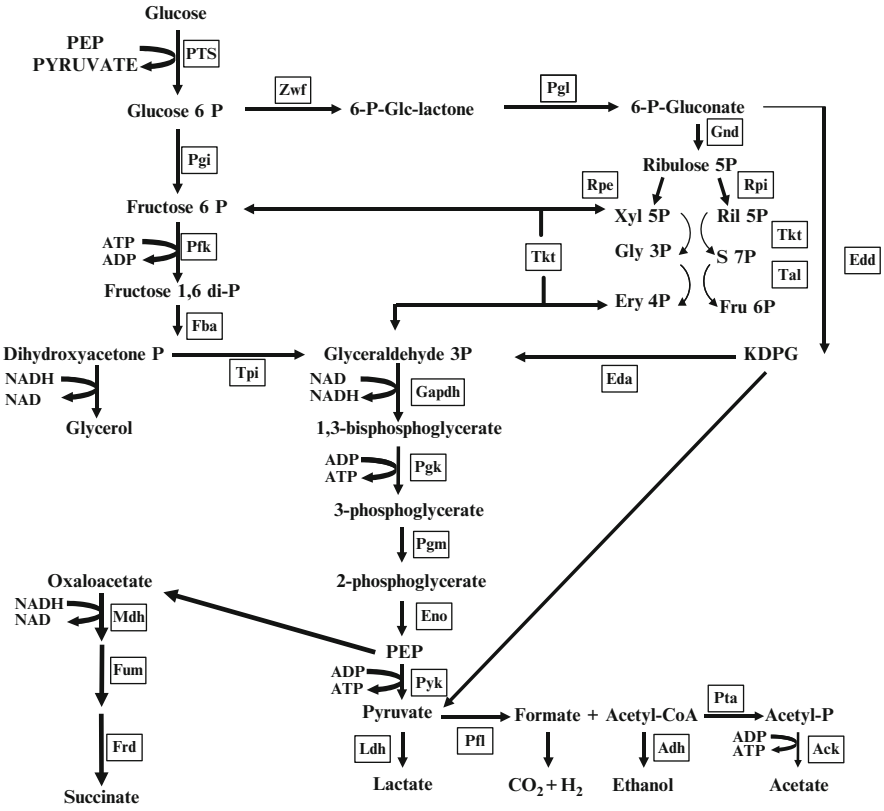


Fig. 1 Carbon central metabolism in wild type *E. coli* under anaerobic conditions. *Rectangles* indicates the main enzymes. Abbreviations: *PTS* glucose – phosphoenolpyruvate phosphotransferase system, *Pgi* phosphoglucose isomerase, *Pfk* phosphofructokinase, *Fba* fructose bisphosphate aldolase, *Tpi* triose phosphate isomerase, *Gapdh* glyceraldehyde-3P dehydrogenase, *Pgk* phosphoglycerate kinase, *PgmA* phosphoglycerate mutase, *Eno* enolase, *Pyk* pyruvate kinase, *Zwf* glucose-6P-1-dehydrogenase, *Gnd* 6-phosphogluconate dehydrogenase, *Rpe* ribulose phosphate epimerase, *Rpi* ribose-5-phosphate isomerase, *Tkt* transketolase, *Tal* transaldolase, *Eda* 2-keto-3-deoxy-phosphogluconate aldolase, *Edd* phosphogluconate dehydratase, *Xyl 5P* xylulose-5-phosphate, *Ril 5P* ribose-5-phosphate, *Gly 3P* glyceraldehyde-3-phosphate, *S 7P* sedoheptulose-7-phosphate, *Ery 4P* erythrose 4-phosphate, *Fru 6P* fructose 6-phosphate, *KDPG* 2-keto-3-deoxy-gluconate-6-phosphate, *Frd* fumarate reductase, *Pfl* pyruvate formate lyase, *Mdh* malate dehydrogenase, *Fum* fumarase, *Adh* alcohol dehydrogenase, *Ldh* lactate dehydrogenase, *PTA* phosphotransacetylase, *Ack* acetate kynase

Because the products have different oxidation states, *E. coli* can modulate its metabolism to grow on different carbon sources, adjusting the proportion of each produced compound [8, 9]. The objective of this chapter is to review the metabolic engineering works carried out using *E. coli* that have manipulated the central carbon metabolism and fermentative pathways to obtain strains that produce metabolites with high titers.

2 Sugar Transport

Culture media in industrial fermentative processes frequently employ sugars that constitute the main raw material for biomass and product generation. Glucose is the most utilized raw material in fermentations with *E. coli*, mostly because it is relatively inexpensive and is the preferred carbon and energy source for this microorganism. Glucose is internalized into the cytoplasm by the phosphoenolpyruvate:sugar phosphotransferase system (PTS). This is a protein system that participates in the transport and phosphorylation of glucose and other sugars [10, 11]. It is composed of the soluble and nonsugar-specific protein components Enzyme I (EI) and the phosphohistidine carrier protein (HPr) (Fig. 2). These proteins relay a phosphoryl group from phosphoenolpyruvate (PEP) to the sugar-specific enzymes IIA and IIB. Component IIC (and in some cases IID) is an integral membrane protein permease that recognizes and transports the sugar molecules, which are phosphorylated by component IIB. There have been 21 separately identified enzyme II complexes encoded in the *E. coli* chromosome, each involved in the transport of several different sugars [12]. In *E. coli*, the glucose-specific II^{Glc} PTS complex is composed of the soluble IIA^{Glc} enzyme and the integral membrane permease $IICB^{Glc}$ [13, 14].

Activity of the PTS results in a tight linkage between sugar transport and its subsequent metabolism due to its dependence on PEP as a phosphate group donor for sugar internalization. When *E. coli* grows in minimal medium containing glucose as the carbon source, PTS consumes 50% of the available PEP [15]. The PTS therefore plays a major role in determining the PEP/PYR ratio and carbon flux distribution originating from these metabolic nodes. It can be expected that

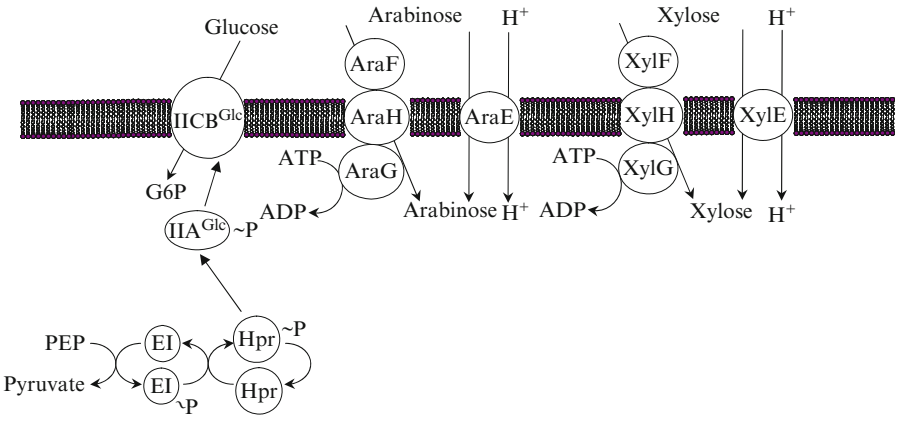


Fig. 2 Glucose, arabinose, and xylose transport systems in *Escherichia coli*. The general energy coupling proteins and the transporting complex for glucose from the phosphoenolpyruvate:sugar phosphotransferase system are shown as well as the ATP-dependent and symporters for arabinose and xylose

modification of PTS activity should have an important impact on carbon flux distribution in central metabolism. Mutant strains lacking EI and HPr exhibit a very limited capacity to transport and phosphorylate glucose and have a very low specific growth rate (PTS^- Glc^- phenotype) [16]. The reconstitution of PTS-independent glucose transport in these PTS^- Glc^- mutants has been achieved using different strategies. Using glucose as the limiting nutrient in a continuous culture it was possible to isolate, from a PTS^- Glc^- strain, mutants capable of internalizing glucose at a rate similar to that of a PTS^+ strain (PTS^- Glc^+ phenotype) [17]. It was determined that in these mutants, glucose is transported and phosphorylated by galactose permease (GalP) and glucokinase (Glk), respectively [17, 18]. Another strategy to augment glucose transport capacity in a PTS^- Glc^- mutant is to increase expression of the genes encoding GalP and Glk [19].

Modification of PTS activity has enabled the improvement of *E. coli* strains to produce useful metabolites derived from central metabolism. *E. coli* strains developed for the production of succinate have been generated by modifying central metabolic pathways to redirect carbon flux from the EMP and TCA pathways to the enzymes isocitrate lyase and succinyl-CoA synthetase, leading to succinate synthesis [20]. It has been determined that inactivation of the gene *ptsG*, which encodes the glucose-specific component $IICB^{Glc}$, resulted in a 23% increase in final succinate titer, a 22% higher specific productivity and a 16% increase in succinate molar yield from glucose when compared to an isogenic $ptsG^+$ strain [21]. The replacement of PTS-dependent activity by GalP and Glk has been shown to cause a twofold increase in the specific rate of acetate production when compared to a PTS^+ strain [19]. These strains were transformed with a plasmid carrying the pdc_{Zm} and $adhB_{Zm}$ genes from *Zymomonas mobilis*, encoding pyruvate decarboxylase and alcohol dehydrogenase II, respectively. These enzymes convert pyruvate to ethanol [22]. It was determined that the PTS^- Glc^+ strain displayed a twofold increase in the specific rate of ethanol production when compared to a PTS^+ strain [19].

3 Central Carbon Metabolism Pathways

3.1 Glycolytic Pathway

For several reasons, glycolysis can be viewed as one of the primordial pathways established in the evolution of metabolism, and many other carbon and energy sources metabolically feed into this pathway. First, the phosphorylated intermediates, the coenzymes and energy carrier molecules (e.g., triosephosphates, NAD/NADH and ADP/ATP) are also used in numerous other pathways, including those participating in biosynthesis. Additionally, glucose likely emerged as the sugar of choice because of its high ring structure stability, which limits potentially deleterious nonenzymatic glycosylation of proteins [5].

From glucose, *E. coli* can obtain the carbon skeletons for every amino acid, nucleotide, coenzyme, fatty acid, or other metabolic products needed for growth. Under both aerobic and anaerobic conditions, glycolysis is the sequence of 10 enzymatic reactions that converts glucose into pyruvate (Fig. 1). In this process, some of the free energy released from glucose is conserved in the form of ATP and NADH. Glycolysis can be divided into two stages. The first involves four consecutive reactions that finally result in the splitting of the glucose molecule (C_6) into two phosphoglyceraldehyde (C_3) molecules. Two ATP molecules are used per mol of metabolized glucose, and these donate the phosphoryl groups present in phosphoglyceraldehyde. The second stage of glycolysis consists of five consecutive reactions that result in the oxidation of phosphoglyceraldehyde to pyruvate. This stage generates four ATP molecules per mol of metabolized glucose, resulting in a net yield of two ATP molecules.

It is well known that *E. coli* controls the synthesis of its respiratory pathway enzymes in response to aerobic and anaerobic cell growth conditions, and that this control depends on whether the alternative anaerobic respiratory substrates are present. In anaerobic respiration, *E. coli* can produce terminal oxidoreductases with the alternative electron acceptors, nitrate, dimethyl sulfoxide (DMSO), trimethylamine-*N*-oxide (TMAO), and fumarate. Each enzyme is able to couple oxidation of NADH to a variety of other electron donors via the cellular quinone pool. Upon oxygen depletion, synthesis of the anaerobic functioning enzymes allows the energetically less favorable electron acceptors to be used for respiratory metabolism in lieu of fermentation. Synthesis of the anaerobic oxidoreductases enzymes is nitrate dependent; nitrate reductase levels are elevated in the presence of this preferred electron acceptor, whereas levels of the DMSO/TMAO reductase and fumarate reductase remain low. When nitrate is depleted from the medium or is absent, the remaining anaerobic respiratory enzymes accumulate to optimal levels. In absence of any of these alternative anaerobic electron acceptor substrates, the cell resorts to a mixed-acid fermentation [23].

Those systems channel electrons from donor to terminal acceptors such that the overall potential difference is maximized for any given growth condition [24]. The adaptive responses are coordinated by a group of global regulators, which include the one-component Fnr (fumarate, nitrate reduction) protein and the two-component Arc (aerobic respiration control) system [5, 23–27].

3.1.1 Transcriptional Regulation of Glycolytic Genes

The preferred nutritional status of glucose for *E. coli* is shown by the observed repression and inhibition exerted by this sugar on gene expression and the activities of enzymes and transporters related to the consumption of other carbon sources [28]. When comparing transcript levels between growth in complex medium with and without glucose, it is common to observe minimal changes in the levels of transcription of glycolytic genes [16, 29, 30]. When they occur, however, they are usually related to energy and NADH-NAD balance, such as that found in the *pfk*,

gapA, and *pykF* genes [31–33]. There are several regulatory proteins, such as ArcA, Crp, FruR, Fnr, and Mlc, that modulate the central metabolic gene expression of *E. coli* at the transcriptional level [1, 34]. Of these, FruR (or Cra) is a global regulator that regulates carbon flow through the central metabolic pathways [35] by controlling the gene expression balance between the glycolytic and gluconeogenic genes [36, 37, 4]. It is known that glycolytic genes, such as *glk*, *pfkA*, *gapA*, *pgk*, *eno*, and *pykF*, are repressed by this protein [38–40]. In addition, *glk* transcription is induced in response to stress caused by the overexpression of foreign proteins [32, 41] and in response to PTS deletion [16].

It has been found that the transcriptional regulatory region of gene *gapA* contains four promoter sequences, three recognized by the vegetative RNA polymerase $E\sigma^{70}$, and one recognized by the heat shock RNA polymerase $E\sigma^{32}$. Transcription of *gapA* by $E\sigma^{32}$ is activated in the logarithmic phase under conditions of starvation and heat shock, and one of the sequences recognized by $E\sigma^{70}$ is subject to catabolite repression. These differentially regulated promoters allow the synthesis of *gapA* transcripts in a wide variety of environmental conditions [42].

3.1.2 Allosteric Regulation of Glycolytic Enzymes

In metabolic pathways, it is common to consider that enzymes catalyzing essentially irreversible reactions are potential sites for control. In glycolysis, the reactions catalyzed by phosphofructokinase (Pfk) and pyruvate kinase (Pyk) are virtually irreversible. In fact, the reaction catalyzed by Pfk was considered the key control step in glycolysis. This enzyme is inhibited by PEP and activated by ADP when either ligand binds to the same allosteric site [43]. Pfk is inhibited by ATP, but the inhibitory effect is reversed by ADP and other phosphonucleosides [44, 45]. Another key enzyme in the glycolytic pathway is Pyk. In *E. coli*, there are two Pyk isozymes, Pyk-I or Pyk-F and Pyk-II or Pyk-A, which differ in their kinetic properties [46]. Both isozymes catalyze pyruvate biosynthesis, but under glycolytic conditions, PykF is the isozyme that displays the highest activity [47]. This isozyme controls the outflow from this pathway and is activated by fructose 1,6-bisphosphate, which enables it to keep pace with the oncoming high flux of intermediates. ATP allosterically inhibits PykF to reduce flux when the energy charge is high.

A metabolite that has an important function in allosteric regulation is PEP. It is the most energetic phosphorylated compound of all molecules involved in glucose catabolism. It has been demonstrated that changes in PEP concentration have an effect on glycolytic and TCA cycle flux in *E. coli* [48, 49]. Several years ago, the only known regulatory effect of PEP on glycolysis was the inhibition of Pfk [44]. Nevertheless, it was recently published that PEP is an inhibitor of the enzymes of the initial reactions of glycolysis. It acts as a potent inhibitor of glucokinase in a competitive manner with respect to ATP, and as a less potent inhibitor of phosphoglucoisomerase (Pgi), Pfk and aldolase (Fba) [50].

Coenzymes NAD and NADH play a central role in catabolism. These nucleotides function as the most important redox carriers involved in metabolism. They not only serve as electron acceptors in the breakdown of catabolic substrates, they also provide the cell with the reducing power needed in energy-conserving redox reactions, such as those that occur in anaerobic and aerobic respiration [51]. The internal redox state is important in the allosteric regulation of glycolysis. It is reported that the flow through GAPDH is regulated by the NADH/NAD ratio [52]. Moreover, NADH is a competitive inhibitor of GAPDH [33].

As mentioned earlier, the control of glycolytic flux has been attributed to the enzymes Pfk and Pyk. Several studies, however, have suggested that flux control may be broadly distributed among most of the enzymes in the glycolytic pathway. Recently, it has been proposed that glycolytic flux is controlled by reactions outside this pathway [53], for instance, fermentative or heterologous routes that can supply the balance between redox and ATP requirements [32, 54].

3.2 Pentose Phosphate Pathway

The pentose phosphate pathway (PPP), also called the phosphogluconate pathway, is an alternative route for glucose degradation and consists of two branches. The oxidative branch converts glucose-6-phosphate to 6-phosphogluconate, with the formation of NADPH as the reducing equivalent for biosynthesis and ribulose-5-phosphate as the precursor for nucleic acids. The nonoxidative branch, with fructose-6-phosphate, glyceraldehyde-3-P, and eritrose-4-P as products, interconnects the glycolytic and aromatic pathways with PPP (Fig. 3). The PPP provides the cell with intermediates for the biosynthesis of amino acids, vitamins, nucleotides and cell wall constituents as the lipopolysaccharide layer.

3.2.1 Enzymes of the Oxidative Branch

The first enzyme of the oxidative branch is glucose-6-phosphate dehydrogenase, encoded by the *zwf* gene. This enzyme catalyzes the transfer of a hydride ion to $NADP^+$ from C-1 of glucose-6-phosphate, producing 6-phosphogluconolactone and NADPH. In *E. coli*, the reaction is irreversible. The enzyme is specific for $NADP^+$ and is inhibited by NADPH, fatty acids, and CoA. Then, 6-phosphogluconolactone is dehydrated by 6-phosphogluconolactonase. This activity may be the rate-limiting step of the carbon flow into the PPP when the carbon source limits growth [55]. 6-Phosphogluconate dehydrogenase catalyzes the oxidative decarboxylation of 6-phosphogluconate, a β-hydroxyacid, to ribulose-5-P and CO_2. This reaction is similar to the one catalyzed by the isocitrate dehydrogenase of tricarboxylic cycle acids, allowing the generation of NADPH under fermentative conditions. In addition, in anaerobic cultures, the growth rate is reduced by 11% and 15% when gluconate or glucose is used, respectively, as the carbon source in

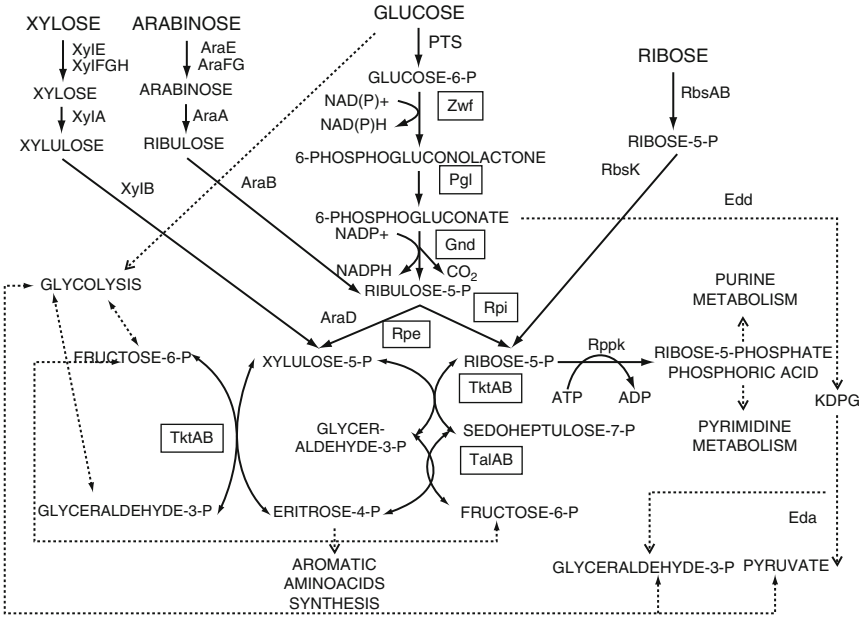


Fig. 3 The pentose phosphate pathway in *Escherichia coli* and catabolic pathways for xylose, ribose, and L-arabinose. *Rectangles* indicates the main enzymes of the pentose phosphate pathway. Abbreviations: *ADP* adenosine diphosphate, *AraA* arabinose isomerase, *AraB* arabinokinase, *AraD* L-ribulose-5-phosphate epimerase, *AraE* arabinose/proton symporter, *AraFG* arabinose ABC (ATP Binding Cassette) transporter, *ADP* adenosine diphosphate, *ATP* adenosine triphosphate, CO_2 carbon dioxide, *Gnd* 6-phosphogluconate dehydrogenase, NAD^+ nicotinamide adenine dinucleotide oxidized, *NADH* nicotinamide adenine dinucleotide reduced, $NADP^+$ nicotinamide adenine dinucleotide phosphate oxidized, *NADPH* nicotinamide adenine dinucleotide phosphate reduced, *Pgl* 6-phosphogluconolactonase, *PTS* phosphoenolpyruvate carbohydrate phosphotransferase system, *RbsAB* ribose/proton symporter, *RbsK* Ribokinase, *Rpe* D-ribulose-5-phosphate epimerase, *Rpi* D-ribose-5-phosphate isomerase, *TalAB* transaldolase, *TktAB* transketolase, *XylA* xylose isomerase, *XylB* xylulokinase, *XylE* xylose/proton symporter, *XylFGH* xylose ABC transporter, *Zwf* glucose-6-phosphate dehydrogenase

a 6-phosphogluconate dehydrogenase mutant [56], indicating that carbon flows through the Entner–Doudoroff and PP pathways.

3.2.2 Enzymes of the Nonoxidative Branch

Ribulose-5-phosphate epimerase (Rpe) catalyzes the conversion of ribulose-5-phosphate to xylulose-5-phosphate. The *rpe* gene is part of a large operon that comprises, among others, genes for the biosynthesis of aromatic amino acids [57]. An *E. coli rpe* mutant is unable to utilize single pentose sugars, and their growth rate is severely impaired when cultured in minimal medium containing glycolytic

carbon sources or gluconate [57]. The ribose-5-phosphate isomerase A is a ubiquitous enzyme that interconverts ribose-5-phosphate and ribulose-5-phosphate [58]. Transketolase (Tkt) transfers two carbon atoms from intermediates of the PPP, thus rearranging the carbon atoms in the molecules that enter this pathway. Like other enzymes that transfer two carbon groups, transketolase requires thiamine pyrophosphate (TPP) as a cofactor. Transaldolase transfers three carbon groups, and thus it is also involved in the rearrangement of the carbon skeletons of the PPP intermediates.

3.2.3 Xylose and Other Pentoses Utilization

The PPP allows *E. coli* to utilize sugars such as D-xylose, D-ribose, and L-arabinose. These sugars cannot be catabolized by other routes [59, 60]. *E. coli* can metabolize all of these sugars to pyruvic acid under anaerobic conditions, and thereafter to a mixture of formate, acetate, lactate, succinate, and a small amount of ethanol [8].

Xylose can be transported through two different systems of inducible permeases. One is a periplasmic protein that represents a high affinity system (Km = 0.3–3.0 μM) and is driven by ATP. The genes encoding this high affinity ABC transporter are organized in a single operon, *xylFGH*. The second is a low affinity system that is energized by proton motive force. This xylose-proton symport system is encoded by the *xyl*E gene, and with *xylA* (xylose isomerase) and *xyl*B (xylulose kinase) belonging to an operon [31] (Figs. 2 and 3).

L-Arabinose is transported by two inducible systems in *E. coli* K12. The first is a low-affinity permease (Km = 100 μM) encoded by *araE* [60], and this system is energized by proton motive force. The second system consists of a high affinity system (Km = 1–3 μM), which is formed by the *araFG* operon [61]. The *araF* gene encodes a periplasmic binding protein with chemostatic receptor function [62], and AraF operates at the expense of a high-energy covalent bond. *araG* encodes an inner membrane protein [63]. Both transport systems are under the control of the *araC* gene product, which is part of the *ara* regulon [64]. L-Arabinose is metabolized by a set of enzymes encoded by the *araBAD* operon, an isomerase (AraA), which converts arabinose to ribulose; a kinase (AraB), which catalyzes the conversion of ribulose to ribulose-5-phosphate; and L-ribulose-5-phosphate epimerase (AraD), which produces xylose-5-phosphate [60]. The expression of this operon is regulated by the AraC protein [65]. Individual isomerase, kinase or epimerase mutants do not growth in complex media when L-arabinose is added. Most strains of *E. coli* cannot grow on D-arabinose, but *E. coli* K12 grows on L-arabinose and *E. coli* B/r grows on D-arabinose [60].

The catabolic genes for ribose uptake and catabolism are organized in a single operon (*rbsDACBK*). The expression of the operon is under control of a repressor encoded by the *rbs*R gene, which is next to the structural genes and is transcribed independently. Ribose, rather than ribose-5-phosphate, is the inducer [60]. The first four genes encode a high affinity ABC transporter and the last encodes a ribokinase. This enzyme (RbsK) is the only enzyme that allows the metabolism of intracellular

ribose. When *E. coli* is grown using xylose, an increased expression of the gene *rbsK* has been observed. This is likely due to the presence of an internal promoter within this operon [31].

3.2.4 Pentose Phosphate Pathway Regulation

Metabolic flow through the pentose phosphate pathway is controlled by the activity of glucose-6-phosphate dehydrogenase, which is controlled by $NADP^+$ availability. Additionally, 6-phosphogluconate dehydrogenase is regulated by the cell growth rate [57]. The pentose phosphate pathway of *E. coli* involves seven enzymes (Fig. 3), most of which are repressed by their final products, i.e., glucose-6-phosphate, fructose-6-phosphate, ribose-5-phosphate, glycerdehyde-3-phosphate, and ribulose-5-phosphate. Activities of glucose-6-phosphate and 6-phosphogluconate dehydrogenases depend of the energy and redox potential of the cell. These enzymes are repressed by ATP, NADH, and NADPH. Reports about transcriptional level regulation of genes from the pentose phosphate pathway indicate that *zwf* is induced by SoxS, MarA, and Rob, *gnd* by GadE, *talA-tktB* by CreB, and *rpiB* by RpiR and LipB [66].

3.3 Entner–Doudoroff Pathway

The Entner–Doudoroff (E–D) pathway is comprised of a series of reactions that catabolize glucose to pyruvate. This pathway occurs only in prokaryotes. Most bacteria use glycolysis and the PPP for glucose metabolism; however, there are few bacteria that substitute glycolysis with the E-D pathway. The distinct feature of this pathway is that it uses 6-phosphogluconate dehydratase (Edd) and 2-keto-3-deoxy-6-phosphogluconate aldolase (Eda) to produce pyruvate from glucose. The E-D pathway has a net yield of 1 mol each of NAD(P)H, glyceraldehide-3-P, and pyruvate. Further metabolism of glyceraldehide-3-P through the triose phosphate pathway yields 1 mol each of ATP, pyruvate, and NADH.

The E-D pathway in *E. coli* is specifically induced by gluconate, allowing its entry into central glycolytic metabolism. The gluconate metabolism involves gluconate transport and phosphorylation, as well as the activity of enzymes of the E-D pathway: Edd and Eda (Fig. 4). An *edd* mutant grows on glucose with a rate identical to that of the wild type, but grows 60–70% slower on gluconate [67]. In contrast, an *eda* mutant does not grow on gluconate or on acid sugars (glucoronate and galactorunate), and grows 25% lower on glucose because 2-keto-3-deoxy-6-phosphogluconate is accumulated and because this metabolite has a bacteriostatic effect [68].

There are two systems for gluconate transporter and phosphorylation in *E. coli*. The GntI system is the main system for gluconate transport, which contains two permeases: one of high affinity (GntT) and the other of low affinity (GntU).

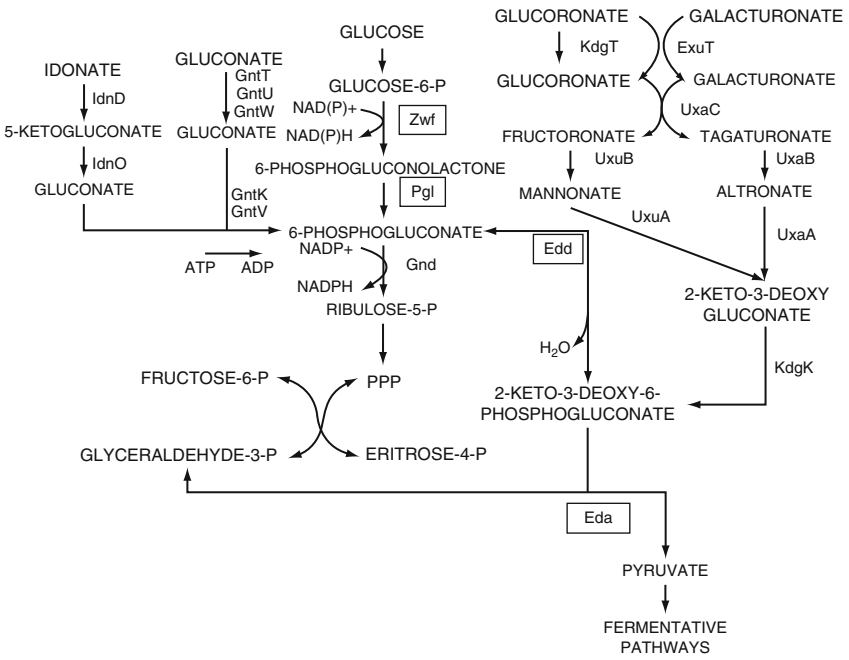


Fig. 4 Entner–Doudoroff pathway in *Escherichia coli* and catabolic pathways for gluconate, glucoronate, galacturonate, idonate, and glucose. *Rectangles* indicates the main enzymes of the Entner–Doudoroff pathway. Abbreviations: *ADP* adenosine diphosphate, *ATP* adenosine triphosphate, CO_2 carbon dioxide, *Eda* 2-keto-3-deoxy-6-phosphogluconate aldolase, *Edd* 6-phosphogluconate dehydratase, *ExuT* hexuronate proton/simporter transporter, *Gnd* 6-phosphogluconate dehydrogenase, *GntT* high affinity gluconate transporter, *GntU* low affinity gluconate transporter, *GntK* gluconokinase I, H_2O water, *IdnD* idonate 5-dehydrogenase, *IdnK (GntV)* gluconokinase II, *IdnO* 5-keto-D-gluconate 5-reductase, *IdnT (GntW)* idonate/5-ketogluconate/gluconate transporter, *KdgK* 2-keto-3-deoxygluconokinase, *KdgT* 2-dehydro-3-deoxy-D-gluconate transporter, *KDPG* 2-keto-3-deoxy-6-phosphogluconate, NAD^+ nicotinamide adenine dinucleotide oxidized, *NADH* nicotinamide adenine dinucleotide reduced, $NADP^+$ nicotinamide adenine dinucleotide phosphate oxidized, *NADPH* nicotinamide adenine dinucleotide phosphate reduced, *Pgl* 6-phosphogluconolactonase, *PPP* Pentose Phosphate Pathway, *UxaA* altronate dehydratase, *UxaB* altronate oxidoreductase, *UxaC* glucuronate isomerase/D-galacturonate isomerase, *UxuA* mannonate dehydratase, *UxuB* mannonate oxidoreductase, *Zwf* glucose-6-phosphate dehydrogenase

This system also contains a thermoresistant glucokinase (GntK). The GntI system is encoded by *gntT* and by the *gntUK* operon [69, 70]. Gluconate acts as the inductor for genes of the GntI system and the E-D pathway (*edd* and *eda*). Both the GntI system and the E-D pathway are negatively regulated by GntR, a repressor protein [71–73]. The other system is GntII, a subsidiary system for gluconate transport, which contains a second high affinity gluconate permease (GntW) and a thermosensitive gluconatokinase (GntV) [69, 71]. The resulting product, 6-phosphogluconate, is further metabolized by two competing enzymes of two alternate pathways. 6-Phosphogluconate is diverted into the oxidative pentose phosphate pathway by

6-phosphogluconate dehydrogenase or into the E-D pathway via Edd and Eda. The expression of *eda* is also subject to negative control by the *kdg*R product, a regulator of the glucoronate and galactorunate metabolism, which also inhibits the expression of the kinase for 2-keto-3-deoxygluconate (KdgK) [74]. The induction of the dehydratase (*edd*) is favored under conditions of low oxygen concentration.

4 Engineering Fermentative Pathways

4.1 Engineering Ethanologenic E. coli *Strains*

4.1.1 First Generation of Metabolic Engineered Ethanologenic Strains

Ethanol is one of the fermentation products generated by *E. coli* under anaerobic conditions. The ethanol yield, however, is very low due to the formation of other fermentation products. Furthermore, the native ethanol pathway is not redox balanced when sugars, like glucose or xylose, are fermented. Only one $NADH^+$ is formed per acetyl-CoA synthesized, but two $NADH^+$ are needed to produce one ethanol molecule via this pathway [8] (Fig. 5). Thus, the native *E. coli* ethanol pathway from acetyl-CoA cannot support homoethanol fermentation due to the need for two NADH molecules per mol of ethanol produced [75].

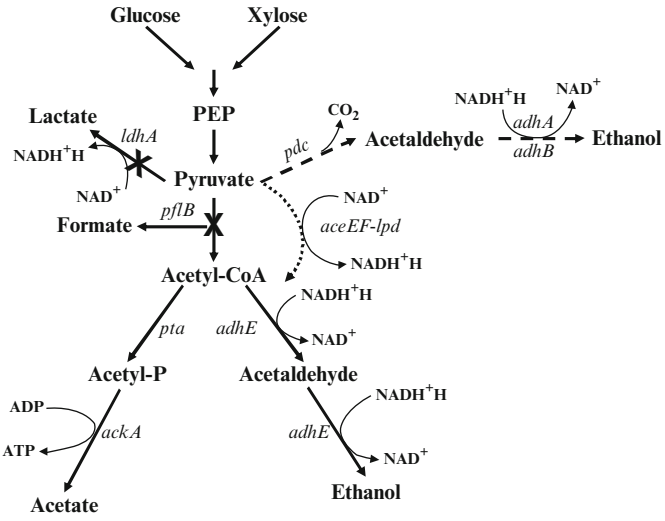


Fig. 5 Fermentation pathways in wild type *E. coli* (*solid arrows*) and engineered ethanologenic strains (*dashed and dotted arrows*). Gene encoding enzymes are indicated by *italics*, the *crossed out pathways* indicate deleted genes to avoid subproducts. *PET* operon pathway (*pdc*, *adh*A, and *adhB*) shown by the *dashed arrow* and novel ethanol pathway in *E. coli* without foreign genes (*aceEF-lpd*) is indicated by the *dotted arrow*

Great interest has recently emerged for producing fuel ethanol from lignocellulosic residues [76, 77]. Lignocellulose (biomass) is a complex substance composed mainly by cellulose, hemicellulose, lignin, and pectin. For ethanol production from the sugars presented in some polymers of the lignocellulose, *E. coli* presents some advantages over other ethanologenic microorganisms (e.g., *Saccharomyces* and *Zymomonas*), primarily a broad substrate-utilization range, such as hexoses (glucose, mannose, galactose, and fructose), pentoses (xylose and arabinose) and uronic acids [60]. In addition, *E. coli* is more resistant to toxic compounds produced during the diluted acid hydrolysis of hemicellulose (alcohols, organic acids, and aldehydes) than other ethanologenic microorganisms [78].

By means of metabolic engineering techniques and metabolic evolution, diverse *E. coli* strains have been engineered for ethanol production. The first engineered *E. coli* strains were transformed with plasmids carrying the *PET* operon, containing the genes of the metabolic pathway for ethanol production from *Zymomonas mobilis*. The ethanol *Z. mobilis* pathway, which includes the genes that encode pyruvate decarboxylase (*pdc*) and alcohol dehydrogenase (*adhB*), allows the production of two ethanol molecules per consumed mol of glucose in a balanced redox reaction (Fig. 5). These engineered *E. coli* strains produced ethanol as the main fermentation product [79–81]; however, plasmid-based recombinants strains are generally less stable than strains in which the foreign genes have been integrated into the host chromosome.

One of the first breakthroughs in *E. coli* metabolic engineering for improving fermentation performance was the design and construction of the ethanologenic strain KO11 derived from *E. coli* W [82, 83]. KO11 was developed by Ingram and coworkers by integrating the *PET* operon into the *E. coli* chromosome under the control of the *pfl* promoter along with an antibiotic resistance marker (chloramphenicol, Cm). The *pflB* locus was chosen because the *pflB* gene is expressed at very high levels during anaerobic growth conditions in *E. coli* [9, 84], and the expression of genes contained in the *PET* operon was increased after several selections for resistance to high levels of Cm. This strain also has an interruption in the *frd* locus to prevent succinate production. Batch fermentation using *E. coli* KO11 in Luria Bertani (LB) broth with 100 g/L of glucose produced 54 g/L ethanol and 42 g/L when 80 g/L of xylose were used. Global volumetric productivity was close to 0.9 g/L h of ethanol and the ethanol yields were greater than 100%, exceeding the maximum theoretical yield (0.51 g ethanol/g sugar) because of the metabolism of nutrients from LB broth by KO11 [83].

4.1.2 Sugar Utilization

E. coli KO11 has been tested for batch ethanol production from diverse biomass residues. Sweet whey (58 g/L sugars) was used without supplements, reaching 38% of the theoretical ethanol yield. When the sweet whey was supplemented with yeast extract and a trace metals mixture, the ethanol yield increased to 100% [85]. Using corn fiber hydrolysates (90 g/L sugars) added with LB broth components or

mixture of sugars (100 g/L), ethanol yields were 80–88%, respectively, and global volumetric productivities of 0.38 and 0.66 g/L h ethanol were observed, without xylose being completely consumed [86]. Takahashi and coworkers [87] carried out the fermentation of sugar cane bagasse hydrolysate (68 g/L sugars) supplemented with tryptone and yeast extract, reaching an ethanol yield of 92% and a volumetric productivity of 0.66 g/L/h ethanol. Utilizing *Pinus sp*. hydrolysate (72 g/L), Barbosa and coworkers [88] produced ethanol with a yield of 85–91% of the theoretical yield (0.8 g/L/h ethanol). Like these examples, many other diverse agricultural residue hydrolysates had been fermented by *E. coli* KO11 for ethanol production [89, 90].

Another set of ethanologenic strains were developed using the *PET* operon in plasmids. These Fermentation Biochemistry Research (FBR) Unit strains are *E. coli* K-12 derivatives that carry mutations for lactate dehydrogenase and pyruvate formate lyase, and hence cannot grow under anaerobic conditions unless these mutations are complemented, for example, with pLOI297. This plasmid contains the *PET* operon under the control of the *lac* promoter from *E. coli* and the tetracycline and ampicillin resistance genes [79]. With this overexpression, the cells recovered their capacity to recycle $NADH^+$ to grow and produce ethanol as their primary fermentation product [76]. Strains transformed using FBR strains (FBR3, FBR4 and FBR5) produce ethanol from glucose, xylose, and arabinose using complex media, with yields of approximately 91% of the maximum theoretical yield, selectively maintaining pLOI297 after several transferences in batch cultures when grown anaerobically [91–93]. The FBR strains have also been used to produce ethanol from biomass hydrolysates supplemented with LB broth components, for example, strain FBR3 fermented corn fiber hydrolysates with sugar concentrations of 66 g/L and has a volumetric ethanol productivity of 0.71 g/L/h.

4.1.3 Second Generation of Metabolic Engineered Ethanologenic Strains

In all the reports described in the two previous paragraphs, it was necessary to supplement the culture medium with complex nutrients to achieve complete fermentation. Furthermore, when *E. coli* KO11 grows in minimal medium, the ethanol yield is lower than 70% of the theoretical maximum, and the ethanol volumetric productivity is half that obtained with LB medium [94]. This is a disadvantage, because complex nutrients are too expensive to be used in fuel ethanol industrial production [22]. Recently, an ethanologenic strain derived from the C lineage of *E. coli* has been reported (strain CCE14) to display a higher ethanol production rate in mineral medium. This is the result of elevated heterologous expression of the chromosomally integrated genes encoding Pdc and Adh from *Z. mobilis*, which results in an ethanol yield that is 90% of the theoretical maximum [32]. This strain displays a higher ethanol production rate, lower organic acid production rate and high glycolytic and ethanologenic fluxes that correlate with the enhanced transcription and enzymatic activity levels of Pdc and Adh. A re-engineered strain derived from *E. coli* KO11 was designated LY160. This strain has all the genes encoding

routes for NAD^+ production deleted, and the complete ethanol pathway of *Z. mobilis* (*pdc*, *adhA*, and *adhB*) was randomly inserted by transposon mutagenesis into its chromosome [95]. After selection in mineral medium supplemented with xylose, a homoethanologenic strain was isolated with genes that code for the ethanol pathway integrated within *rrlE,* a ribosomal gene highly expressed in *E. coli* at high and low growth rates. Performance of LY160 in NBS minimal medium with 1M betaine [96] and 90 g/L xylose shows that ethanol titer and yield are similar to that of *E. coli* KO11 in LB broth. An ethanol yield of 95% of the theoretical maximum and a volumetric productivity of 0.61 g/L/h ethanol was reported for LY160 [95].

Strain *E. coli* SZ420, derived from *E. coli* B, was engineered to produce ethanol from the native pathway [97]. Genes coding for fumarate reductase (*frdABCD*), lactate dehydrogenase (*ldhA*), acetate kinase (*ackA*), and pyruvate formate lyase (*pflB*) were eliminated. The *pflB* promoter was used to drive high anaerobic expression of the *pdh* operon (*aceEF-lpd*), and *aceE*, *aceF*, and *lpd* expression increased 24-, 18- and 30-fold, respectively, in reference to the parent strain. In addition, enzymatic activity of the Pdh complex was 33-fold higher than *E. coli* B under anaerobic conditions. Interestingly, the anaerobic Pdh activity of *E. coli* SZ420 was 64% higher than the aerobic Pdh activity of the parent strain. *E. coli* SZ420 metabolized glucose or xylose (50 g/L) using LB broth into ethanol (19 g/L) in 192 h. Nevertheless, it showed a lag phase, and xylose was not totally consumed after 192 h.

A novel ethanologenic strain that lacks heterologous genes was derived from *E. coli* K-12, obtained by chemical mutagenesis. This strain, denominated *E. coli* SE2378, is able to grow under anaerobic conditions due to a mutation in the pyruvate dehydrogenase (*pdh*) operon that activates its transcription under anaerobic conditions. The *pflB*, *adhE*, *ldhA*, and *aceF* genes have been deleted from the SE2378 strain. It can produce ethanol from glucose and xylose with a yield of 82% of the theoretical maximum [75]. Due to the *pdh* operon mutation, *E. coli* SE2378 can generate an additional NADH from pyruvate to acetyl-CoA and produce 2 mol of ethanol per mol of glucose under anaerobic conditions. This strain uses the native *E. coli* pathway, reducing acetyl-CoA to acetaldehyde and then to ethanol by alcohol dehydrogenase (AdhE) (Fig. 5). In LB medium added with glucose (50 g/L), *E. coli* SE2378 produced 22 g/L ethanol in 72 h with a maximum specific ethanol productivity (q_p) of 1.34 g/g cells/h, whereas the q_p with 50 g/L xylose was higher than glucose (2.24 g/g cells/h). SE2378, however, requires acetate and glutamate for growth in glucose-minimal medium.

4.1.4 Stability of Ethanologenic *E. coli* Strains

For commercial production of ethanol, batch and continuous culture fermentation are broadly utilized, the latter producing a greater quantity of ethanol. One disadvantage of the continuous culture is the generation of mutant strains during the culture time [98]. It is clear that, for successful long-term continuous fermentation, a high degree of stability of the recombinant strain is essential.

E. coli KO11 has demonstrated to be a good biocatalyst for ethanol production, with high yields from pentose and hexose sugars and biomass hydrolysates. When *E. coli* KO11 and other transformed strains with pLOI297 are transferred several times in batch culture [99] or cultivated in continuous cultures [100], however, these strains show instability with regard to ethanol production, resulting in a reduction in ethanol yield. This occurs in the absence of antibiotics in LB medium or when using pentoses as carbon sources. When *E. coli* KO11 grows in LB medium with glucose, the ethanol yield decreases to around 50% of the theoretical maximum in only 12 generations [99]. Using glucose- or xylose-limited continuous cultures at dilution rates (D) of 0.14 1/h and 0.07 1/h, respectively, these authors showed that *E. coli* B (pLOI297) and *E. coli* KO11 exhibited a rapid loss of ethanologenicity, with or without the presence of antibiotics in the feed medium. Even the ethanologenicity of *E. coli* KO11 was lost when the concentration of Cm was increased from 40 to 300 mg/L [101]. In another investigation, Dumsday and coworkers [100] evaluated the ethanologenic stability of *E. coli* KO11 in batch and continuous culture fermentations ($D = 0.06$ 1/h). They found that this strain was stable on glucose, mannose, xylose, and galactose (20 g/L) in batch cultures even in the absence of selective antibiotics (12 generations approximately). In continuous cultures only on glucose, the strain was stable. On mannose, xylose, and xylose–glucose mixtures, however, KO11 lost its ethanologenic capacity after 10 days. It has been suggested that the reduction or loss of ethanologenicity of *E. coli* KO11 in continuous culture is due to genetic instability [99] and loss of the heterologous ethanologenic genes [101, 102]. On the other hand, using *E. coli* FBR5, Martin and coworkers [102] carried out a continuous culture with LB broth containing xylose or glucose (50 g/L) without antibiotics at dilution rates of 0.045 1/h and 0.075 1/h, respectively. This strain maintained a stable ethanol yield of 80–85% of the theoretical maximum on both sugars over 26 days. In a continuous fluidized bed reactor with FBR5 the plasmid loss increased in free cells, whereas the amount of immobilized cells did not change. This strain has the drawback that under aerobic conditions and in the absence of antibiotics in the medium, the plasmid pLOI297 is lost.

4.1.5 Tolerance of Ethanologenic Strains

Since ethanol is a low-value product, its final concentration in the culture medium for commercial production must be the highest possible to minimize the cost of ethanol purification [77]. A higher resistance to ethanol toxicity is therefore a desired trait in the microorganisms employed for commercial production. It has been proposed that the main damage site for ethanol in *E. coli* is the cell membrane, causing damage to the peptidoglycan assembly in the growing cell wall and the associations between cross-linking enzymes and the cell membrane. Moreover, high concentrations of alcohols solubilize lipids and denature proteins, leading to membrane damage [103]. The toxicity of alcohols is directly related to their chain length and hydrophobicity, indicating a hydrophobic site of action. It has also been observed that elevated temperatures reduce ethanol tolerance [103]. An ethanol

tolerant strain derived from *E. coli* KO11 was obtained utilizing random mutation techniques and metabolic evolution. Strain LY01 was obtained when *E. coli* KO11 was cultivated in LB medium with 50 g/L glucose in the presence of exogenous ethanol. After sequential transfers performed while increasing the ethanol concentration (35–50 g/L), the cells were spread on solid medium containing Cm to select mutants, which retained efficient ethanol production and tolerance. These colonies were used to inoculate broth with ethanol for continuing selection [104]. After several transfers, among the selected strains, *E. coli* LY01 showed resistance to growth inhibition by 45 g/L ethanol. LY01 was able to produce over 60 g/L ethanol from 140 g/L xylose-LB (72 h, ethanol yield of 85%). This titer is the highest reported with ethanologenic *E. coli*. The survival ratio was 50% (CFU) when LY01 was exposed to 100 g/L ethanol over 0.5 min, whereas *E. coli* KO11 only survived less than 10% [104]. Ethanol tolerance in microorganisms correlates with the cell membrane lipid composition, including increases in the amount of unsaturated fatty acids (vaccenic acid), shifts in phospholipid composition, and changes in the phospholipids to protein ratios [105]. These changes, along with some factors such as the capacity of a strain to produce osmolites like trehalose, increase the cell's resistance to ethanol damage [106].

Some toxic molecules are generated during the thermochemical treatments of biomass: furane derivates, organic acids, aldehyde, alcohols, and aromatic compounds that inhibit the cell growth and ethanol production [106, 107]. It was found that, in many cases, the degree of toxicity is directly related to hydrophobicity [78]. The combinations of furfural with aromatic aldehydes and ethanol were synergistic in toxicity for growth, and furfural was the only aldehyde that caused a strong inhibition of ethanol production. *E. coli* KO11 and LY01, however, have a native ability to transform furfural to the less toxic furfuryl alcohol [108]. LY01 was more resistant than KO11 to furfural and 5-HMF, and the toxic effect of aldehydes on cells did not affect the membrane integrity [78]. Aliphatic acids appear to inhibit both growth and ethanol production by collapsing ion gradients and increasing the internal anion concentration [109]. Unlike organic acids and aldehydes, alcohols damaged the cellular membrane by leakage of magnesium from the cells and inhibiting the cell growth [103, 110]. The toxicity of hemicellulose hydrolysate results from a combination of compounds rather than from a single toxic compound [108]. Despite the multiple toxic effects caused by organic acids, aldehydes and alcohols on *E. coli* growth and ethanol production, ethanologenic *E. coli* strains have a greater resistance to these toxic effects compared to other ethanologenic strains, such as *S. cereviciae* and *Z. mobilis* [78, 107, 110].

4.2 Metabolic Engineering of E. coli *for* L*-Alanine Production*

L-Alanine is an important amino acid used in the field of food, cosmetic and pharmaceuticals [111], including its use in pre- and postoperative therapies [112]. L-Alanine is generally produced by enzymatic conversion of L-aspartic acid by

L-aspartate β-decarboxylase from *Pseudomonas dacunhae*. This method is very efficient and uses either immobilized or free cells, reaching conversion yields greater than 90% [113]. Nevertheless, fumaric acid used to obtain L-aspartic acid by enzymatic catalysis is a product derived from petroleum [112]. L-Alanine can also be produced with microorganisms. *Corynebacterium tumescens* or *Arthrobacter sp*. uses an L-alanine dehydrogenase (ALD) with the growth and amino acid production phases clearly separated to produce L-alanine. A D-alanine auxotrophic *E. coli* strain uses lactic acid and an amino donor to generate this amino acid [111, 114]. A recombinant strain of *Z. mobilis* expressing L-alanine dehydrogenase (*alaDh*) of *Bacillus sphaericus*, under thiamine limitation, has also been reported for L-alanine synthesis [115]. The alanine yields obtained from these microorganisms, however, are low [112].

In *E. coli*, at least three pathways are involved in the generation of L-alanine. One is through the enzyme valine-pyruvate aminotransferase (AvtA), utilizing valine as the amino donor and pyruvate as the amino acceptor [116]. In the latter, the enzyme cysteine desulfurase (IscS) catalyzes the transfer of sulfur and selenium from cysteine and selenocysteine to molecules like tRNA and to sulfur and selenium-dependent proteins, thereby generating L-alanine. Nevertheless, this pathway is not an important contributor to L-alanine synthesis [117]. Finally, the main and the most direct pathway for L-alanine synthesis is catalyzed by the enzyme glutamate-pyruvate aminotransferase (AlaB), which catalyzes the reversible amination of pyruvate to L-alanine with glutamate as the amino donor in a NADH dependent form. Thus, when alanine is produced through this pathway, the yield is improved by O_2 limitation [118] (Fig. 6). Furthermore, D-alanine is formed by racemization of L-alanine catalyzed by two enzymes, Alr and DadX, with anabolic and catabolic functions, respectively [116, 119].

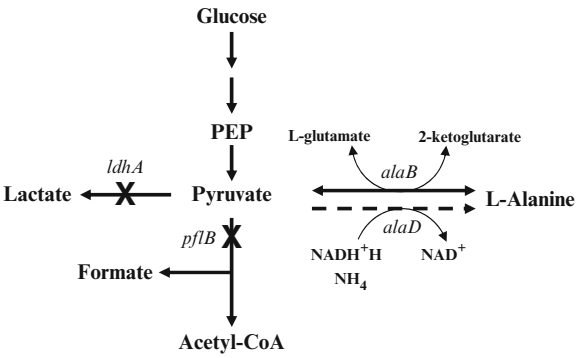


Fig. 6 Alanine pathway in wild type *E. coli* (*ala B*, *solid arrow*) and heterologous alanine dehydrogenase pathway (*alaD*, *dashed arrow*). Gene encoding enzymes are showed by *italics*, the *crossed out pathways* indicate deleted genes to avoid subproducts

4.2.1 Process Development for L-Alanine Production with Engineered *E. coli* Strains

Some L-alanine production strains have been engineered by cloning the *alaD* gene from *Arthrobacter sp*. or *Bacillus sphaericus* with diverse results. A decade ago, an *E. coli* strain was modified for DL-alanine production by cloning the gene coding for ALD from *Arthrobacter sp*. HAP1 under control of *lac* promoter in a plasmid. The titer reached was 8.1 g/L of DL-alanine in mineral medium (30 g/L glucose) using shake flasks under oxygen-limited conditions [111]. Lee and coworkers [120] developed an *E. coli* strain containing the *alaD* gene from *B. sphaericus* cloned into the vector pTrc99A. This strain also had mutations in *aceF* and *ldhA* genes to prevent acetyl-CoA and lactate formation from pyruvate. Using a two-phase process, first an aerobic phase (20% dissolved oxygen) and then a limited oxygen phase with complex medium and glucose-NH_4Cl feeding, these authors obtained an L-alanine yield based on glucose consumed of 0.69 g/g (maximum theoretical yield for alanine is 0.989 g/g glucose). A maximum volumetric rate of L-alanine production of 1.54 g/L/h was reported. Further work for improving culture conditions, i.e., the addition of glucose-NH_4Cl pulses during the production stage, resulted in a maximum volumetric rate of 2.0 g/L/h alanine. This yielded 32 g/L of L-alanine in 27 h. The overall yield on glucose was 0.63 g/g. Nevertheless, the addition of IPTG and antibiotics to the cultures makes this process impractical for commercial L-alanine production.

In another two-phase process, Smith and coworkers [121] produced DL-alanine from glucose in an *E. coli* mutant strain that over-expressed the *alaD* gene from *B. sphaericus* in plasmid pTrc99A (Fig. 6) and has deleted the genes coding for enzymes that compete for pyruvate to form products like formate (*pflB*), phosphoenolpyruvate (*pps*), acetate (*poxB*), lactate (*ldhA*), and acetyl-CoA (*aceEF*) [120]. The process consisted of cell growth under aerobic conditions using 20 g/L glucose in complex medium with antibiotics, followed by anaerobic conditions with the addition of glucose. This strategy allowed the production of 34 g/L DL-alanine in 13 h with a DL-alanine yield on glucose of 0.86 g/g and an overall volumetric productivity of 2.1 g/L/h in batch mode. To avoid excessive acetate formation during the aerobic phase, the same strain was cultured under fed-batch mode at $\mu = 0.15$ 1/h and then switched to a nongrowing anaerobic phase. During the anaerobic phase, 88 g/L of DL-alanine was produced at a high volumetric rate (4.0 g/L/h) and with a DL-alanine yield of 100% of the theoretical maximum.

An H^+-ATPase- and lactate dehydrogenase defective *E. coli* mutant expressing the alanine dehydrogenase gene (*adh*) from *Bacillus stearothermophilus* has also been used for DL-alanine production [122]. The mutation in the H^+-ATPase-operon increased the glycolytic flux, which stimulated pyruvate production [123]. The strain produced 20 g/L DL-alanine from 50 g/L glucose after 24 h of fermentation using complex medium with antibiotics under aerobic conditions. The DL-alanine yield on glucose was only 0.41 g/g due to the great amount of pyruvate that was produced (16 g/L), which indicates a limiting step is the conversion of pyruvate

to DL-alanine that is likely due to the aerobic conditions utilized during fermentation or favored by the high glycolytic flux of this strain.

Finally, the largest L-alanine concentration produced by *E. coli* has been reported by Zhang and coworkers [112], who utilized a derivative strain of *E. coli* W, called XZ132, to produce L-alanine. Strain XZ132 had deletions in the *frd*, *pflB*, *adhE*, *ackA*, *mgsA*, and *dadX* genes to prevent the production of succinate, formate, ethanol, acetate, lactate, and the conversion of L-alanine to D-alanine, respectively. In XZ132, the native *ldhA* gene of *E. coli* was replaced by the *alaD* gene, which codes for alanine dehydrogenase of *Geobacillus stearothermophilus* XL-65-6 (Fig. 6). The strain construction included driving *alaD* transcription from the native *ldhA* promoter for anaerobic L-alanine production. A metabolically engineered strain was evolved by serial transfers in mineral medium, with increasing glucose concentrations until XZ132 was selected. Strain XZ132 produces L-alanine with 99.5% of chiral purity as the main product during batch fermentation (48 h), reaching concentrations of 114 g/L from 120 g/L glucose in mineral medium [112]. These authors reported an L-alanine yield of 0.95 g/g glucose and a volumetric rate of 2.37 g/L/h. Xylose (50 g/L) was used for fermentation in the same mineral medium with XZ132. Results show that this strain produced 43 g/L of L-alanine with 99.5% of chiral purity within 72 h with an L-alanine yield of 0.85 g/g xylose. The work carried out with strain *E. coli* XZ132 has demonstrated that the generation of new strains for L-alanine production at high concentrations from glucose is possible without the use of plasmids, rich media or two-phase cultures (aerobic and anaerobic).

4.3 Pathway Engineering for Lactate Production

4.3.1 D-Lactate

L-Lactate is the most used isomer in the food industry, and it has been reported that it is the most used isomer in the emerging poly lactate (PLA) industry [124]. Homolactic fermentation in *E. coli* has been achieved by a number of different approaches. Gupta and Clark [125] reported that *adh* and *pta* double mutants regained the ability to grow anaerobically on hexoses by homolactic fermentation. *E. coli* is capable of equilibrating the redox balance when it grows on hexoses or pentoses by reducing pyruvate to lactate, but it is not able to grow on sorbitol or gluconate. This shows that is only possible to regenerate the redox potential generated by sugars but not by sugar alcohols or acid sugars. To enhance lactate flux, other pyruvate consuming pathways must be eliminated (Fig. 7). The expression of the *focA-pflB* operon is regulated by ArcA and Fnr and it is active under low or no oxygen conditions. The reaction catalyzed by pyruvate formate lyase (Pfl) is the major pyruvate-consuming pathway under anaerobic conditions. Thus, most of the strategies to achieve homolactic fermentation are designed to include elimination of the Pfl activity. Since it is the major acetyl-CoA formation reaction under

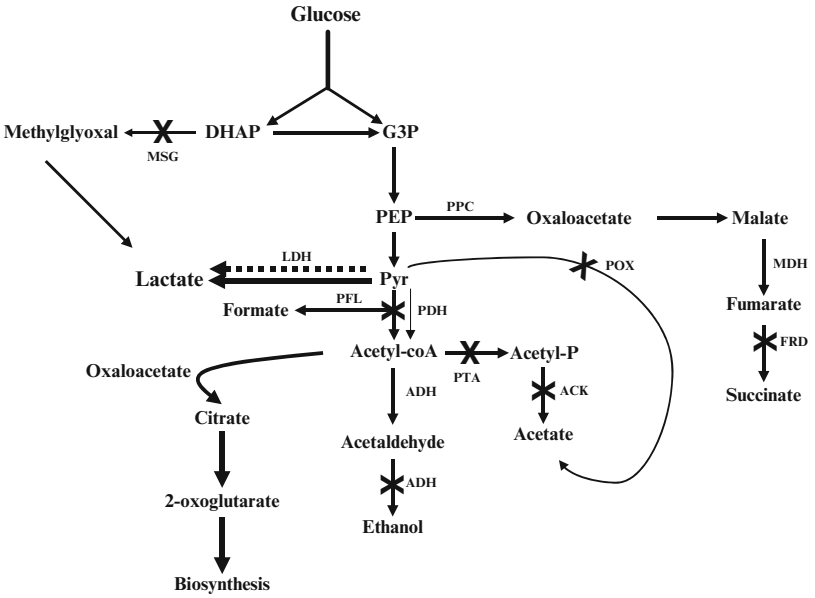


Fig. 7 Fermentation pathways of *Escherichia coli*: redirecting flux to lactate. *Solid lines* indicate homologous enzymes, *dotted lines* indicate heterologous enzymes. Abbreviations: *ACK* acetate kinase, *ADH* alcohol dehydrogenase, *DHAP* dihydroxyacetone phosphate, *FRD* fumarate reductase, *G3P* glyceraldehyde 3-phosphate, *LDH* lactate dehydrogenase, *MDH* malate dehydrogenase, *MSG* methylglyoxal synthase, *PDH* pyruvate dehydrogenase, *PEP* phosphoenol pyruvate, *PFL* pyruvate formate lyase, *POX* pyruvate oxidase, *PPC* phosphoenol pyruvate carboxylase, *PTA* phosphotransacetylase

anaerobic conditions, the *pfl* knockout strains show a low growth rate and low cell mass formation, and some of them are unable to grow on glucose under anaerobic conditions without acetate supplementation [126, 127].

Zhou et al. [126] developed an *E. coli* W3110 based biocatalysts with *pfl* and other deletions from genes that code for fermentative enzymes (Frd and Adh). Strains SZ40 (*focA-pflB*, *frdBC*) and SZ58 (*focA-pflB*, *frdBC, adhE*) were able to produce optically pure D-lactate in mineral media with a 98% yield. As previously mentioned in *pfl* mutants, however, the cell yield is drastically reduced, leading to a severe reduction in the volumetric productivity. The cell yield was improved by a further mutation in the acetate kinase gene (SZ63 strain), likely due to a reduction in the consumption of acetyl-CoA by the acetate pathway. Volumetric productivity was improved by an initial aeration phase or by adding 10 mM of acetate.

4.3.2 L-Lactate

Dien and coworkers [128] developed L-lactate producing strains by the heterologous expression of the lactate dehydrogenase from *Streptococcus bovis* in

nonfermenting *E. coli* strains (*Δpfl and Δldh*). Strains were evaluated for plasmid maintenance without antibiotic selection, the lactate pathway being the only NADH reduction reaction. Recombinant strains were able to produce lactic acid as the major fermentation product, with a 93% yield and productivities up to 2.33 g/L/h using complex medium. Using a *ptsG* knockout to abolish catabolite repression, simultaneous consumption of glucose and xylose was achieved in a *pfl*/*ldh*/*frd* knockout strain (FBR119). The heterologous expression of the *ldh* of *S. bovis* in FBR119 cultures in complex media supplemented with acetate allowed the lactate yield of 0.77% from 100 g/L of added sugars (50 g/L glucose + 50 g/L xylose) [129].

E. coli naturally produces the D-lactate isomer through the native lactate dehydrogenase (Ldh), an L-lactate producing SZ63 derivative generated by replacing the native *ldhA* gene with the chromosomal integration of the L-lactate dehydrogenase (*ldhL*) of *Pediococcus acidilactici* [130]. The resulting strain, SZ79, was able to produce L-lactate with a high yield but poor productivity. An adaptive evolution process led to a mutant with a higher growth rate. The SZ85 mutant is capable of growing faster and converting glucose or xylose with a high yield and productivity. The *ldhL* gene was sequenced, and the results showed that mutations arose in the upstream and coding region of the heterologous *ldhL* gene [130].

4.3.3 The Importance of the Energetic Yield

The metabolic effect of the knockout of *pfl* genes under microaerobic conditions was studied by Zhu and Zhimizu [127]. It was found that *pflA*$^-$ and *pflB*$^-$ strains produce large amounts of D-lactate. Enzyme activities and intracellular metabolite concentrations were measured. In the *pfl*$^-$ cells, ATP generation was reduced as a consequence of the lack of the acetate production pathway. It was shown, however, that the enzymatic activity of the acetate kinase (Ack) was highly increased. Thus, it is possible that the Ack-Pta pathway could play a role in producing acetyl-CoA and therefore consuming ATP. The lower ATP/AMP ratio promotes an increase in glycolytic flux [131], and since no increase in succinate production was found, even with a supply of CO_2, it was established that the energetic demand is favored over redox balance. A significantly higher NADH/NAD relationship was found in *pfl*$^-$ strains. Thus, homolactic fermentation is carried out to equilibrate the redox balance and to produce energy through glycolysis.

During anaerobic growth, ATP is derived from substrate level phosphorylation. The acetate forming reaction generates an extra mol of ATP. On xylose, 2 mol of ATP are used for transport and phosphorylation. Under anaerobic conditions, an ATP-dependent transport is used. In this case, the net yield of ATP of converting xylose to pyruvate or lactate is 0.67 mol of ATP/mol of xylose. Due to the low energy yield, *pfl* and *ack* mutants are unable to grow anaerobically on xylose. *pfl* or *ack* mutants, however, are able to grow on arabinose, because arabinose is transported into the cell by the use of a sugar/proton symporter (*araE*). Thus, 1 mol of ATP is conserved, which enables cell growth [132].

4.3.4 Adaptative Evolution and *In Silico* Design

Adaptive evolution is a strategy that has been widely used in the past few years to develop D- or L-lactate producing strains. Adaptive evolution has been useful to overcome the growth limitations resulting from *pfl* deletion. Zhou et al. [133] modified the ethanologenic *E. coli* strain KO11 for D-lactate production. Knocking out ethanologenic and cellobiose utilization genes along with *focA-pflB*, *adhE,* and *ackA*, followed by a metabolic evolution process in LB medium with 10% glucose, strain SZ186 was obtained. This SZ186 strain has native sucrose utilization genes, and no heterologous genes are expressed in it. Further improvement was obtained by adding 1 mM of betaine. It was found that this osmoprotector increased sugar and lactate tolerance. Betaine addition also increased cell yield and doubled specific productivity, thus substantially increasing the volumetric productivity [134]. SZ184 was further improved by another metabolic evolution round in mineral media with 10% glucose, leading to strain SZ194. This strain produces 1 M of D-lactate in 72 h using mineral salts medium supplemented with 1 mM of betaine in simple batch culture [135]. Further knocking out of the methylglyoxal synthase gene (*msgA*) in SZ194 allowed the elimination of chiral impurities (strain SZ195), and a derivative strain obtained by adaptive evolution (TG114) produces D-lactate with >99% optical purity. L-Lactate production strains derived from SZ194 were also obtained by integration from *ldhL* of *P. acidilactici* with an *msgA* deletion, followed by an adaptive evolution (strain TG108). The TG114 and TG108 strains are able to convert 12% glucose to 1.3 mol of D- and L-lactate, respectively, with a maximum productivity of 2.88 g/L/h and a 98% yield in batch fermentation [96, 136].

In silico design has been used to develop lactic acid producing strains. Fong et al. [137] used the OptKnock algorithm [138] to identify a multiple gene deletion combination that couples growth and the production of D-lactic acid. Three different designs were selected based on this algorithm: *pta-adhE, pta-pfk*, and the *pta-adhE-pfk-glk* knockout strains. The resulting knockout strains were subjected to adaptive evolution to improve growth rates. In all strains, growth rate increases occurred over the course of adaptive evolution, and results showed that the obtained phenotypes were consistent with the computationally determined solution spaces. Experiments were carried out in mineral medium, and the glucose concentration used was 2 g/L. Lactate titers ranged from 0.87 to 1.76 g/L and the secretion rates were directly coupled to the growth rates.

4.3.5 Dual Culture Strategies to Improve Lactate Productivity

Chang and coworkers [139] reported D- and L-lactic acid production with *pta ppc* mutants of *E. coli* in a two-phase process (aerobic and then anaerobic) using complex medium. A *pta ldhA* double mutant harboring the L-*ldh* gene from *Lactobacillus casei* produced optically pure L-lactate as the major fermentation product. Using this approach, volumetric productivities of 1.04 g/L/h were achieved [139].

Using an aerobic growth phase to an OD_{600} of 30 that was followed by anaerobic nongrowth production phase, Zhu and coworkers [140] produced 137 g/L of D-lactate in defined media using strain ALS947, an *aceEF, poxB, pfl, frdABCD E. coli* mutant. Since Pdh and Pfl activities are absent in ALS947, this strain shows auxotrophy for acetate and isoleucine. Succinate, as by product, was found to be forming from acetate feed as a product of the glyoxylate shunt of the tricarboxylic acid cycle. This approach showed the highest volumetric productivity (6.3 g/L/h) and titer reported so far.

4.4 *Recombinant* E. coli *Engineered for Production of Succinate*

4.4.1 Basic Mutations and Dual-Phase Cultures

NZN111 is a nonfermenting strain of *E. coli* that has *pfl* and *ldh* knocked out. A spontaneous mutant (AFP111) regained the ability to grow under anaerobic conditions due to succinate production. It was found that a mutation that inactivates *ptsG* was responsible for the phenotypic change. *ptsG* inactivation caused PEP to be available for carboxylation and further succinate conversion, such that NAD^+ is regenerated and glucose fermentation can occur [6]. A malic enzyme that is NADH-dependent was overexpressed from a multicopy plasmid, and the resulting strain was able to ferment glucose. Succinate was the major fermentation product [141].

Plasmid pTrc99A-*pyc* harboring the pyruvate carboxylase gene from *Rhizobium etli* was introduced to both the NZN111 and AFP111 strains, after which they were evaluated under anaerobic and dual-phase growth conditions. Results showed that AFP111/pTrc99A-*pyc* was best for succinate production, with a mass yield of 0.90 g of succinate per gram of glucose-rich media under dual phase conditions [142]. This strain was further characterized under dual-phase fermentation conditions, and different transition times from aerobic to anaerobic were evaluated. Results show that transition times have a great effect on succinate yield and productivity. Using the best transition time, an overall yield of 1.1 g/g, a titer of 99.2 g/L, and a productivity of 1.3 g/L/h were achieved in a glucose fed-batch rich media culture [143]. Due to enhanced anaplerotic activities, when the NZ111 strain was grown aerobically on acetate, it was able to convert glucose to succinate with a yield of 1.28 mol/mol and a productivity of 1.13 g/L/h in the anaerobic stage [144].

The *E. coli* strain AFP184 is a C600 derivative that lacks the genes coding for Pfl, Ldh, and PTS [145]. This was used to produce succinate in dual-phase fermentations in a medium containing 15 g/L of corn step liquor (CSL). The following different carbon sources and sugar mixtures were used: glucose, fructose, xylose, and sucrose. AFP 184 was able to convert all sugars and sugar combinations to succinate, except for sucrose. Results showed that higher yields were obtained from glucose (0.88 g/g) in the anaerobic phase, and fructose and xylose showed lower

yields (0.66 and 0.5 g/g). In the sugar mixtures, no catabolite repression was detected. The highest succinate titter (40 g/L) and overall productivity (1.27 g/L/h) were achieved with glucose as a carbon source [146].

4.4.2 Production Under Anaerobic Conditions and Cofactor Studies

Overexpression of PEP carboxylase (Ppc) and PEP carboxykinase were evaluated by Millard and coworkers [147]. It was found that overexpressing PPC resulted in a higher succinate production under anaerobic conditions for an *E. coli* strain that has no modifications in the fermentative pathways.

Lin and coworkers [148] evaluated the overexpression of pantothenate kinase (Pank) to increase the acetyl-CoA and CoA pools. Acetyl-CoA is an activator of Ppc and Pyc. In this work, Pank was overexpressed along with Ppc from *Sorghum vulgare* and Pyc from *Lactococcus lactis*. Results showed that coexpression of Pank with Ppc and Pank with Pyc increased succinate production compared to the individual overexpression of Ppc or Pyc. Lin and coworkers [21] evaluated the individual expression of Ppc and Pyc and coexpression of both enzymes to improve succinate production in *E. coli*. This coexpression was also applied to the *ldh pfl* double mutant of *E. coli*. The coexpression of Ppc and Pyc showed a higher succinate production when compared to the individual expression of each enzyme. Furthermore, it was found that the elimination of lactate and acetate pathways with the coexpression of Ppc and Pyc was the most effective strategy to redirect the carbon flux to succinate (Fig. 8). The effect of different carbon sources, glucose, xylose, and sorbitol, was evaluated for succinic acid production in different host strains with *Sorgum* Ppc overexpression. Xylose showed a better succinate yield than glucose, which was likely due to a higher PEP availability. Sorbitol, a more reduced substrate, showed higher succinate production but also showed higher ethanol production [21].

Strain SBS110MG is a *ldh adhE* double mutant of *E. coli* that was evaluated for succinate production with the expression of a pyruvate carboxylase (Pyc) from *Lactococcus lactis*. This strain is able to produce succinate from glucose with a molar yield of 1.3 mol/mol. The results showed that, by the expression of Pyc, succinate was the main NAD^+ regeneration pathway. Thus, glucose consumption was improved fourfold [149].

A novel pathway was designed by Sanchez and coworkers [150] using a glyoxylate shunt to increase the succinate theoretical yield by reducing NADH requirements. Strain SBS550MG was constructed by inactivation of competing pathways (*ldh, adhE, ackA*) and isocitrate lyase repressor (*iclR*). Simultaneous overexpression of *pyc* from *L. lactis* and an NADH-insensitive citrate synthase (*citZ*) from *B. subtilis* were evaluated. The resulting strain was able to convert glucose into succinate with a molar yield of 1.6. In a complex medium and with fed-batch techniques, a succinate titer of about 40 g/L and a productivity of 1.18 g/L/h were achieved.

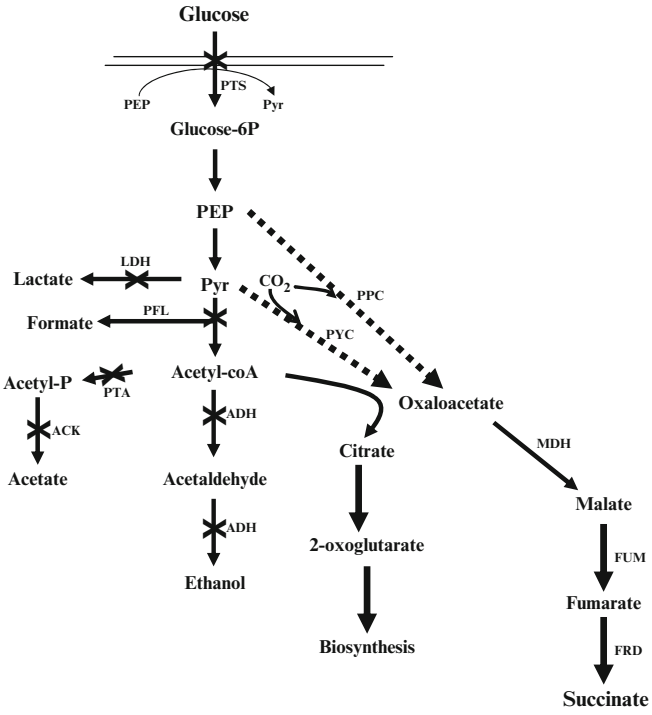


Fig. 8 Fermentation pathways of *Escherichia coli*: redirecting flux to succinate. *Solid lines* indicate homologous enzymes, *dotted lines* indicate overexpressed and heterologous enzymes. Abbreviations: *ACK* acetate kinase, *ADH* alcohol dehydrogenase, *FRD* fumarate reductase, *FUM* fumarase, *LDH* lactate dehydrogenase, *MDH* malate dehydrogenase, *PEP* phosphoenolpyruvate, *PFL* pyruvate formate lyase, *PPC* phosphoenol pyruvate carboxylase, *PTS* phosphotransferase system, *PYC* pyruvate carboxylase, *Pyr* pyruvate

4.4.3 *In Silico* Design and Adaptative Evolution

Based on metabolic pathway comparison between *E. coli* and *Mannheimia succiniciproducens, E. coli* strains were engineered for succinate production. Five genes or operons, *ptsG*, *pykF*, *sdhA*, *mqo*, and *aceBAK*, were eliminated. The resulting strains did not increase succinate production. An in silico metabolic analysis based on linear programming was used to design a succinate producer. Pyruvate forming enzymes *ptsG*, *pykF*, and *pykA* were disrupted to generate the W3110GFA strain, which produces succinate as the major fermentation product in LB-glucose under anaerobic conditions. Additional deletions were done for the *pfl* and *ldhA* genes. The *pfl* deletion resulted in a higher succinate molar ratio but in less titer and slower growth. The *ldhA* deletion resulted in a strain with marginal growth under anaerobic conditions [151].

E. coli C derivatives were engineered for succinate and malate production by deleting central anaerobic genes (*adhE*, *ldhA*, *and ackA*), followed by growth-based

selection over 2,000 generations. In the resulting strains, succinate and malate remained as the primary route for NAD^+ regeneration. Growth selected mutants produced higher levels of other organic acids. Further improvements were achieved by the deletion of the *pflB*, *poxB*, *and msgA* genes. The best succinate biocatalyst produced up to 622 mM of succinate, with a molar conversion yield of 1.6 per mole of glucose metabolized [152]. All FRT sites were removed from the KJ073 strain to create KJ091 (*ΔldhA*, *ΔadhE*, *ΔackA*, *ΔfocA-pflB*, *ΔmgsA*, *and ΔpoxB*), which lacks foreign DNA. The KJ091 strain was further improved by deleting threonine decarboxylase (*tdcD*; acetate kinase homologue) and 2-ketobutyrate formate-lyase (*tdcE*; pyruvate formate lyase homologue). These deletions reduced the acetate level by 50% and increased succinate yield (1.3 mol/mol glucose). In addition, by removing aspartate aminotransferase (*aspC*) and the NAD^+ linked malic enzyme (*sfcA*), the succinate molar yield was increased to 1.5, the succinate titer increased to 700 mM, and the average volumetric productivity increased to 0.9 g/L/h in simple batch fermentation in mineral media. Residual pyruvate and acetate were reduced by further deletion of *pta* to produce strain KJ134 (*ΔldhA, ΔadhE, ΔfocA-pflB, ΔmgsA, ΔpoxB, ΔtdcDE, ΔcitF, ΔaspC, ΔsfcA*, *and Δpta-ackA*) [153].

5 Conclusions

This chapter shows that basic knowledge of fermentative metabolism, coupled with metabolic engineering techniques, constitute powerful tools to understand and explore the fundamental cellular capacity of *E. coli* and to exploit its capacity to be applied as an industrial biocatalyst to produce a wide array of commodity chemicals. It is possible to employ *E. coli* with a wide array of organic compounds as terminal electron acceptors for fermentative metabolism, such as hexose sugars (e.g., glucose, mannose, galactose, fructose, and others), pentoses (e.g., xylose, arabinose, ribose, xylulose, and others), sugar acids (e.g., glucuronate and galacturonate), some uronic acids, and, although not discussed in this chapter, *E. coli* also has been engineered for disaccharide catabolism.

Under oxygen-deprived conditions, different substrates are converted into biochemical products that represent valuable molecules to society. A wide variety of metabolic engineering strategies have been employed to construct homofermentative strains, displaying high yield and productivity for the target product. With different degrees of success, high yield, and productivity, homofermentative strains had been developed, such as those described in this chapter for ethanol, lactate, alanine, and succinate. These studies provide the basis for the implementation of appropriate genetic modifications to increase further the product yield, such as disrupting pathways that compete for metabolite production. This is not restrictive, however, as *E. coli* fermentative metabolism can also be engineered to obtain other potential products with high titers, such as formate, hydrogen, acetate, acetone, propionate, propanol, propanodiol, butyrate, butanol, bioelectricity, and others.

Characterization and process development with metabolically engineered strains have shown that it is possible to obtain homofermentative strains with titters reaching values above 1 M. Volumetric productivity is a function of cell density. Some reports, therefore, focus only on the construction of a robust biocatalyst with high specific productivities to be employed in cultures under anaerobic conditions or at low cell densities, but channeling most of the carbon to product formation. Two-stage culture strategies that are based on the development of high quantities of biomass under aerobic conditions, followed by a switch to anaerobic conditions and complemented with feed batch cultures, however, have been used to attain higher increases in volumetric productivity. Furthermore, when strain design by intuitive, knowledge-based, or synthetic biological methods does not allow for high yield-productivity strains, adaptive evolution has become a very practical tool for strain improvement and gives the ability to understand how *E. coli* adapts and evolves under different conditions.

Thus far, it has been demonstrated that the enzyme activity of central carbon metabolism pathways under fermentative conditions are sufficient to contend with increases in the rates of sugar consumption and product rate formation. Redox balance remains a key factor, and glycolytic flux is still a complex physiological parameter that appears widely controlled by ATP demand. It also remains to be seen how many of the strains developed by the metabolic engineering of fermentative metabolism and derived processes reach industrial scale as oil reserves decline, petroleum increases in price, and more eco-friendly technologies are required for chemical production.

Acknowledgements Support from grants UNAM-PAPIIT-DGAPA: IN220908 and CONACyT′ – Estado de Morelos MOR-2007-COL-80360 is acknowledged.

References

1. Perrenoud A, Sauer U (2005) Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*. J Bacteriol 187:3171–3179
2. Fuhrer T, Fischer E, Sauer U (2005) Experimental identification and quantification of glucose metabolism in seven bacterial species. J Bacteriol 187:1581–1590
3. Gottschalk G (1986) Bacterial metabolism, 2nd edn. Springer, New York, p 237
4. Sakar D, Shimizu K (2008) Effect of *cra* gene knockout together with other genes knockouts on the improvement of substrate consumption rate in *Escherichia coli* under microaerobic condition. Biochem Eng J 42:224–228
5. Lin ECC, Iuchi S (1991) Regulation of gene expression in fermentative and respiratory systems is *Escherichia coli* and related bacteria. Annu Rev Genet 25:361–387
6. Chatterjee R, Millard CS, Champion K et al (2001) Mutation of the *ptsG* gene results in increased production of succinate in fermentation of glucose by *Escherichia coli*. Appl Environ Microbiol 67:148–154
7. Clark DP (1989) The fermentation pathways of *Escherichia coli*. FEMS Microbiol Rev 63:223–234

8. Böck A, Sawers G (1996) Fermentation. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella*. Cellular and molecular biology, 2nd edn. ASM, Washington DC, USA
9. Sawers G, Bock A (1988) Anaerobic regulation of pyruvate formate-lyase from *Escherichia coli* K-12. J Bacteriol 170:5330–5336
10. Postma PW, Lengeler JW, Jacobson GR (1996) Phosphoenolpyruvate: carbohydrate phosphotransferase systems. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella*. Cellular and molecular biology, 2nd edn. ASM, Washington DC, USA
11. Saier MH (2000) Vectorial metabolism and the evolution of transport systems. J Bacteriol 182:5029–5035
12. Tchieu JH, Norris V, Edwards JS et al (2001) The complete phosphotranferase system in *Escherichia coli*. J Mol Microbiol Biotechnol 3:329–346
13. Misset O, Blaauw M, Postma PW et al (1983) Bacterial phosphoenolpyruvate-dependent phosphotransferase system. Mechanism of the transmembrane sugar translocation and phosphorylation. Biochemistry 22:6163–6170
14. Stock JB, Waygood EB, Meadow ND et al (1982) Sugar transport by the bacterial phosphotransferase system. The glucose receptors of the *Salmonella typhimurium* phosphotransferase system. J Biol Chem 257:14543–14552
15. Holms WH (1986) The central metabolic pathway of *Escherichia coli*: relationship between flux and control at a branch point, efficiency of conversion to biomass, and excretion of acetate. In: Horecker BL, Stadtman ER (eds) Current topics in cell regulation. Academic, New York
16. Flores N, Flores S, Escalante A et al (2005) Adaptation for fast growth on glucose by differential expression of central carbon metabolism and *gal* regulon genes in an *Escherichia coli* strain lacking the phosphoenolpyruvate: carbohydrate phosphotransferase system. Metab Eng 7:70–87
17. Flores N, Yong-Xiao J, Berry A et al (1996) Pathway engineering for the production of aromatic compounds in *Escherichia coli*. Nat Biotechnol 14:620–623
18. Flores S, Gosset G, Flores N et al (2002) Analysis of carbon metabolism in *Escherichia coli* strains with an inactive phosphotransferase system by ^{13}C labeling and NMR spectroscopy. Metab Eng 4:124–137
19. Hernández-Montalvo V, Martinez A, Hernández-Chávez G et al (2003) Expression of *galP* and *glk* in a *Escherichia coli* PTS mutant restores glucose transport and increases glycolytic flux to fermentation products. Biotechnol Bioeng 83:687–694
20. Lin H, Bennett GN, San KY (2005) Effect of carbon sources differing in oxidation state and transport route on succinate production in metabolically engineered *Escherichia coli*. J Ind Microbiol Biotechnol 32:87–93
21. Lin H, San KY, Bennett GN (2005) Effect of *Sorghum vulgare* phosphoenolpyruvate carboxylase and *Lactococcus lactis* pyruvate carboxylase coexpression on succinate production in mutant strains of *Escherichia coli*. Appl Microbiol Biotechnol 67:515–523
22. Martinez A, York SW, Yomano LP et al (1999) Biosynthetic burden and plasmid burden limit expression of chromosomally integrated heterologous genes (*pdc*, *adhB*) in *Escherichia coli*. Biotechnol Prog 15:891–897
23. Gunsalus R (1992) Control of electron flow in *Escherichia coli*: coordinated transcription of respiratory pathway genes. J Bacteriol 174:7069–7074
24. Shalel-Levanon S, San KY, Bennet GN (2005) Effect of ArcA and FNR on the expression of genes related to the oxygen regulation and the glycolysis pathway in *Escherichia coli* under microaerobic growth conditions. Biotechnol Bioeng 92:147–159
25. Guest JR, Green J, Irvine AS et al (1996) The FNR modulon and FNR-regulated gene expression. In: Lin ECC, Lynch AS (eds) Regulation of gene expression in *Escherichia coli*. Chapman & Hall, New York
26. Lynch AS, Lin ECC (1996) Regulation of aerobic and anaerobic metabolism by the Arc system. In: Lin ECC, Lynch AS (eds) Regulation of gene expression in *Escherichia coli*. Chapman & Hall, New York

27. Park SJ, Gunsalus RP (1995) Oxygen, iron, carbon, and superoxide control of the fumarase *fumA* and *fumC* genes of *Escherichia coli*: Role or *arcA*, *fnr* and *soxR* gene products. J Bacteriol 177:6255–6262
28. Gutierrez-Ríos RM, Freyre-Gonzalez JA, Resendis O et al (2007) Identification of regulatory network topological units coordinating the genome-wide transcriptional response to glucose in *Escherichia coli*. BMC Microbiol 7:53
29. Báez-Viveros J, Flores N, Juárez K et al (2007) Metabolic transcription analysis of engineered *Escherichia coli* strains that overproduce L-phenylalanine. Microbial Cell Fact 6:30
30. Franchini AG, Egli T (2006) Global gene expression in *Escherichia coli* K-12 during short-term and long term adaptation to glucose-limited continuous culture conditions. Microbiology 152:2111–2127
31. Gonzalez R, Tao H, Shanmugam KT et al (2002) Global gene expression differences associated with changes in glycolytic flux and growth rate in *Escherichia coli* during the fermentation of glucose and xylose. Biotechnol Prog 18:6–20
32. Orencio-Trejo M, Flores N, Escalante A et al (2008) Metabolic regulation analysis of an ethanologenic *Escherichia coli* strain based on RT-PCR and enzymatic activities. Biotech Biofuels 1:8
33. Zhu J, Shimizu K (2005) Effect of a single-gene knockout on the metabolic regulation in *Escherichia coli* for D-lactate production under microaerobic condition. Metab Eng 7:104–115
34. Saier MH (1996) Cyclic AMP-independent catabolite repression in bacteria. FEMS Microbiol Lett 138:97–103
35. Ow DSW, Lee RMY, Nisson PM et al (2007) Inactivating FruR global regulator in plasmid-bearing *Escherichia coli* alters metabolic gene expression and improves growth rate. J Biotechnol 131:261–269
36. Saier MH, Ramseier T (1996) The catabolite repressor/activator (Cra) protein of enteric bacteria. J Bacteriol 178:3411–3417
37. Crasnier-Mednansky M, Park MC, Studley WK et al (1997) Cra-mediated regulation of *Escherichia coli* adenylate cyclase. Microbiology 143:785–792
38. Chin AM, Feldheim DA, Saier MH (1989) Altered transcription patterns affecting several metabolic pathways is strains of *Salmonella typhimurium* which overexpress the fructose regulon. J Bacteriol 171:2424–2434
39. Ramseier TM, Bleding S, Michotey V et al (1995) The global regulatory protein FruR modulates the direction of carbon flow in *Escherichia coli*. Mol Microbiol 16:1157–1169
40. Saier MH, Ramseier TM, Reizer J (1996) Regulation of carbon utilization. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella*. Cellular and molecular biology, 2nd edn. ASM, Washington DC, USA
41. Arora KK, Pedersen PL (1995) Glucokinase of *Escherichia coli*: induction in response to the stress of overexpressing foreign proteins. Arch Biochem Biophys 319:574–578
42. Charpentier B, Branlant C (1994) The *Escherichia coli gapA* gene is transcribed by the vegetative RNA polymerase holoenzyme $E\sigma^{70}$ and the heat shock RNA polymerase $E\sigma^{32}$. J Bacteriol 176:830–839
43. Fenton AW, Reinhart GD (2002) Isolation of a single activating allostering interaction in phosphofructokinase from *Escherichia coli*. Biochemistry 41:13410–13416
44. Blangly H, Buc H, Monod J (1968) Kinetics of the allosteric interactions of phosphofructokinase from *Escherichia coli*. J Mol Biol 31:13–35
45. Fenton AW, Parichartanakul NM (2003) Identification of substrate contact residues important for the allosteric regulation of phosphofructokinase from *Escherichia coli*. Biochemistry 42:6453–6459
46. Fraenkel DG (1996) Glycolysis. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella*. Cellular and molecular biology, 2nd edn. ASM, Washington DC, USA

47. Ponce E, Flores N, Martinez A et al (1995) Cloning of the two pyruvate kinase isoenzymes structural genes from *Escherichia coli*: the relative roles of these enzymes in pyruvate biosynthesis. J Bacteriol 177:5719–5722
48. Peng MJ, Arauzo-Bravo SK (2004) Metabolic flux analysis for a *ppc* mutant *Escherichia coli* based on ^{13}C-labelling experiments together with enzyme activity assays and intracellular metabolite measurements. FEMS Microbiol Lett 235:17–23
49. Yang C, Hua Q, Baba T et al (2003) Analysis of *Escherichia coli* anaplerotic metabolism and its regulation mechanism from the metabolic responses to altered dilution rates and phosphoenolpyruvate carboxykinase knockout. Biotechnol Bioeng 84:129–144
50. Ogawa T, Mori H, Tomita MY et al (2007) Inhibitory effect of phosphoenolpyruvate on glycolytic enzymes in *Escherichia coli*. Res Microbiol 158:159–163
51. De Graef MR, Alexeeva S, Snoep JL et al (1999) The steady-state internal redox state (NADH/NAD) reflects the external redox state and is correlated with catabolic adaptation in *Escherichia coli*. J Bacteriol 181:2351–2357
52. Garrigues C, Loubiere P, Lindley ND et al (1997) Control of the shift from homolactic acid to mixed-acid fermentation in *Lactococcus lactis*: predominant role of the NADH/NAD$^+$ ratio. J Bacteriol 179:5282–5287
53. Koebmann BJ, Westerhoff HV, Snoep JL et al (2002) The glycolytic flux in *Escherichia coli* is controlled by the demand for ATP. J Bacteriol 184:3909–3916
54. Vemuri GN, Eiteman MA, Altman E (2005) Increased recombinant protein production in *Escherichia coli* strains with overexpressed water-forming NADH oxidase and a deleted ArcA regulatory protein. Biotechnol Bioeng 94:538–542
55. Wang Y, Wu SL, Hancock WS et al (2005) Proteomic profiling of *Escherichia coli* proteins under high cell density fed-batch cultivation with overexpression of phosphogluconolactonase. Biotechnol Prog 21:1401–1411
56. Fraenkel DG, Vinopal RT (1973) Carbohydrate metabolism in bacteria. Annu Rev Microbiol 27:69–100
57. Sprenger GA (1995) Genetics of pentose-phosphate pathway enzymes of *Escherichia coli* K-12. Arch Microbiol 164:324–330
58. Zhang R, Andersson CE, Savchenko A et al (2003) Structure of *Escherichia coli* ribose-5-phosphate isomerase: a ubiquitous enzyme of the pentose phosphate pathway and the Calvin cycle. Structure 11:31–42
59. Fraenkel DG (1987) Glycolysis, pentose phosphate pathway, and Entner-Doudoroff pathway. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella*. Cellular and molecular biology. ASM, Washington DC, USA
60. Lin ECC (1996) Dissimilatory pathways for sugar, poliols and carboxylates. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella*. Cellular and molecular biology, 2nd edn. ASM, Washington DC, USA
61. Macpherson AJS, Jones-Mortimer MC, Henderson PJF (1981) Identification of area transport protein of *Escherichia coli*. Biochem J 196:269–283
62. Scripture JB, Hogg RW (1983) The nucleotides sequences defining the signal peptides of the galactose-binding protein and the arabinose-binding protein. J Mol Biol 258: 10853–10855
63. Kosiba BE, Schleif R (1982) Arabinose-inducible from *Escherichia coli*: its cloning from chromosomal DNA, identification as the *araFG* promoter and sequence. J Mol Biol 156:53–66
64. Kolodrubetz D, Schleif R (1981) Regulation of the L-arabinose transport operons in *Escherichia coli*. J Mol Biol 151:215–227
65. Ogden S, Haggerty D, Stoner CM et al (1980) The *Escherichia coli* L-arabinose operon: binding sites of the regulatory proteins and a mechanism of positive and negative regulation. PNAS 77:3346–3350
66. Ratushny AV, Smirnova OG, Usuda Y et al. (2006) Regulation of the pentose phosphate pathway in *Escherichia coli*: gene network reconstruction and mathematical modeling of

metabolic reaction. The fourth international conferences of bioinformatics of genome and structure 2006
67. Bausch C, Peekhaus N, Utz C et al (1998) Sequence analysis of the GntII system for gluconate metabolism reveals a novel pathway for L-idonic acid catabolism in *Escherichia coli*. J Bacteriol 180:3704–3710
68. Fuhrman LK, Wanken A, Nickerson KW et al (1998) Rapid accumulation of intracellular 2-keto-3-deoxy-6-phosphogluconate in an Entner-Doudoroff aldolase mutant results in bacteriostasis. FEMS Microbiol Lett 159:261–266
69. Porco A, Alonso G, Istúriz T (1998) The gluconate high affinity transport of GntI in *Escherichia coli* involves a multicomponent complex system. J Basic Microbiol 38:395–404
70. Tong S, Porco A, Istüriz T, Conway T (1996) Cloning and molecular genetic characterization of the *Escherichia coli gntR, gntK* and *gntU* genes of GntI, the main system for gluconate metabolism. J Bacteriol 178:3260–3269
71. Bachi B, Kornberg HL (1975) Genes involved in the uptake and catabolism of gluconate by *Escherichia coli*. J Gen Microbiol 90:321–335
72. Istúriz T, Palmero E, Vitelli-Flores J (1986) Mutations affecting gluconate catabolism in *Escherichia coli*. Genetic mapping of the locus for the thermosensitive gluconokinase. J Gen Microbiol 132:3209–3219
73. Peekhaus N, Conway T (1998) What's for dinner? Entner-Duodoroff metabolism. J Bacteriol 180:3495–3502
74. O'Neill MC (1989) *Escherichia coli* promoters I. Consensus as it relates to spacing class, specificity, repeat substructure, and three-dimensional organization. J Biol Chem 264:5531–5534
75. Kim YK, Ingram LO, Shanmugam KT (2007) Construction of an *Escherichia coli* K-12 mutant for homoethanologenic fermentation of glucose or xylose without foreign genes. Appl Environ Microbiol 73:1766–1771
76. Dien BS, Cotta MA, Jeffries TW (2003) Bacteria engineered for fuel ethanol production: current status. Appl Microbiol Biotechnol 63:258–266
77. Galbe M, Zacchi G (2002) A review of the production of ethanol from softwood. Appl Microbiol Biotechnol 59:618–628
78. Zaldivar J, Martínez A, Ingram LO (1999) Effect of selected aldehydes on the growth and fermentation of ethanologenic *Escherichia coli*. Biotechnol Bioeng 65:24–33
79. Alterthum F, Ingram LO (1989) Efficient ethanol production from glucose, lactose, and xylose by recombinant *Escherichia coli*. Appl Environ Microbiol 55:1943–1948
80. Ingram LO, Conway T (1988) Expression of different levels of ethanologenic enzymes from *Zymomonas mobilis* in recombinant strains of *Escherichia coli*. Appl Environ Microbiol 54:397–404
81. Neale AD, Scopes RK, Kelly JM (1988) Alcohol production from glucose and xylose using *Escherichia coli* containing *Zymomonas mobilis* genes. Appl Microbiol Biotechnol 29:162–167
82. Jarboe LR, Grabar TB, Yomano LP et al (2007) Development of ethanologenic bacteria. Adv Biochem Eng/Biotechnol 108:237–261
83. Ohta K, Beall DS, Mejia JP et al (1991) Genetic improvement of *Escherichia coli* for ethanol production: chromosomal integration of *Zymomonas mobilis* genes encoding pyruvate decarboxylase and alcohol dehydrogenase II. Appl Environ Microbiol 57:893–900
84. Sawers G, Bock A (1989) Novel transcriptional control of the pyruvate formate-lyase gene: upstream regulatory sequences and multiple promoters regulate anaerobic expression. J Bacteriol 171:2485–2498
85. Leite AR, Guimaraes WV, Fernandes de Araújo E et al (2000) Fermentation of sweet whey by recombinant *Escherichia coli* KO11. Brazilian J Microbiol 31:212–215
86. Dien BS, Hespell RB, Ingram LO et al (1997) Conversion of corn milling fibrous co-products into ethanol by recombinant *Escherichia coli* strains KO11 and SL40. World J Microbiol Biotech 13:619–625

87. Takahashi CM, de Carvalho Lima KG, Takahashi DF et al (2000) Fermentation of sugar cane bagasse hemicellulosic hydrolysate and sugar mixtures to ethanol by recombinant *Escherichia coli* KO11. World J Microbiol Biotechnol 16:829–834
88. Barbosa MF, Beck MJ, Fein JE, Potts D et al (1992) Efficient fermentation of *Pinus sp*. acid hydrolysates by an ethanologenic strain of *Escherichia coli*. Appl Environ Microbiol 58:1382–1384
89. Asghari A, Bothast RJ, Doran JB et al (1996) Ethanol production from hemicellulose hydrolysates of agricultural residues using genetically engineered *Escherichia coli* strain KO11. J Ind Microbiol 16:42–47
90. Doran Peterson J, Ingram LO (2008) Anaerobic respiration in engineered *Escherichia coli* with an internal electron acceptor to produce fuel ethanol. Ann NY Acad Sci 1125:363–372
91. Dien BS, Hespell RB, Wyckoff H et al (1998) Fermentation of hexose and pentose sugars using a novel ethanologenic *Escherichia coli* strain 1. Enz Microbial Technol 23:366–371
92. Dien BS, Iten LB, Bothast RJ (1999) Conversion of corn fiber to ethanol by recombinant *E. coli* strain FBR3. J Ind Microbiol Biotechnol 22:575–581
93. Dien BS, Nichols NN, O'bryan PJ et al (2000) Development of new ethanologenic *Escherichia coli* strains for fermentation of lignocellulosic biomass. Appl Biochem Biotechnol 84–86:181–196
94. Huerta-Beristain G, Utrilla J, Hernández-Chavez G et al (2008) Specific ethanol production rate in ethanologenic *Escherichia coli* strain KO11 is limited by pyruvate decarboxylase. J Mol Microbiol Biotechnol 15:55–64
95. Yomano LP, York SW, Zhou S et al (2008) Re-engineering *Escherichia coli* for ethanol production. Biotech Lett 30:2097–2103
96. Martínez A, Grabar TB, Shanmugam KT et al (2007) Low salt medium for lactate and ethanol production by recombinant *Escherichia coli* B. Biotechnol Lett 29:397–404
97. Zhou S, Iverson AG, Grayburn WS (2008) Engineering a native homoethanol pathway in *Escherichia coli* B for ethanol production. Biotechnol Lett 30:335–342
98. Bayrock DP, Ingledew WM (2005) Ethanol production in multistage continuous, single stage continuous, *Lactobacillus*-contaminated continuous, and batch fermentations. World J Microbiol Biotechnol 21:83–88
99. Lawford HG, Rousseau JD (1995) Loss of ethanologenicity in *Escherichia coli* B recombinants pLOI297 and KO11 during growth in the absence of antibiotics. Biotech Lett 17:751–756
100. Dumsday GJ, Zhou B, Yaquin W et al (1999) Comparative stability of ethanol production by *Escherichia coli* KO11 in batch and chemostat culture. J Ind Microbiol Biotech 23:701–708
101. Lawford HG, Rousseau JD (1996) Factors contributing to the loss of ethanologenicity of *Escherichia coli* B recombinants pLOI297 and KO11. Appl Biochem Biotechnol 57 (58):293–305
102. Martin GJO, Knepper A, Zhou B et al (2006) Performance and stability of ethanologenic *Escherichia coli* strain FBR5 during continuous culture on xylose and glucose. J Ind Microbiol Biotechnol 33:834–844
103. Ingram LO, Buttke T (1984) Effects of alcohols on micro-organisms. Adv Microb Physiol 25:253–300
104. Yomano LP, York SW, Ingram LO (1998) Isolation and characterization of ethanol-tolerant mutants of *Escherichia coli* KO11 for fuel ethanol production. J Ind Microbiol Biotech 20:132–138
105. Dombek KM, Ingram LO (1984) Effects of ethanol on the *Escherichia coli* plasma membrane. J Bacteriol 157:233–239
106. Zaldivar J, Nielsen J, Olsson L (2001) Fuel ethanol production from lignocellulose: a challenge for metabolic engineering and process integration. Appl Microbiol Biotech 56:17–34

107. Klinke HB, Thomsen AB, Ahring BK (2004) Inhibition of ethanol-producing yeast and bacteria by degradation products produced during pre-treatment of biomass. Appl Microbiol Biotechnol 66:10–26
108. Zaldivar J, Martínez A, Ingram LO (2000) Effect of alcohol compounds found in hemicellulose hydrolysate on the growth and fermentation of ethanologenic *Escherichia coli*. Biotechnol Bioeng 68:524–530
109. Peterson JD, Ingram LO (2008) Anaerobic respiration in engineered *Escherichia coli* with an internal electron acceptor to produce fuel ethanol. Ann NY Acad Sci 1125:363–372
110. Zaldivar J, Ingram LO (1999) Effect of organic acids on the growth and fermentation of ethanologenic *Escherichia coli* LY01. Biotechnol Bioeng 66:203–210
111. Katsumata R, Hashimoto S (1996) Process for producing alanine. US Patent 5559016
112. Zhang X, Jantama K, Moore JC et al (2007) Production of L-alanine by metabolically engineered *Escherichia coli*. Appl Microbiol Biotech 77:355–366
113. Shibatani T, Kakimoto T, Chibata I (1979) Stimulation of L-aspartate β-decarboxylase formation by L-glutamate in *Pseudomonas dacunhae* and improved production of L-alanine. Appl Environ Microbiol 38:359–364
114. Hashimoto S, Katsumata R (1999) Mechanism of alanine hyperproduction by *Arthrobacter oxydans* HAP-1: metabolic shift to fermentation under nongrowth aerobic conditions. Appl Environ Microbiol 65:2781–2783
115. Uhlenbusch I, Hermann S, Sprenger GA (1991) Expression of an L-alanine dehydrogenase gene in *Zymomonas mobilis* and excretion of L-alanine. Appl Environ Microbiol 57:1360–1366
116. Reitzer LJ (1996) Ammonia assimilation and the biosynthesis of glutamine, glutamate, aspartate, asparagine, L-alanine and D-alanine. In: Neidhart FC (ed) *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd edn. ASM, Washington DC, USA
117. Kambampati R, Lauhon CT (2000) Evidence for the transfer of sulfane sulfur from IscC to ThiI during the *in vitro* biosynthesis of 4-thiouridine in *Escherichia coli* tRNA. J Biol Chem 275:10727–10730
118. Wang M, Buckley L, Berg CM (1987) Cloning of genes that suppress an *Escherichia coli* K-12 alanine auxotroph when present in multicopy plasmids. J Bacteriol 169:5610–5614
119. Wild J, Hennig J, Lobocka M et al (1985) Identification of the *dadX* gene coding for the predominant isozyme of alanine racemase in *Escherichia coli* K-12. Mol Gen Genetics 198:315–322
120. Lee M, Smith GM, Eiteman MA et al (2004) Aerobic production of alanine by *Escherichia coli aceF ldhA* mutants expressing the *Bacillus sphaericus* alaD gene. Appl Microbiol Biotechnol 65:56–60
121. Smith GM, Lee SA, Reilly KC et al (2006) Fed-batch two-phase production of alanine by a metabolically engineered *Escherichia coli*. Biotech Lett 28:1695–1700
122. Wada M, Narita K, Yokota A (2007) Alanine production in an H^+-ATPase- and lactate dehydrogenase-defective mutant of *Escherichia coli* expressing alanine dehydrogenase. Appl Microbiol Biotechnol 76:819–825
123. Causey TB, Shanmugam KT, Yomano LP et al (2004) Engineering *Escherichia coli* for efficient conversion of glucose to pyruvate. PNAS 101:2235–2240
124. Narayanan N, Roychoudhury PK, Srivastava A (2004) L(+) Lactic acid fermentation and its product polymerization. Elec J Biotechnol 7:167–179
125. Gupta S, Clark DP (1989) *Escherichia coli* derivatives lacking both alcohol dehydrogenase and phosphotransacetylase grow anaerobically by lactate fermentation. J Bacteriol 171:3650–3655
126. Zhou S, Causey TB, Hasona A et al (2003) Production of optically pure D-lactic acid in mineral salt medium by metabolically engineered *Escherichia coli* W3110. Appl Environ Microbiol 69:399–407
127. Zhu J, Zhimizu K (2004) The effect of *pfl* gene knockout on the metabolism for optically pure D-lactate production by *Escherichia coli*. App Microbiol Biotechnol 64:367–375

128. Dien BS, Nichols NN, Bothast RJ (2001) Recombinant *Escherichia coli* engineered for production of L-lactic acid from hexose and pentose sugars. J Ind Microbiol Biotechnol 27:259–264
129. Dien BS, Nichols NN, Bothast RJ (2002) Fermentation of sugar mixtures using *Escherichia coli* catabolite repression mutants engineered for production of L-lactic acid. J Ind Microbiol Biotechnol 29:221–227
130. Zhou S, Shanmugam KT, Ingram LO (2003) Functional replacement of the *Escherichia coli* D-(–)-lactate dehydrogenase gene (ldhA) with the L-(+)-lactate dehydrogenase gene (ldhL) from *Pediococcus acidilactici*. Appl Environ Microbiol 69:2237–2244
131. Utrilla J, Gosset G, Martinez A (2009) ATP limitation in a pyruvate formate lyase mutant of *Escherichia coli* MG1655 increases glycolytic flux to D-lactate. J Ind Microbiol Biotechnol 36:1057–1062
132. Hasona A, Kim Y, Healy FG et al (2004) Pyruvate formate lyase and acetate kinase are essential for anaerobic growth of *Escherichia coli* on xylose. J Bacteriol 22:7593–7600
133. Zhou S, Yomano LP, Shanmugam KT et al (2005) Fermentation of 10% (w/v) sugar to D(–)-lactate by engineered *Escherichia coli*. Biotechnol Lett 27:1891–1896
134. Zhou S, Grabar TB, Shanmugan KT et al (2006) Betaine tripled the volumetric productivity of D-(–)-lactate by *Escherichia coli* strain SZ132 in mineral salts medium. Biotechnol Lett 28:671–676
135. Zhou S, Shanmugam KT, Yomano LP et al (2006) Fermentation of 12% (w/v) glucose to 1.2 M lactate by *Escherichia coli* strain SZ194 using mineral salts medium. Biotechnol Lett 28:663–670
136. Grabar TB, Zhou S, Shanmugam KT et al (2006) Methylglyoxal bypass identified as source of chiral contamination in L(+) and D(–) lactate fermentations by recombinant *Escherichia coli*. Biotechnol Lett 28:1527–1535
137. Fong FS, Burgard AP, Herring CD et al (2005) In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. Biotechnol Bioeng 91:643–648
138. Burgard AP, Pharkya P et al (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. Biotechnol Bioeng 84:647–657
139. Chang DE, Jung HC, Rhee JS et al (1999) Homofermentative production of D(–) or L(+) lactate in metabolically engineered *Escherichia coli* RR1. Appl Environ Microbiol 65:1384–1389
140. Zhu Y, Eiteman MA, DeWitt K et al (2007) Homolactate fermentation by metabolically engineered *Escherichia coli* strains. Appl Environ Microbiol 73:456–464
141. Stols L, Donnelly MI (1997) Production of succinic acid through overexpression of NAD dependent malic enzyme in an *Escherichia coli* mutant. Appl Environ Microbiol 63:2695–2701
142. Vemuri GN, Eiteman MA, Altman E (2002) Effects of growth mode and pyruvate carboxylase on succinic acid production by metabolically engineered strains of *Escherichia coli*. Appl Environ Microbiol 68:1715–1727
143. Vemuri GN, Eiteman MA, Altman E (2002) Succinate production in dual phase *Escherichia coli* fermentations depends on the time of transition from aerobic to anaerobic conditions. J Ind Microbiol Biotechnol 28:325–332
144. Wu H, Li ZM, Zhou L et al (2007) Improved succinic acid production in the anaerobic culture of an *Escherichia coli pflB ldhA* double mutant as a result of enhanced anaplerotic activities in the preceding aerobic culture. Appl Environ Microbiol 73:7837–7843
145. Donnelly MI, Sanville-Millard CY, Nghiem NP (2004) Method to produce succinic acid from raw hydrolysates. US Patent 6,743,610
146. Andersson C, Hodge D, Berglund KA et al (2007) Effect of different carbon sources on the production of succinic acid using metabolically engineered *Escherichia coli*. Biotechnol Prog 23:381–388

147. Millard CS, Chao YP, Liao JC et al (1996) Enhanced production of succinic acid by overexpression of phosphoenolpyruvate carboxylase in *Escherichia coli*. Appl Environ Microbiol 62:1808–1810
148. Lin H, Vadali RV, Bennett GN et al (2004) Increasing the Acetyl-CoA pool in the presence of overexpressed phosphoenolpyruvate carboxylase or pyruvate carboxylase enhances succinate production in *Escherichia coli*. Biotechnol Prog 20:1599–1604
149. Sanchez AM, Bennett GN, San KY (2005) Efficient succinic acid production from glucose through overexpression of pyruvate carboxylase in an *Escherichia coli* alcohol dehydrogenase and lactate dehydrogenase mutant. Biotechnol Prog 21:358–365
150. Sanchez AM, Bennett GN, San KY (2005) Novel pathway engineering design of the anaerobic central metabolic pathway in *Escherichia coli* to increase succinate yield and productivity. Metab Eng 7:229–239
151. Lee SJ, Lee DY, Kim TY et al (2005) Metabolic engineering of *Escherichia coli* for enhanced production of succinic acid, based on genome comparison and *in silico* gene knockout simulation. Appl Environ Microbiol 71:7880–7887
152. Jantama K, Haupt MJ, Svoronos SA et al (2008) Combining metabolic engineering and metabolic evolution to develop nonrecombinant strains of *Escherichia coli* C that produce succinate and malate. Biotechnol Bioeng 99:1140–1153
153. Jantama K, Zhang X, Moore JC et al (2008) Eliminating side products and increasing succinate yields in engineered strains of *Escherichia coli* C. Biotechnol Bioeng 101:881–893
154. Bothast RJ, Nichols NN et al (1999) Fermentations with new recombinant organisms. Biotechnol Prog 15:867–875
155. Hernández-Montalvo V, Valle F, Bolivar F et al (2001) Characterization of sugar mixtures by an Escherichia coli mutant devoid of the phosphotransferase system. Appl Microbiol Biotechnol 57:186–191
156. Korner H, Sofia HJ, Zumft WG (2003) Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. FEMS Microbiol Rev 27:559–592
157. Martinez A, Rodríguez ME, Wells ML et al (2001) Detoxification of dilute acid hydrolysates of lignocellulose with lime. Biotechnol Prog 17:287–293
158. Nichols NN, Dien BS, Bothast RJ (2001) Use of catabolite repression mutants for fermentation of sugar mixtures to ethanol. Appl Microbiol Biotechnol 56:120–125
159. Novotny MJ, Frederickson WL, Waygood EB et al (1985) Allosteric regulation of glycerol kinase by enzyme III glc of the phosphotransferase system in Escherichia coli and Salmonella typhimurium. J Bacteriol 162:810–816
160. Plumbridge J (2002) Regulation of gene expression in the PTS in Escherichia coli: the role and interactions of Mlc. Curr Opin Microbiol 5:187–193
161. Prüb BM, Campbell JW, Van Dyk TK et al (2003) FlhD/FlhC is a regulator of anaerobic respiration and the Entner-Doudoroff pathway through induction of the methyl-accepting chemotaxis protein Aer. J Bacteriol 185:534–543
162. Sheehan J, Himmel M (1999) Enzymes, energy, and the environment: a strategic perspective on the U.S. Department of Energy's research and development activities for bioethanol. Biotechnol Prog 15:817–827

Adv Biochem Engin/Biotechnol (2010) 121: 109–138
DOI: 10.1007/10_2009_64

Published online: 23 March 2010

Modeling Languages for Biochemical Network Simulation: Reaction vs Equation Based Approaches

Wolfgang Wiechert, Stephan Noack, and Atya Elsheikh

Abstract Biochemical network modeling and simulation is an essential task in any systems biology project. The systems biology markup language (SBML) was established as a standardized model exchange language for mechanistic models. A specific strength of SBML is that numerous tools for formulating, processing, simulation and analysis of models are freely available. Interestingly, in the field of multidisciplinary simulation, the problem of model exchange between different simulation tools occurred much earlier. Several general modeling languages like Modelica have been developed in the 1990s. Modelica enables an equation based modular specification of arbitrary hierarchical differential algebraic equation models. Moreover, libraries for special application domains can be rapidly developed. This contribution compares the reaction based approach of SBML with the equation based approach of Modelica and explains the specific strengths of both tools. Several biological examples illustrating essential SBML and Modelica concepts are given. The chosen criteria for tool comparison are flexibility for constraint specification, different modeling flavors, hierarchical, modular and multidisciplinary modeling. Additionally, support for spatially distributed systems, event handling and network analysis features is discussed. As a major result it is shown that the choice of the modeling tool has a strong impact on the expressivity of the specified models but also strongly depends on the requirements of the application context.

Keywords Biochemical network modeling, Modelica, Modeling languages, Object oriented modeling, SBML

W. Wiechert (✉), S. Noack, and A. Elsheikh
Institut für Biotechnologie 2, Forschungszentrum Jülich, 52428, Jülich, Germany
e-mail: wiechert@fz-juelich.de

Contents

Abbreviations

DAE	differential algebraic equation
PDAE	partial differential algebraic equation
PDE	partial differential equation
SBML	systems biology markup language
XML	extended markup language

1 Introduction

1.1 Biochemical Network Modeling

Only 10 years ago the mathematical modeling of cellular processes and, particularly, biochemical network dynamics played a rather subordinate role in biology. Although such models have been built up for more than half a century (mostly in the fields of theoretical biology and biophysics) and some simple mathematical descriptions like, e.g., the Michaelis–Menten or the Monod Model are a common heritage of biology, the use of mathematical methods to describe and analyze the behavior of complex intracellular networks was not really accepted at that time. The original idea of establishing a discipline of theoretical biology was to supply biology with a theoretical foundation in analogy to physics and chemistry. It was proposed that biological processes should have a mathematical description like other phenomena in the world of mechanics, electromagnetism, or chemistry [1]. Unfortunately, all these projects were substantially hampered by the non-availability of reliable quantitative data and the difficulty to realize a reductionistic approach to complex cellular systems.

The situation significantly changed about a decade ago when the miniaturization and automation of standard laboratory procedures led to the development of high throughput techniques for analyzing the intracellular environment. With the availability of lots of so-called "omics" data the experimental biologist was faced with a new kind of problem: mathematical tools and methods are indispensable to develop, analyze, and explain the complex intracellular interactions mirrored by the measured data sets and to design informative new experiments. This, finally, gave rise to the advent of systems biology as a new biological discipline [2, 3]. Being a systems science, engineering methods play an important role in this emerging field and, particularly, modeling and simulation became a central concern.

1.2 Standardization for Model Exchange

Dynamic biochemical network modeling is one of the most prominent approaches to describe the interaction of molecular species in the intracellular environment. This is done in a so-called mechanistic way by describing the cellular state with concentrations of those molecular species involved in the investigated processes and by specifying a reaction kinetic term per reaction as a function of substrate, product, cofactor, inhibitor, and activator concentrations [4].

The vast majority of biochemical network models that are currently available are based on the continuum and homogeneity assumptions. This means that:

1. All chemical species involved in the considered processes have such a high copy number that a continuous concentration value can be used to describe it.

2. Diffusion processes are very fast compared to chemical reactions so that concentrations can be considered to be spatially homogeneous.

As a consequence of these two assumptions, most mathematical models of biochemical networks built up to date have more or less the same underlying mathematical structure:

$$\dot{\mathbf{c}} = \mathbf{N} \cdot \mathbf{v}(\mathbf{c},\mathbf{s},\boldsymbol{\alpha}), \quad \mathbf{c}(0) = \mathbf{c}_0 \tag{1}$$

with concentration state vector **c**, stoichiometric matrix **N**, external (possibly time dependent) concentrations **s**, and kinetic parameter vector $\boldsymbol{\alpha}$.

According to (1) the whole system dynamics is governed by the set of balance equations for the changing metabolite pools together with the enzyme kinetic terms for all reaction steps. This results in a set of ordinary differential equations. In steady state investigations this system degenerates to a set of nonlinear algebraic equations. A simple running example (Fig. 1) clarifies this familiar concept.

Since biochemical network modeling has now become a common and widely applied methodology, it is not surprising that we are facing a rapidly increasing number of published models and modeling tools [5]. At the same time the question arose of how to:

- collect different models in a common knowledge repository
- make an easy exchange of models between different research groups possible
- exchange models between different computational tools for systems analysis, simulation, parameter estimation, etc.
- combine existing models in order to build up more complex ones

Clearly, these requirements immediately gave rise to a call for standardization.

Based on a general structure (1), it was actually not so difficult to establish a standard representation for the majority of "mainstream" biochemical network models. The efforts converged in the specification of the systems biology markup

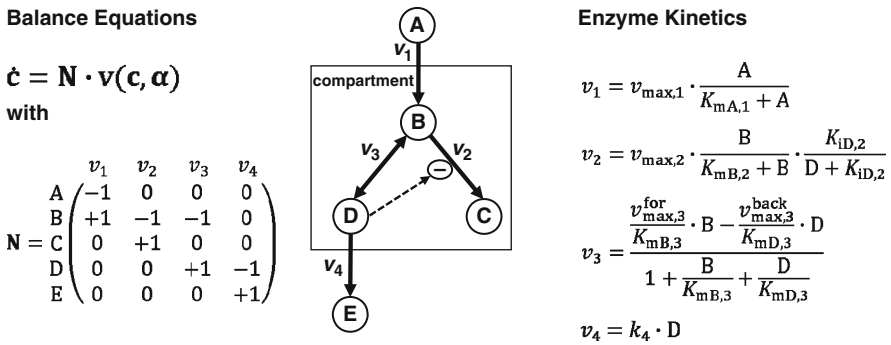


Fig. 1 Example of a biochemical network model using enzyme kinetic expressions for the description of metabolic conversions

language (SBML). The first version of SBML was published in [6] soon followed by an improved and extended second release [7]. The following sections are based on SBML 2 which has quickly become a commonly accepted standard throughout the field of systems biology.

1.3 Universal Modeling Languages

Interestingly, the same standardization problems occurred about 20 years before in the fields of mechanical, electrical, and process engineering. Quasi standards for the exchange of electrical network models, e.g., SPICE electrical part lists [8], were already established in the 1980s and different engineering fields finally found some more or less standardized ways to communicate their models. As a common experience it was not really possible to transfer large models between different research teams using different commercial and/or academic tools.

With the advent of multidisciplinary engineering fields like mechatronics, medical engineering, environmental engineering or bioprocess engineering, the need for multidisciplinary modeling and simulation tools being able to cover different physical domains in one single modeling environment became urgent [9]. At the same time, object orientation became the dominant paradigm in software engineering [10]. Object oriented system design aims at modularity, component reusability and hierarchical architectures. Clearly, these requirements are essential for mathematical modeling as well.

In the 1990s all these activities converged in several efforts to establish standardized model exchange languages for multidisciplinary modeling from which the VHDL-AMS [11] and the Modelica [12] initiatives are currently the most prominent ones. A similar academic approach which should also be mentioned here is ProMoT [13], another is introduced in [14]. With the availability of first commercial multidisciplinary simulation tools based on VHDL-AMS or Modelica, these two standards are now becoming commonly accepted and widely used in many technical disciplines. It is not an exaggeration to say that modern multidisciplinary modeling languages represent the essence of four decades of experience in engineering simulation. From now on the chapter focuses on Modelica, which is in fact quite similar to VHDL-AMS.

From a simulationist's viewpoint it is quite interesting to compare a universal multidisciplinary modeling language with a disciplinary approach like SBML. From a technical viewpoint this means to compare equation based with reaction based modeling. This is done in the following sections and as a result several important differences will be pointed out. It will be shown that the underlying features of both approaches have direct consequences for the way models are build up, the tools for model processing that have to be developed, and the future prospects of the modeling language.

2 A Brief Overview of SBML

2.1 XML Dialects for Model Specification

SBML is an extended markup language (XML) format [15] and thus supplies a hierarchical structure for the exchange of biochemical network data. The SBML standard specifies how a document must be structured, how the different entries must be named and which data must be included. Although XML is a human readable text format, it is time-consuming to extract the desired information from a huge document. For this purpose, different hierarchical browsers for displaying XML document contents like Xerlin [16] are available. However, for a special XML format like SBML the user will rarely use a universal XML editor for model specification but rather will use a domain-specific tool for biochemical network modeling [17, 18].

The great success of XML as a vehicle to carry any type of information between heterogeneous computer systems is explained by the availability of numerous software tools for specifying the structure of XML documents, parsing and extracting information from them, or modifying and transforming the structured data [15]. For this reason XML formats are used today to store configuration files of software systems, represent formatted texts, tables or graphics, exchange data or to specify communication protocols. Thus, apart from being "fashionable," it is not surprising that XML is now also being used as a state-of-the-art tool for representing structured mathematical models. Particularly, in the case of SBML, a well sorted set of software tools for reading, manipulating, and translating XML models as well as code generation for simulation are available [19].

2.2 SBML Structure

It makes little sense to give an in depth description of SBML in this chapter. To this end the reader is referred to the official SBML specification documents [20]. In order to illustrate the basic concepts, the simple example from Fig. 1 will be used. The different sections of a hierarchically structured SBML description are given as follows (cf. Fig. 2):

1. The whole system might be spatially structured into different disjoined homogeneous compartments. A list of these compartments can be given in the compartment section.
2. The species section gives a list of all chemical substances involved in the reaction network (here: A, B, etc.). For each species it must be specified to which spatial compartment it belongs. Additionally, the biochemical name of a species and its initial concentration can be supplied.
3. All network reactions are arranged in the reaction section (here: v_1, v_2, etc.). Each reaction definition includes a list of all reactants and products as wells as a

```
<?xml version="1.0" encoding="UTF-8" standalone="no" ?>
- <sbml xmlns="http://www.sbml.org/sbml/level2/version3" level="2"
    metaid="metaid_01" version="3">
  - <model id="model_01" metaid="metaid_02" name="SimpleNetwork">
    + <annotation>
    + <listOfCompartments>
    - <listOfSpecies>
        <species compartment="compartment_01" id="species_01"
          initialAmount="10.0" metaid="metaid_02" name="A" />
        <species compartment="compartment_01" id="species_02"
          metaid="metaid_03" name="B" />
      </listOfSpecies>
    - <listOfReactions>
      - <reaction id="reaction_01" metaid="metaid_04" name="v1" reversible="false">
        - <listOfReactants>
            <speciesReference name="A" species="species_01" />
          </listOfReactants>
        - <listOfProducts>
            <speciesReference name="B" species="species_02" />
          </listOfProducts>
        - <kineticLaw>
          + <math xmlns="http://www.w3.org/1998/Math/MathML">
          - <listOfParameters>
              <parameter id="parameter_01" name="vmax" value="10.0" />
              <parameter id="parameter_02" name="km" value="0.05" />
            </listOfParameters>
          </kineticLaw>
        </reaction>
      </listOfReactions>
    </model>
  </sbml>
```

Fig. 2 Segment of an SBML file representing a part of the biochemical network shown in Fig. 1

mathematical term for a kinetic model describing the biochemical conversion. Corresponding kinetic parameters and their values are specified in the parameter list (here: $v_{\max,1}$, $k_{\mathrm{mA},1}$, etc.).

Given this information, the complete model shown in Fig. 1 can be assembled. Using an SBML based biochemical network modeling tool like COPASI [18] the information can be supplied in a much more user friendly form from which the SBML document is automatically generated.

3 A Brief Overview of Modelica

3.1 Equation Based Simulation Languages

The Modelica language initiative was a result of an increasing desire in the scientific computing community for supporting multidisciplinary modeling. Since mathematics is the only language common to all scientific disciplines, Modelica is

based on the elementary building blocks of algebraic and differential equations. For this reason Modelica is called an equation based language. Equations can be assembled into physically meaningful building blocks which, again, can be assembled into large system models. Particularly, Modelica is open to implement special libraries for any type of physical application domain. Libraries for mechanical multibody systems, electric circuits, hydraulics, control engineering, thermodynamics, and many other domains are already available (www.modelica.org). A European consortium is currently enhancing an already existing large set of free libraries with other libraries with further application domains [21].

As a consequence of its universal scope, Modelica must be based on domain neutral concepts which do not already imply any physical application. In the present case, the chemical reaction steps used in SBML as elementary building blocks are obviously not domain neutral. In Modelica, physical meaning is generated by implementing application specific libraries from basic equation blocks.

Essentially, Modelica looks like an object oriented programming language with the difference that the compiler target is not an executable computer program but an executable simulation code. Like with program source files, a well structured and documented Modelica code is human readable, particularly if a special programming editor with syntax highlighting is used (cf. Fig. 3). Moreover, Modelica – like modern code documentation systems – defines a standardized way to generate

```
model SimpleNetwork
  //Declaration part
  //Metabolite concentrations
  Real A(start=10.0), B(start=0.1), ...;
  //Reaction rates
  Real v1, v2, ...;
  //Kinetic parameters
  parameter Real vmax1=10.0, kmA1=0.05, ...;
equation
  //Balances
  der(A)=-v1;
  der(B)=v1-v2-v3;
  der(C)=v2;
  der(D)=v3-v4;
  der(E)=v4;
  //Kinetics
  v1=vmax1*A/(A+kmA1);
  v2=vmax2*B/(kmB2+B)*kiD2/(D+kiD2);
  v3=(vmax3f/kmB3*B-vmax3b/kmD3*D)
            /(1+B/kmB3+D/kmD3);
  v4=k4*D;
end SimpleNetwork;
```

Fig. 3 Flat implementation of the network shown in Fig. 1 with Modelica. Some repetitive parts are abbreviated by *dots*

HTML-documentation that can be browsed by HTML-browsers or advanced Modelica environments.

Clearly, from the facts mentioned above, there is no previously specified structure in which a biochemical network model has to be encoded in Modelica. The general philosophy behind Modelica is that the language itself is the standard for model exchange. Consequently, the modeler can express his interpretation's perspectives and point of views in a Modelica specification. Here, the most basic representation is a so-called flat code in which all required algebraic and differential equations are sequentially specified. A straightforward implementation of the running example may look as shown in Fig. 3. This file already contains all the necessary information to generate an executable simulation program. Moreover, the role of identifiers as state variables, external variables, or parameters can be understood from the context and is recognized by the compiler tool.

3.2 Building Components and Libraries

Although the specification of equation systems by flat models (i.e., lists of equations) is straightforward, this approach is not really recommendable for the practical development of large models or even reusable components to be used in other models. To specify models in a structured way, Modelica supplies powerful mechanisms for modular and hierarchical modeling. Basically, Modelica is based on the structural features of object oriented programming languages as far as this makes sense in the simulation context. For this purpose, equation systems can be broken down into independent subsets, constituting the components. Every component can be connected to other components using a special type of interface object called a connector.

Once a component has been specified, it can be reused many times in a system model. Modelica language formalism automatically takes care of the unique identification of quantities in each sub model of a large model. To this end a familiar hierarchical "dot notation" is used.

Before the mechanism of assembling independent components to meaningful system models is explained in detail, the concept of a connector must be clarified. Essentially, connectors are bidirectional information channels, which identify variables from one component with those of another. In the running example a connector carries two variables, a substance concentration and a reaction flux (Fig. 4). The meaning of the keyword *flow* will be explained soon.

Once the connector's definition is provided, it is possible to specify a model for Michaelis–Menten reactions as shown in Fig. 5. This reaction is connected to its substrate S and product P by two connectors. The meaning of the two associated flux variables is the amount of material withdrawn from S on the one hand and supplied to P on the other. Clearly, these two fluxes are related by the stoichiometry of the reaction which in this case simply means the identity of both fluxes. In Fig. 5 this identity has a minus sign because the standard convention of Modelica is

```
connector cp "Chemical port"
  Real c "Metabolite concentration";
  flow Real v "Reaction rate";
end cp;
```

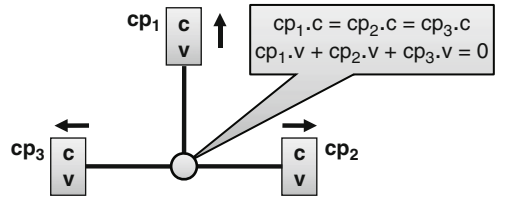


Fig. 4 Implementation and meaning of connectors in Modelica. Connecting three connectors assembles the equations shown. These connectors may belong to independent components (e.g., metabolite, reaction, effector)

```
model DynMet
  cp M "Chemical port";
  //Initial value definition
  parameter Real c_0=0.1;
  //Metabolite concentration
  Real c(start=c_0);
equation
  //Balance
  der(c)=M.v;
  M.c=c;
end DynMet;
```

```
model MM_1S
  cp S, P "Chemical ports";
  Real v "Reaction rate";
  //Kinetic parameters
  parameter Real vmax=10.0;
  parameter Real km=0.05;
equation
  //Kinetic model
  v=vmax*S.c/(km+S.c);
  S.v=v;
  P.v=-v;
end MM_1S;
```

Fig. 5 The implementation of Modelica components representing metabolites and reactions as basic parts of metabolic networks

that positive fluxes are always in the direction from inside to outside of a component. Thus, there is an inwards and an outwards flow with a positive and negative sign, respectively.

Other reaction types can be introduced in a quite similar manner. The implementation of reactions with many substrates or products is straightforward. Even when effectors like inhibitors or activators are present, this case can be easily handled by declaring an effector's connector with zero flux. This simply means that effectors influence a reaction rate but do not participate in the reaction itself (cf. Fig. 6).

Having defined the reaction steps, the component specification for chemical substance pools is surprisingly simple (cf. Fig. 5). A chemical compound has only a single connector. It will soon become clear that this connector expresses the result of the material balance of fluxes around this pool. Knowing this, the change of the pool size is simply given by the flux variable itself which can have a positive or negative sign.

The reason why this simple description of pools works is the inherent Modelica mechanism for connecting components. Essentially, two components are not directly

```
model MM_1S1I
  //Chemical ports
  cp S, P;
  cp_I I;
  Real v "Reaction rate";
  //Kinetic parameters
  parameter Real vmax=0.6;
  parameter Real km=1.0;
  parameter Real ki=0.4;
equation
  //Kinetic model
  v=vmax*S.c/(km+S.c)*ki/(I.c+ki);
  S.v=v;
  P.v=-v;
  //No conversion of inhibitor
  I.v=0;
end MM_1S1I;
```

Fig. 6 Modelica component for modeling an inhibited reaction

connected by one connector but there is a coupling node between all connected components. It is possible to connect arbitrarily many connectors to a coupling node as shown in Fig. 4. In the Modelica language, this is simply achieved with the *connect* statement. As an example, the complete network specification for the example of Fig. 1 is shown in Fig. 7. Basically, this is a system parts list together with a wiring scheme.

3.3 Automatic Code Generation

What does it mean to couple three different connectors (cp) with associated concentration (c) and flux variables (v) to a single node? The answer depends on the type of variable given in the connector specification (Fig. 4). If the keyword *flow* is missing in a variable specification, the variable is called a potential variable. This means that the respective variables of all linked connectors are equal (i.e., they denote the same physical quantity). In our case this means that, at the coupling node, the identity equations

$$\mathrm{cp}_1.\mathrm{c} = \mathrm{cp}_2.\mathrm{c} = \mathrm{cp}_3.\mathrm{c} \tag{2}$$

are generated. On the other hand, if the keyword *flow* is present, the variable is treated as a flow quantity and a Kirchhoff-like material balance is generated:

$$\mathrm{cp}_1.\mathrm{v} + \mathrm{cp}_2.\mathrm{v} + \mathrm{cp}_3.\mathrm{v} = 0 \tag{3}$$

```
model SimpleNetwork
  //Metabolites
  DynMet A(c_0=10.0);
  DynMet B, C, D, E;
  //Kinetic models
  MM_1S v1(vmax=10.0, km=0.05);
  MM_1S1I v2(vmax=0.6, km=1.0, ki=0.4);
  MM_1S1P v3(vmaxf=0.3, vmaxb=0.4,
             kmS=0.4, kmP=0.5);
  Trp_1S v4(k=0.5);
equation
  connect(A.M, v1.S);
  connect(v1.P, B.M);
  connect(B.M, v2.S);
  connect(v2.P, C.M);
  connect(D.M, v2.I);
  connect(B.M, v3.S);
  connect(v3.P, D.M);
  connect(D.M, v4.S);
  connect(v4.P, E.M);
end SimpleNetwork;
```

Fig. 7 The implementation of the complete network of Fig. 1 in Modelica

The resulting flat equation system for the running example, which is generated from the wiring scheme of the components, is shown in Fig. 8. Obviously, the Kirchhoff laws introduce exactly the right balance equations for each substance pool, particularly if the sign convention is applied properly.

The sketched equation generation procedure explains how a fully specified system model is assembled from its components and wiring scheme. The final result is a flat model which can also be displayed to the user. In this flat model the Kirchhoff laws and the identification of potential variables introduce a lot of superfluous variables. However, model simplification by variable elimination and automatic formula manipulation is automatically done by advanced Modelica compilers [22, 23]. In fact, even nonlinear relations are simplified based on state of the art computer algebraic algorithms. In addition, high-level optimization methods based on graph algorithms for system decomposition are utilized [9, 24]. However, the end user is not aware of all these operations running in the background.

This is the basic procedure which enables the Modelica user to specify his models on a high modular level from which a flat model is automatically generated in a completely transparent way. Once a library for biochemical network simulation is implemented, a non-specialist user can simply assemble his models on a high-level where each component has a domain-specific meaning. It is also possible to extend the library if necessary. Moreover, available graphical modeling environments make the library components access easy and support a non-specialist user in the model building process (cf. Fig. 9).

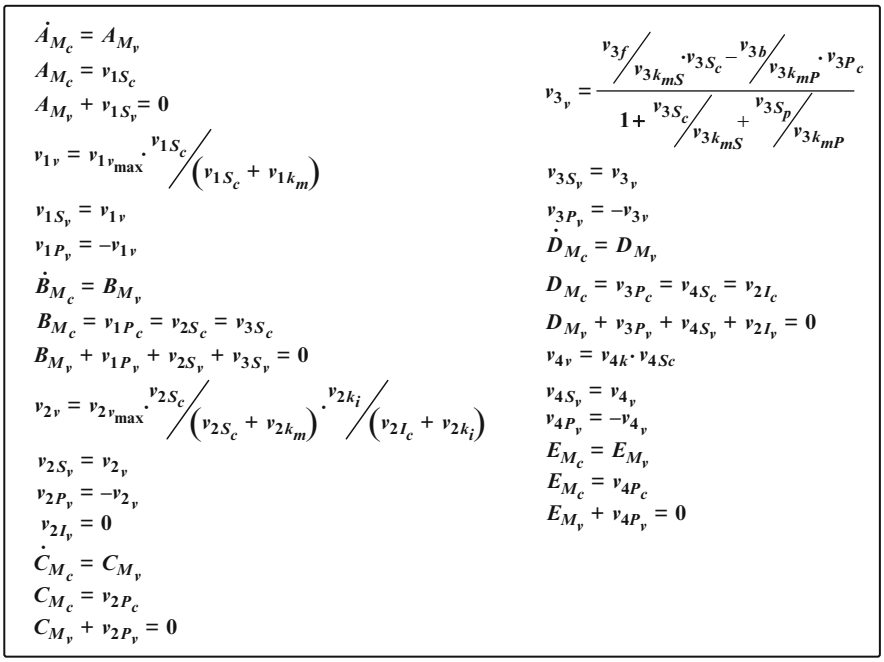


Fig. 8 Resulting flat equation system obtained by applying the connector-laws to the components wiring scheme of Fig. 7. Each component whether a reaction (v_1, v_2, ..., v_5) or a metabolite (A, B, ..., E) communicates with the outside world via two interfaces (S)ubstrate and (P)roduct. Each interface is characterized by two quantities, reaction rate (v), and concentration (c). For example, v_{1S_c} describes the concentration of the substance connected to the reaction v_1 through the interface S

4 Further Constraints on Biochemical Network Models

In this and the following sections the reaction and equation based approaches to biochemical network modeling are compared with respect to different criteria. It will soon become clear that they are not really competitive approaches but rather are complementary with specific strengths in the one or the other application context.

4.1 Different Types of Constraints in SBML

The standard way to assemble a biochemical network model from substance balance equations and reaction kinetics is not always sufficient to express any assumption made on the system or some other conditions which are not directly related to biochemistry. This makes it necessary to have a mechanism for extending

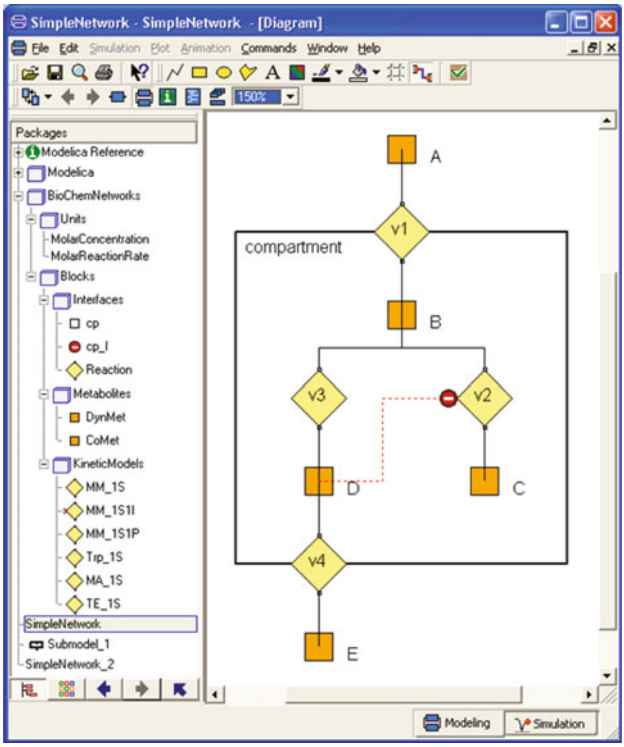


Fig. 9 Dymola GUI for building up and simulating system models using self-defined Modelica libraries

standard network models with additional constraint equations. In the *rules* section of an SBML model different kinds of equation constraints are allowed:

1. *AlgebraicRule*: $0 = f(\mathbf{w})$
2. *AssignmentRule*: $x = f(\mathbf{u})$
3. *RateRule*: $\dot{x} = f(\mathbf{w})$

Here, x is a variable, f is some arbitrary function returning a numerical result, $\mathbf{u}$ is a vector of variables that does not include x, and w is a vector of variables that may include x.

Essentially, by introducing the rules section, equation based modeling is introduced into SBML. Although, any additional model feature can be expressed by using rules, this takes place in an unstructured way.

Some simple examples shall be given to illustrate the occurrence of additional equation constraints.

1. The first example is thermodynamic equilibrium which might be an assumption for one reaction step A $\rightarrow$ B. If thermodynamic equilibrium is considered as a constraint it is not necessary to specify a reaction rate. Instead, an algebraic

relation between the concentrations of both pools A and B can be used which is given by

$$v_f \cdot A = v_b \cdot B \tag{4}$$

Here, v_f and v_b denote the forward and backward rate constants, respectively.

2. Another type of constraints is given by conservation laws for conserved moieties. If, for example, the dynamics of the energy metabolites AMP, ADP, ATP is not explicitly described in the model, a conservation relation might still be given:

$$\mathrm{AMP} + \mathrm{ATP} = 2\,\mathrm{ADP} \tag{5}$$

3. In many cases, the intracellular biochemical network is modeled together with the surrounding environment such as, for example, in bioreactor models. In this case the bioreactor balances, which cannot be straightforwardly expressed as a reaction system, must be added to the cell model.
4. In [25] a multiscale model was presented that combines steady state stoichiometric equations and overall growth kinetics with a detailed mechanistic description of a specific pathway. Additionally, to fill in the remaining degrees of freedom for a unique model solution, some phenomenological relations between concentrations were added based on the available measurement data. The resulting model can only be expressed in SBML if most of the equations are given in the *rules* section.

4.2 Realizing Constraints in Modelica

Since Modelica is an equation based modeling language, the incorporation of any type of equations into a model poses no problem at all. Particularly, no additional concepts have to be introduced to mix differential equations with algebraic equations. After assembling a model from both types of equations a DAE system emerges. This is exactly the type of mathematical model structure Modelica is made for.

Since the 1980s the theory of DAEs has made great progress and both the similarities and differences to ODEs without purely algebraic parts are well understood [26]. As a consequence several powerful numerical DAE solvers are available today [27]. From the start, the development of Modelica was based on these advanced solvers. Particularly, both special cases of pure ODE models or pure algebraic models (including steady state and quasi steady state systems) are covered.

It is more interesting to see how additional equation constraints can be added to a model on the level of modular modeling. This is illustrated again with the running

```
model TE_1S
  cp S, P "Chemical ports";
  Real v "Reaction rate";
  //Kinetic parameters
  parameter Real kf=1.0;
  parameter Real kb=0.5;
equation
  //Thermodynamic equilibrium
  kf*S.c=kb*P.c;
  S.v=v;
  P.v=-v;
end TE_1S;
```

Fig. 10 Constraint based modeling in Modelica exemplified by a thermodynamic equilibrium relation

example. Here, a rapid equilibrium assumption can be represented as a Modelica reaction step component as shown in Fig. 10. This component can replace any reaction kinetic component with the same number of substrates and products.

Likewise, a conservation law or other non-local constraints can be added to a network by a separate component that again is connected to the respective substrate pools. If cells are modeled together with a bioreactor, this is clearly done by establishing a component library for bioreactors and coupling it with a cell model via substrate uptake rate, growth rate, product formation, etc. In such cases the strength of Modelica to represent multidisciplinary systems becomes obvious.

4.3 Equation Sorting and Consistency Checking

Specifying additional constraints on a given model can render the equation system inconsistent. There might be an over or under specification which can be detected by the model assembly algorithm [26, 28]. Even more challenging is the so-called index problem of DAEs [29, 30]. To illustrate this problem, the example system of Fig. 1 is used again, but now the Michaelis–Menten term of reaction v_1 is replaced by an equilibrium relation between A and B (see (4)):

$$\dot{A} = -v_x, \quad \dot{B} = v_x - v_2 - v_3, \quad v_x = f(A,B) \tag{6}$$

Obviously, the equilibrium condition in this model does not directly give the information about how to solve for the unknown net reaction rate v_x of the corresponding reaction. The missing information can be reconciled from the original model by the additional equation differentiation. By differentiating the equilibrium constraint with respect to time a new differential equation is obtained:

$$v_\mathrm{f} \cdot \dot{A} = v_\mathrm{b} \cdot \dot{B} \Leftrightarrow v_x = -v_\mathrm{b}/v_\mathrm{f} \cdot \dot{B} \tag{7}$$

This solves the problem, because (7) establishes a direct relation between the unknown variable v_x and the state variables A, B. From the theory of DAEs it is well known when and how such a procedure works [30]. However, it requires an automatic manipulation of equation systems by computer algebraic methods. Modelica contains powerful computer algebraic machinery and thus can automatically diagnose and solve inconsistency and index problems (if possible). Clearly, any SBML tool dealing with additional rules will need the same algorithms to assist the user in assembling proper and meaningful models.

5 Different Modeling Flavors

5.1 Modeling as a Creative Act

Although standardized system representations have been developed for certain types of physical domains like multibody systems, electrical circuits, or chemical reaction networks, there is no unique way of how to model a given real system. In any case, modeling is also a creative activity in which a real system is interpreted in terms of mental concepts. Particularly, there might be good reasons for not using a standard system representation. Examples are given below.

When comparing modeling languages with programming languages, the language of chemical reaction equations might be considered as the "assembler language" of cellular process modeling. This means that the most basic process in the cellular environment is used as the building block for network modeling, much like the elementary computational operations of a computer processor are used as the building blocks to execute any program. Clearly, if homogeneity is assumed, anything happening in a cell can be expressed on this level. However, the human user might like to introduce a higher level language that builds up models from aggregates of simple biologically meaningful components.

As an example, a simple model for the expression of a gene is used here which can be described on the reaction level as follows:

$$\dot{mRNA} = f_{\mathrm{tr}}(R) - d_{\mathrm{mRNA}} \cdot mRNA, \quad \dot{P} = k_{\mathrm{tl}} \cdot mRNA - d_{\mathrm{P}} \cdot P \tag{8}$$

Here, the mRNA concentration depends on a transcription function $f_{\mathrm{tr}}(R)$ with repressor R and the corresponding protein P is synthesized with a linear law. Both quantities underlie degradation processes (constants $d_{\mathrm{mRNA}}, d_{\mathrm{P}}$).

Since this sequence is the same for any gene in the system, the whole process might be called "gene expression" thus introducing a new concept into the modeling framework. Generally, the introduction of a new terminology to describe a system on a higher level of abstraction is a very common procedure and is constitutive for the general methodology of conceptual system modeling.

As an example in [31] a highly developed conceptual framework for the modeling of biochemical networks has been given which classifies the processes taking

place in a living cell in terms of control engineering concepts. Thereby a clear distinction is made between substance storages (with and without genetic information) and substance transformers (e.g., enzymatic reactions, degradation, and polymerization processes). A further modeling object named signal transformer can be used for the description of signal transduction and processing.

Obviously, conceptual modeling approaches to cellular systems are not directly implementable within SBML because the vocabulary of this modeling language is fixed. This is quite different with Modelica which allows one to introduce arbitrary new concepts and even to build different types of libraries which conceptualize the same system in different ways. For example, a component describing gene expression is shown in Fig. 11.

The clear advantage of conceptual modeling is that Modelica offers ultimate flexibility for the modeler to express his interpretation of the system structure. On the other hand, this means that different frameworks will lead to different Modelica libraries which might be more or less accepted and, in particular, are not compatible with each other.

5.2 *Hierarchical Modular and Object Oriented Modeling*

One major concern in systems biology is to obtain a structured representation of complex biological systems in terms of hierarchically organized components. For

```
model GE
  cp R, P "Chemical ports";
  Real mRNA(start=0.1) "mRNA concentration";
  Real vtr "Transcription rate";
  Real vtl "Translation rate";
  //Kinetic parameters
  parameter Real kR=0.5;
  parameter Real m=3.0;
  parameter Real ktl=0.3;
  parameter Real dmRNA=0.1;
  parameter Real dP=0.1;
equation
  //Balances
  der(mRNA)=vtr-dmRNA*mRNA;
  der(P.c)=vtl-dP*P.c;
  //Kinetic models
  vtr=1-R.c^m/(kR^m+R.c^m);
  vtl=ktl*mRNA.c;
  //No conversion of repressor
  R.v=0;
end GE;
```

Fig. 11 Applying conceptual modeling in Modelica by formulating components of higher abstraction levels, e.g., gene expression

example, there are hierarchies of single reactions, pathways and parts in metabolism or genes, operons and regulons in the genetic system. These hierarchies should be represented in a model by an analogous representation. Again, the SBML approach – remaining on the assembler level – is very limited to represent such a hierarchical system structure.

This is quite different with Modelica which, from the start, was designed for the modeling of engineering systems having a natural modular and hierarchical architecture. As a simple example a series of single reaction steps can be combined to a pathway by simply assembling them into a submodel (Fig. 12). In this way, any model can be assembled to larger models which again can be used as building blocks for the next hierarchical level. This approach has been extensively demonstrated in [32].

It should be noticed that hierarchical modeling was already implemented in rather early signal flow oriented tools like Simulink. However, signal flow modeling is based on the concept of causality: for each component it has to be specified what the input and output signals are. This already determines a block-wise sequence of computation. In contrast, Modelica and SBML are non-causal, meaning that the equation system is first completely generated and then numerically treated as a whole. Essentially, it is the responsibility of the Modelica compiler to determine the flow of information among system's variables. For this reason, the non-causal approach is more powerful but also more demanding than the signal flow approach [9].

A modern extension to modular and hierarchical modeling is the object oriented paradigm. Modelica is based on object oriented concepts as far as this makes sense with an equation based approach. The most characteristic feature of object oriented systems is certainly class abstraction. Similar models can be generalized to an abstract model class which is later specialized by implementing the different details. This gives additional power for conceptual modeling but also for a structured extension of existing libraries by implementing new submodels derived from abstract classes and implementing them in a system model.

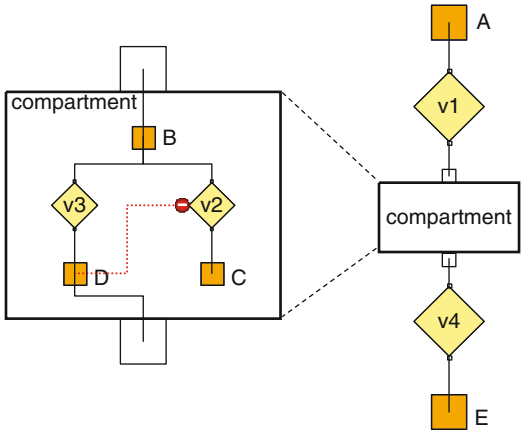


Fig. 12 Hierarchical modeling with Modelica. All reactions inside the compartment of the example network of Fig. 1 are combined into a submodel that is connected to the remaining network parts

As an example, a reaction step is considered. Without having a concrete law for the reaction kinetics, an abstract reaction step can first be introduced as shown in Fig. 13. Here, only the component interface is specified by the connectors, but the formula for the reaction kinetics is missing. Later on, concrete reaction mechanisms can be plugged in by class derivation. This is shown in Fig. 13 for a simple mass action law, a Michaelis–Menten mechanism, and a thermodynamic equilibrium assumption. By using class derivation, reaction laws can be quickly exchanged in a model while still keeping the interfacing structure and the connectivity fixed.

5.3 *Multidisciplinary Modeling*

As mentioned before, in different physical domains modeling has become quite standardized today. For example, there is nothing to be discussed about how to formulate the equations of mechanical mass-spring-damper systems by application of Newton's law, or how to formulate equations for resistance–capacity–inductivity electrical networks by using Kirchhoff's law. This is also the reason why specialized easy-to-use disciplinary simulation tools are highly developed today. Biochemical networks are just another such domain. By using SBML one can rely on a standardized model structure and thus use powerful specially adapted programs for modeling and simulation.

However, once the borders of a discipline are crossed because another physical domain enters the system, domain-specific simulation tools run into problems. This has already been demonstrated above for the SBML rules section. There are many

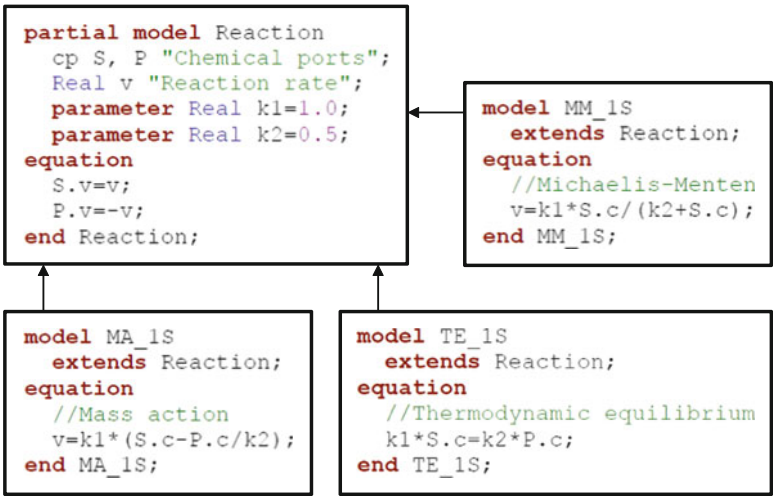


Fig. 13 The concept of class derivation in Modelica. By specifying an abstract reaction class different kinetic models can be derived by only supplementing the specific reaction mechanism

examples where the formalism of standard biochemical network modeling needs to be extended:

1. Cellular metabolism is coupled to the dynamics of the surrounding environment.
2. Several different organisms are interacting.
3. Electrochemical effects have to be described.
4. The functioning of a measurement device must be modeled.

One frequently chosen way to deal with multidisciplinary problems is simulator coupling. In this case disciplinary simulators for different domains are driven in parallel and the interfacing data is exchanged by using application programming interfaces (APIs). This rather cumbersome procedure requires a lot of manual work, a good knowledge of the APIs, as well as some understanding of the underlying numerical algorithms and synchronization procedures. However, once the next version of a simulator is available, the established simulator coupling might have to be adapted. It is not surprising that model interfacing und simulator coupling is currently being discussed in the SBML community.

This problem does not occur with multidisciplinary simulation tools because they were made for solving exactly this problem. In Modelica it is possible to integrate models from different physical domains into one multidisciplinary system model. The only new elements that might have to be introduced are the interfaces between the domains. These interfaces are usually given by components that connect to variables of different physical units and describe the translation from one world to the other. This, for example, holds true for almost any sensor or actuator device.

6 Spatially Distributed Models

6.1 From DAEs to PDAEs

So far, the restriction was made that all systems are spatially homogenous with the consequence that each species can be described by one single concentration value. A first extension to this situation is a multicompartmental system. Spatial transport processes between different compartments of the system (e.g., between cytosol and mitochondrium) are then modeled as a pseudo reaction with the conceptual difference that the transported species remains unchanged while it changes the compartment.

The situation is quite different when slow diffusion processes occur in the system leading to continuous concentration gradients. Taking a rigorous approach this means that any variable becomes dependent on the spatial coordinates. The DAE system then has to be replaced by partial differential algebraic equation (PDAE) systems. This is the most general class of continuous time system descriptions that can be handled by standard software tools today.

Although the Modelica language provides constructs for specifying partial differential equations (PDEs), the Modelica community is still waiting for a rigid implementation for simulating PDEs from one of the standard compiler available. Nevertheless, several attempts for supporting PDE have been performed by several research groups independently [33–35]. However, if diffusion processes need only an approximate representation on a rough spatial discretization grid, it is possible to describe it within the reaction network framework using PDAEs.

As an example, consider a one-dimensional spatial domain x as illustrated in Fig. 14. It is discretized with a step size of Δx. If diffusion is slow the substances described by the concentration variables c_i can be assumed as homogenously distributed in each single micro compartment. Diffusion then can be described by diffusion flows between spatial compartments. These transport processes are approximately given by Fick's law as

$$J = -D \cdot \frac{(A_i - A_{i-1})}{\Delta x} \tag{9}$$

where J is the flux and D the diffusion coefficient. In this way transport can be interpreted as a reaction rate of an inter-compartmental pseudo reaction step.

Summarizing, any spatially extended reaction diffusion system can be modeled within the formalism of chemical reaction networks. However, there are two practical problems:

1. If there are many reacting species, the whole reaction network has to be duplicated within each single compartment. This means that multiple identifiers have to be introduced denoting the same substance in different compartments (cf. Fig. 14).
2. If the spatial discretization step size Δx is changed, the whole equation system has to be newly generated which makes automatic model generation procedures mandatory.

In SBML this approach is, in principle, possible but additional tools are required to generate the spatially distributed model from a given one-compartment network. Clearly, the generated model would no more be readable for a human reader.

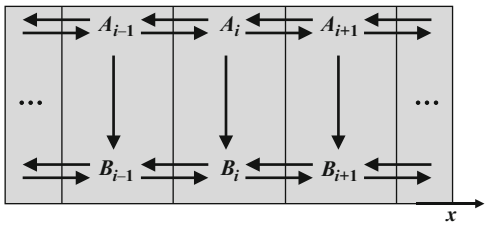


Fig. 14 Discretization of a diffusion process of two reactants along a one-dimensional spatial domain

6.2 Spatially Distributed Systems in Modelica

Interestingly, this is quite different with the Modelica approach because multiple copies of the same model structure are inherently possible in this language. The enumeration of multiple copies of the same variables in different compartments is automatically managed by the system. Moreover, Modelica contains enhanced language elements to generate dynamically a model structure in dependence of user specified parameters like step sizes. For example, the diffusion model shown in Fig. 14 can be easily implemented in Modelica by using a for-loop that automatically establishes the network connectivity as shown in Fig. 15. It should be clear that it is also possible to implement a copy of a complete reaction network in each of the compartments without any efforts.

It is well known that PDEs in more than one dimension lead to new problems if the shape of the underlying domain is not rectangular. Non-rectangular grids and appropriate numerical discretization schemes must then be used to approximate the system equations. This leads, for example, to the finite volume or finite element methods.

When interpreting these two approaches in the right way it becomes apparent that both methods essentially do the same as the one-dimensional approach shown above: For any two adjacent volumes a transport flow is defined in terms of the concentrations in both volumes. However, the corresponding laws look much more complicated than in the simple one-dimensional example. Nevertheless, the finite elements and finite volumes methods can be implemented in Modelica in a similar manner as that demonstrated for the one-dimensional case. The details are presented in [36] and would exceed the scope of this chapter.

```
model Diffusion
  parameter Integer N=100;
  DynMet[N+1] A, B;
  MM_1S1P[N] RAB, RA, RB;
equation
  for i in 1:N loop
    connect(A[i].M,RAB[i].S);
    connect(B[i].M,RAB[i].P);
    connect(A[i].M,RA[i].S);
    connect(A[i+1].M,RA[i].P);
    connect(B[i].M,RB[i].S);
    connect(B[i+1].M,RB[i].P);
  end for;
end Diffusion;
```

Fig. 15 The implementation of a diffusion model as an example for spatially distributed modeling in Modelica

7 Events

7.1 Events in Continuous Time Systems Simulation

The term "event" denotes a point on the time scale where a discontinuous change of the system state happens in zero time. Usually it is distinguished between time events for which the time is known in advance and state events which happen in dependence of the current state of the system. Clearly, in a continuous world discrete changes of variables do not really happen. However, the transient processes might be so fast in comparison to the time constants of the overall system dynamics that it makes sense to condense them to one single time event. Different examples are:

- switching operations in an experiment as for example switching between two substrates
- external variables modeled by spline functions (i.e., case distinctions)
- threshold processes modeled by discontinuous kinetic terms
- genetic switches

In practice an event is specified by a Boolean condition as a function of time and state variables. At the moment this condition becomes true, an action is specified by a function computing a discontinuous jump in the state space. Afterwards the differential equation solvers are reinitialized with the changed values. The incorporation of events into differential equation solvers is state-of-the-art but is not free of problems because state events must be automatically and precisely located and events that happen shortly one after the other must be distinguished [37].

7.2 Events in SBML and Modelica

SBML introduces an optional list of *Event* objects, that describe the time and form of explicit instantaneous discontinuous state changes in the model. An *Event* definition requires two parts, a *Trigger* condition and a list of *EventAssignment*. The *Trigger* condition represents a mathematical Boolean expression. If the *Trigger* evaluates to true, the *Event* is fired, and the *EventAssignment* takes place. The *EventAssignment* represents a mathematical expression which a concerned object is assigned to. The semantics of an *Event* imposes the non-confliction of *EventAssignments*, i.e., the effect of one assignment should not affect the result of another assignment. In Modelica, this semantics is guaranteed at the syntax level. Such confliction is detected by the Modelica compiler at static level, i.e., compile time.

In Modelica events are integrated into the modeling language and can be specified everywhere in a model or its submodels. The when statement is used within equations, e.g.:

$$\mathrm{der}\,(\mathrm{M.c}) = \mathrm{when}\,(\mathrm{M.c} > 0)\,\mathrm{M.v}\,\mathrm{else}\,0$$

An event occurs when the condition evaluates to true and is handled accordingly by the DAE solver used by the Modelica Compiler.

8 Systems Analysis

8.1 Sensitivity Analysis

Simulation tools not only perform numerical solutions based on the system equations but also assist the modeler in systems analysis. Doubtlessly the most important systems analysis tool is sensitivity analysis. It linearizes a system around a reference state and, thus, makes any kind of linear method – at least in an approximate sense – available for systems analysis. For example, sensitivity analysis is required for parameter fitting, statistical regression analysis, experimental design, and metabolic control theory.

In the context of biochemical network simulation the most wanted sensitivities are those of the state variables with respect to kinetic parameters and initial values ($\mathrm{d}c/\mathrm{d}\alpha$, $\mathrm{d}c/\mathrm{d}c_0$, cf. (1)) and the linearization around a reference state ($\mathrm{d}\dot{c}/\mathrm{d}c$).

Implementation of sensitivity analysis can be done in three different ways:

1. The simplest approach is numerical differentiation based on approximation formulas for the derivatives which use only function evaluations. The best known approximation of course is the differential quotient or finite difference approximation

$$\frac{\mathrm{d}c}{\mathrm{d}\alpha} \approx \frac{c(\alpha + \Delta\alpha) - c(\alpha)}{\Delta\alpha} \tag{10}$$

Numerical differentiation is very easy to implement because no internal details about a simulation tool or a model need be known to use it. The simulator has just to be started several times. On the other hand, finite differences have several well known numerical drawbacks among which are low error order and ill condition with respect to numerical errors. Step size control algorithms and higher order formulas are available, all of which need higher computing time due to multiple function evaluations [38].

2. Another approach to compute sensitivities is to use a computer algebra system that calculates the sensitivity equations explicitly [39]. Basically this generates a new DAE system from the original one. In the case of biochemical reaction networks, the derived sensitivity equation system has some structural features in common with the original biochemical network model. This becomes evident from the parameter sensitivities of (1):

$$\frac{\partial \dot{\mathbf{c}}}{\partial \alpha} \approx \mathbf{N} \cdot \left(\frac{\partial \mathbf{v}}{\partial \mathbf{c}} \frac{\partial \mathbf{c}}{\partial \boldsymbol{\alpha}} + \frac{\partial \mathbf{v}}{\partial \mathbf{s}} \frac{\partial \mathbf{s}}{\partial \boldsymbol{\alpha}} + \frac{\partial \mathbf{v}}{\partial \boldsymbol{\alpha}} \right), \quad \frac{\partial \mathbf{c}}{\partial \boldsymbol{\alpha}}(0) = 0 \tag{11}$$

Obviously, only the reaction velocity derivatives $\partial \mathbf{v}/\partial \boldsymbol{\alpha}$ need to be computed by computer algebra whereas all other ingredients are readily available. However, once the model is extended with additional constraints the computer algebraic approach tends to produce extremely lengthy solutions.

3. The state of the art method to compute sensitivities is so called automatic differentiation (AD) [40]. It has some similarity with the computer algebraic approach but is substantially more efficient because intermediately computed results are reused. Essentially, AD directly takes a program written in some programming language and generates a second program which computes the required derivatives. AD can be performed at the level of raw C-code [40] or at the higher level of equation based modeling language [41]. Due to the efficiency of the AD approach, a special model structure is not needed.

Because Modelica is very similar to a programming language, AD can be generally implemented for the Modelica language. Particularly, the AD tool described in [41, 42] generates a derived Modelica model from the original one for computing the sensitivities. The derived model can then be processed by the same Modelica compiler used for the original model.

8.2 Network Analysis

In recent years many network based systems analysis methods have been developed. Among them are graph algorithms for connectivity, path analysis and stoichiometry based algorithms for computing elementary modes or optimal flux solutions [43]. To carry out these methods, the knowledge of a biochemical network as represented by the stoichiometric matrix $\mathbf{N}$ in (1) is essential.

This is the application domain where the reaction based approach has clear advantages. Due to the special model structure the network structure can be readily extracted from an SBML specification. In contrast, a Modelica model must be based on a previously specified library in order to implement automatic procedures for network retrieval. However, it should be noticed that any information stored in the *rules* section of an SBML model experiences the same problem.

9 Modeling Tools

9.1 Graphical Network Representation

Although raw Modelica code is much easier to read than raw SBML code, both languages are not really intended for a human reader when large or even huge models are built up. In this case, graphical tools are needed to represent the models structure in an intuitively understandable way. Clearly, the most suitable graphical

representation for SBML models should orient on the chemical reaction network itself. The different data items are then connected to the nodes and edges of the network graph. Only the rules and event sections must be handled in a different way and cannot directly be associated to the network. Available tools for building up SBML models like CellDesigner [17] already take this approach.

The Modelica language already has an built in structural feature which makes it possible to generate a graphical representation for any type of system model automatically. This structure is given by the components and their connections together with the model hierarchy. For each instance of a model class, a graphical symbol shown as a box with connecting ports is generated. The user might also assign an icon with this box. Connectors are shown as lines. If a model is hierarchically composed from submodels, clicking on the model box will open a new display that shows the internal architecture. This concept is very familiar from the popular Simulink simulation tool.

However, it must be pointed out that the standard layout of graphical system models produced in that way might not be intuitively understandable in a specific application domain. As an example, looking at the model architecture in Fig. 9, it stands out that the substance pools are drawn in between the reactions but are connected to them in a non-intuitive way. This way of course represents the structure of the underlying mathematical equations but not their interpretation in their specific application domain.

Moreover, since connectors are usually drawn with horizontal and vertical lines, a network graph might look fine for an electrical network but not really for a biochemical network. Clearly these esthetic problems can be solved if special application specific tools are developed that generate Modelica code based on a previously fixed library. However, this causes a high implementation effort.

9.2 Availability

The specifications of Modelica and SBML are open source, i.e., everyone is invited to use them for his tools and model exchange. Moreover, both specifications are subject to intensive discussion and continuous improvements. On the one hand, SBML has established a large community in systems biology due to a wide range of free and commercial SBML-based tools, parsers and editors for model design, visualization, simulation of discrete and continuous systems and network analysis on various platforms [19].

On the other hand, due to the complexity of the Modelica language as well as the wide range of algorithms and concepts on which Modelica is based, there are still few compilers and programming environments for Modelica available [44]. Currently, free compiler such as OpenModelica [45] are less efficient compared to commercial programming environments like Dymola (Dynasim AB), which can also be used for constructing, simulating, and analyzing large models. Additionally,

the first steps were taken in implementing a library for modeling biological and biochemical systems [46].

10 Summary

The arguments given for the two different modeling approaches in the special field of biochemical network modeling are summarized in Table 1. It becomes obvious that neither of the two approaches is superior. As long as modeling projects are restricted to standard network structures, SBML certainly is the most direct way to reach the goals. If it is to be expected that multidisciplinary and hierarchical aspects become important, Modelica is the better alternative. Finally, it should be pointed out that future versions of SBML will incorporate more and more of the features that are already implemented in the more mature tool Modelica.

Due to the availability of an advanced SBML-Parser like LibSBML [47], transformation of an SBML model into a Modelica model based on a pre-given and well-suited biochemical library seems to be a straightforward task, at least for a basic subset of SBML-constructs. However, by converting Modelica models based on highly abstraction concepts to SBML models a lot of information may get lost. The computational biology community can benefit a lot from a tool adopting the first direction, as already large SBML-repositories exist. These models are then subject to consistency checking, validation, and visualization using various levels of abstraction blocks, simulation, and analysis with the help of many assisting tools for Modelica.

Table 1 Comparison of different features of SBML and Modelica

Feature[a]	SBML	Modelica
Standardization	+	−
Extendability	−	+
Hierarchical modeling	−	+
Multidisciplinary modeling	−	+
Spatially distributed modeling	−	(+)
Conceptual modeling	(+)	+
Intuitive graphical user interface	+	+
Available tools for:		
Model compilation	+	+
Network analysis	+	−
Sensitivity analysis	+	+

[a]Support
+ Yes
(+) Partly
− No

Acknowledgment This work was funded by German Ministry of Education and Research (BMBF) within the SysMAP Project (0313704).

References

1. von Bertalanffy L (1932) Theoretische biologie. Bd. 1: allgemeine theorie, physikochemie, aufbau und entwicklung des organismus. Gebrüder Borntraeger, Berlin
2. Kitano H (2001) Foundations of systems biology. MIT Press, Cambridge
3. Klipp E, Herwig R, Kowald A et al (2005) Systems biology in practice. Wiley-VCH, Weinheim
4. Heinrich R, Schuster S (1996) The regulation of cellular systems. Kluwer Academic, Dordrecht
5. Snoep JL (2005) The Silicon Cell initiative: working towards a detailed kinetic description at the cellular level. Curr Opin Biotechnol 16:336–343
6. Hucka M, Finney A, Sauro H et al. (2001) Systems biology markup language (SBML) level 1: structures and facilities for basic model definitions. www.sbml.org
7. Hucka M, Finney A, Sauro HM et al (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19:524–531
8. Vladimirescu A (1994) The SPICE book. Wiley, NY
9. Cellier FE (1991) Continuous system modeling. Springer, NY
10. Booch G (1990) Object oriented design with applications. Benjamin-Cummings, CA
11. Ashenden PJ, Peterson GD, Teegarden DA (2002) The systems designers guide to VHDL-AMS. Morgan Kaufmann, CA
12. Fritzson P (2003) Principles of object-oriented modeling and simulation with Modelica 2.1. Wiley-IEEE Computer Society, New York
13. Ginkel M, Kremling A, Nutsch T et al (2003) Modular modeling of cellular systems with ProMoT/Diva. Bioinformatics 19:1169–1176
14. Mann H (1994) Equation formulation and solution methods behind DYNAST – a multipurpose simulation tool. In: IMACS symp. on mathematical modelling MATHMOD, Vienna, pp 938–939
15. Bill E, Kent S, Thiru T et al (2007) Professional XML. Wiley-VCH, Heidelberg
16. Campbell KA (2002) Xerlin 1.1 user guide. www.xerlin.org
17. Funahashi A, Tanimura N, Morohashi M et al (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. BIOSILICO 1:159–162
18. Hoops S, Sahle S, Gauges R et al (2006) COPASI – a COmplex PAthway SImulator. Bioinformatics 22:3067–3074
19. Systems biology markup language. www.sbml.org
20. Hucka M, Hoops S, Keating SM et al (2008) Systems biology markup language (SBML) level 2: structures and facilities for model definitions. www.sbml.org
21. EUROSYSLIB Modelica libraries for embedded systems modelling and simulation. http://www.itea2.org/public/project_leaflets/EUROSYSLIB_profile_oct-07.pdf
22. Elmqvist H (1993) Object-oriented modeling and automatic formula manipulation in Dymola. In: SIMS'93, Scandinavian Simulation Society, Kongsberg, Norway
23. Maffezzoni C, Girelli R, Lluka P (1996) Generating efficient computational procedures from declarative models. Simul Pract Theory 4:303–317
24. Murota K (1987) Systems analysis by graphs and matroids. Springer, Berlin
25. Lieres E, Petersen S, Wiechert W (2004) A multi-scale modeling concept and computational tools for the integrative analysis of stationary metabolic data, vol 1. J Integr Bioinform
26. Cellier FE, Ernesto K (2006) Continuous system simulation. Springer, New York

27. Ascher UM, Petzold LR (1998) Computer methods for ordinary differential equations and differential algebraic equations. SIAM, Philadelphia
28. Leitold A, Hangos KM (2001) Structural solvability analysis of dynamic process models. Comput Chem Eng 25:1633–1646
29. Mattsson SE, Söderlind G (1993) Index reduction in differential-algebraic equations using dummy derivatives. SIAM J Sci Comput 14:677–692
30. Pantelides CC (1988) The consistent initialization of differential-algebraic systems. SIAM J Sci Stat Comput 9:213–231
31. Kremling A, Jahreis K, Lengeler JW et al (2000) The organization of metabolic reaction networks: a signal-oriented approach to cellular models. Metab Eng 2:190–200
32. Kremling A, Gilles ED (2001) The organization of metabolic reaction networks. II. Signal processing in hierarchical structured functional units. Metab Eng 3:138–150
33. Olsson H, Tummescheit H, Elmqvist H (2005) Using automatic differentiation for partial derivatives of functions in Modelica. In: Modelica 2005, Hamburg, Germany, pp 105–112
34. Saldamli L, Bachmann B, Fritzson P (2005) A framework for describing and solving PDE models in Modelica. In: Modelica 2005, Hamburg, Germany, pp 113–122
35. Tiller M (2001) Introduction to physical modeling with Modelica. Kluwer Academic, Norwell
36. Wiechert W, Treude M (2007) Teaching finite elements: an alternative approach using Modelica. In: EUROSIM, Lubljana
37. Plank J (1997) State events in continuous modelling and simulation – concepts, implementation and new methodology. In: ASIM/ARGESIM/SCS (ed) Advances in simulation
38. Ames WF (1977) Numerical methods for partial differential equations. Academic, New York
39. Hurlebaus J, Buchholz A, Alt W et al (2002) MMT – a pathway modeling tool for data from rapid sampling experiments. In Silico Biol 2:467–484
40. Griewank A, Walther A (2009) Evaluating derivatives: principles and techniques of algorithmic differentiation. SIAM, PA
41. Elsheikh A, Wiechert W (2008) Automatic sensitivity analysis of DAE-systems generated from equation-based modeling languages. Advances in automatic differentiation. Lecture notes in computational science and engineering, vol 64. Springer, Berlin
42. Elsheikh A, Noack S, Wiechert W (2008) Sensitivity analysis of Modelica applications via automatic differentiation. In: 6th international Modelica conference, Bielefeld, Germany
43. Bernhard P (2006) Systems biology: properties of reconstructed networks. Cambridge University Press, New York
44. Modelica and the Modelica Association. www.modelica.org
45. The OpenModelica Project. www.ida.liu.se/projects/OpenModelica
46. Nilsson EL, Fritzson P (2005) Biochemical and metabolic modeling and simulation with Modelica. In: BioMedSim. Linköping, Sweden
47. Bornstein BJ, Keating SM, Jouraku A et al (2008) LibSBML: an API library for SBML. Bioinformatics 24:880–881

Adv Biochem Engin/Biotechnol (2010) 121: 139–162
DOI: 10.1007/10_2009_63

Published online: 12 February 2010

Impact of Thermodynamic Principles in Systems Biology

J.J. Heijnen

Abstract It is shown that properties of biological systems which are relevant for systems biology motivated mathematical modelling are strongly shaped by general thermodynamic principles such as osmotic limit, Gibbs energy dissipation, near equilibria and thermodynamic driving force. Each of these aspects will be demonstrated both theoretically and experimentally.

Keywords Black box kinetics, Gene regulation, Lin-log kinetics, Metabolic networks, Model reduction, Pseudo-steady state, Stoichiometry, Systems biology, Thermodynamics

Contents

J.J. Heijnen
Department Biotechnology, Bioprocess technology, Delft University of Technology, Julianalaan 67, 2628 BC Delft, The Netherlands
e-mail: j.j.heijnen@tudelft.nl

1 Introduction

Thermodynamic principles apply to all physical and chemical systems, including biological systems. In this chapter it will be shown how these principles shape their properties, especially from a quantitative model based, systems biology, point of view.

2 Thermodynamic Principles in Mathematical Models of Biological Systems

The key aspect of a living cell is the formation of new cells, called growth.

Growth requires that a cell produce each of the molecules present in the newly formed cells. This occurs in a large and complex metabolic (reaction) network. This network is composed of many reactions, which consume and produce small molecules, called metabolites. Each reaction is catalysed by a specific enzyme, which is under genetic/environmental control. Prediction of growth requires a mathematical model of this network for which the fundamental equations are the mass balances of intracellular metabolites. In vector notation:

$$\frac{\mathrm{d}X}{\mathrm{d}t} = SV\,(e, X, p) - \mu X, \tag{1}$$

where X is the vector containing the individual intracellular metabolites X_j.

V is the vector of the rates of enzyme catalyse reactions, with v_i the rate of reaction catalysed by enzyme present at an amount e_i. The rate v_i depends on the amount of enzyme present, e_i, on the kinetic effect of metabolites X_j involved (e.g. substrate, product and possible allosteric effectors) and the parameters p (e.g. $V^{\max}$, affinities, Hill coefficient etc.). S is the so-called stoichiometric matrix which represents the structure of the reaction network. Its rows represent the metabolites, its columns the reactions [44]. The term μX is the so-called dilution term.

Equation (1) requires information on the dynamic behaviour and values of metabolite concentrations, on the values of stoichiometric coefficients, on enzyme levels resulting from genetic regulation and on the shape/algebraic nature of the

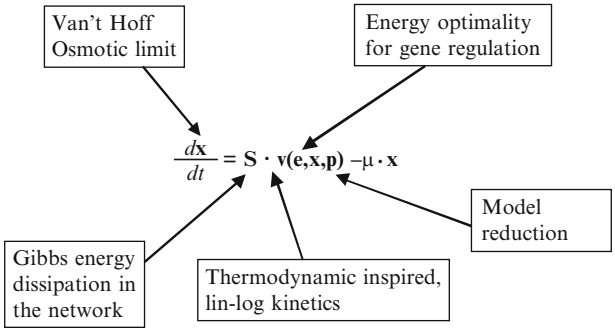


Fig. 1 Impact of thermodynamic principles in systems biology

enzyme kinetic relations, and (1) is the basis of parameter estimation from experimental data and the associated need for model reduction.

In the following sections we will show that thermodynamic principles can be used to shed light on this information (Fig. 1):

- Metabolite concentration levels (X_j) and their control mechanisms
- The stoichiometry of growth (S)
- The genetic regulation of enzyme levels (e_i)
- Principles of model reduction
- The kinetics of enzyme catalysed reactions based on thermodynamic driving force

3 The Osmotic Limit Dictates Low Concentrations of Intracellular Metabolites

Cells have a genome which contains about 5,000 genes. These genes code for about 5,000 proteins, of which about 2,000 are enzymes. Therefore, in a cell, one can expect about 2,000 different metabolites which are small molecules (e.g. metabolites in central metabolism and in pathways for amino acid, nucleotide, lipid and carbohydrate/cell wall synthesis). Many of these metabolites are negatively charged (having phosphate and carboxylate groups) and therefore there are also considerable concentrations of counter cations (K^+, Mg^{2+}). The sum concentration of all these small molecules is limited by a thermodynamic property called osmotic pressure [1]. Because cells contain a cell membrane that is water permeable, the presence of a high intracellular sum concentration of membrane impermeable anionic/cationic small molecules leads to a water activity inside cells that is lower than outside. This creates a flow of water into the cell, leading to increase of intracellular pressure. The water inflow stops when the pressure has reached the

osmotic pressure. According to van 't Hoff the osmotic pressure is linear in the sum concentration C of all intracellular metabolites ($P_{\text{osm}} = CRT$). For example, for C = 1 mol/L it follows that $P_{\text{osm}} = 25$ bar ($C = 1$ mol/L $= 1{,}000$ mol/m^3, $R = 8.314$ J/mol K, $T = 298$ K gives $P_{\text{osm}} = 24.8 \times 10^5 \text{N/m}^2 = 25$bar). From a mechanical point of view, this pressure is counteracted by the mechanical strength of the cell membrane/cell wall, which is obviously bound to physical limits. Therefore there must exist a maximal sum concentration of small intracellular molecules. Assuming a limit of 25 bar gives for this maximal sum $C \approx 1.0$ mol/L, which gives a sum concentration of organic (anionic) metabolites of order 0.5 mol/L. Assuming the presence of about 1,000 different metabolites in cells gives, for the average intracellular metabolite concentration X_j, a value of about 10^{-3} mol/L. Of course, there will be a wide distribution of concentrations, so we can expect an intracellular concentration range of 10^{-2} to 10^{-4} mol/L, which is equivalent to 20–0.20 μmol/g dry biomass. These values are indeed found as shown in Table 1.

4 Consequences of Low Metabolite Concentrations from a Systems Biology Point of View

The general property of low intracellular metabolite concentrations has very important consequences at system level.

A first consequence is the near *absence of spontaneous reactions*. Usually the metabolic network is considered to be totally enzyme catalysed and one assumes implicitly that non-enzymatic reactions (which occur spontaneously) are absent. Given the multitude of reactive molecules inside cells, one would expect much more spontaneous reactions. Such reactions would be disadvantageous because they are not under genetic control and they cause loss of material. The key to suppress such reactions, in favour of enzyme catalysed reactions, is to have a very low metabolite concentration (which kinetically "kills" the rate of a spontaneous reaction) in combination with matching high affinities of enzymes. This is indeed found. So one could state that the osmotic limit enforces *high affinity enzymes*.

A second general consequence is the *need of active export*.

Many biological systems are used in industrial processes (antibiotics, fuels, amino acids, organic acids etc.). From an economic point of view, one aims at high ($\approx$ 1 M) extracellular concentrations of product. This implies that the final step of product metabolism, export, has to deal with an unfavourable concentration gradient of about 10^{-3} M inside and 1 M outside. Clearly, this requires active export [6].

Another general aspect of metabolites is a *fast, order of seconds, turnover time* (t.o.t.) of each metabolite. The t.o.t. of a metabolite X_j is defined as $(\text{t.o.t.})_{X_j} = \frac{X_j}{V_{\text{sum}}}$, with X_j the metabolite concentration and V_{sum} the sum of all production rates of this

Table 1 Intracellular metabolite concentrations and turnover time in glucose limited aerobic cultures of several organisms (*Saccharomyces cerevisiae* from [2], *Penicillium chrysogenum* from [3, 4] *E. coli* from [5])

Metabolites	Intracellular level (μmol/gDW)			Turnover time (s)		
	P. chrysogenum	*S. cerevisiae*	*E. coli*	*P. chrysogenum*	*S. cerevisiae*	*E. coli*
Central metabolites						
G6P	4.64	5.2	1.42	23.3	17	3.6
F6P	0.71	1.4	0.38	5.7	7.3	1.2
T6P	0.55		0.13	47.8		NA
M6P	1.95		0.48			NA
6PG	0.25	0.48	0.10	3.7	4.5	1.1
Mannitol-1P			0.99			NA
G3P		0.13	0.17		57	13.1
FBP	0.9	0.64	0.82	7.2	3.2	2.5
F2,6bP	0.01		0.35			NA
2PG+3PG	0.59	2.8	1.65	2.3	6.6	2.5
PEP	0.24	2.3	1.61	0.9	5.7	2.7
Pyruvate	0.22	1.1	0.75	0.9	1.7	1.5
α-Ketoglutarate	2.05		0.31	22.1		0.6
Succinate	0.23	4.0	2.65	3.3	20	8.9
Fumarate	0.65	0.85	0.22	13.0	4.1	0.7
Malate	3.33	7.3	0.94	19.0	30	2.8
Amino acids						
Alanine	21.7	32	1.34	269	3,268	76.7
Asparagine	1.5	4.7	0.58	459	1,142	81.7
Aspartate	16.3	21	2.57	717	577	35.0
Glutamate	53.0	170	74.69	658	1,112	229.0
Glutamine	28.7	64	6.14	1,243	2,401	80.0
Glycine	2.1	2.9	1.51	244	247	31.0
Histidine	0.72	6.0	0.15	432	3,141	53.8
Isoleucine	0.33	1.6	0.11	111	140	12.9
Leucine	0.73	1.0	0.36	131	125	27.1
Lysine	1.2	4.1	1.21	356	619	119.7
Methionine	0.14	0.20	0.05	58.8	66	10.5
Phenylalenine	0.19	1.6	0.13	61.2	430	23.8
Proline	0.95	3.9	0.66	206	925	101.4
Serine	5.7		0.53	453		8.0
Threonine	5.9	4.0	0.47	758	220	29.3
Tryptophan	0.11	0.51	0.02	130	788	11.9
Tyrosine	0.26	1.6	0.18	145	832	44.3
Valine	2.1	10	0.51	243	490	40.9
Ornithine		4.1	0.49		502	49.1
Adenine nucleotides						
ATP	7.39	7.0	5.95	NA	1.4	2.0
ADP	1.03	1.3	2.31	NA	0.25	0.8
AMP	0.27	0.6	0.91	NA	3.1	9.4

metabolite. Because X_j is low and V_{sum} can be high, one indeed finds t.o.t. of the order of seconds (Table 1).

These *fast t.o.t.* have several important consequences:

- Considering product formation, where a substrate molecule is processed along a multistep pathway to the secreted product, it follows that the time between substrate entrance and product leaving the cell is only of order minutes. Clearly cell factories follow the *just in time* principle.
- Considering the *metabolite mass balances* (1), we can safely neglect:
  - The dilution term μX_i, which is orders of magnitude smaller than the synthesis term SV.
  - The dynamic term $\frac{dX_i}{dt}$ for time scales larger than minutes (which follows from t.o.t. of order seconds). This leads to *pseudo-steady state*.
- This *pseudo-steady state* property, which is a direct consequence of the low metabolite levels due to an osmotic limit, is one of the most important network properties. It allows one to write for the metabolite mass balances:

$$SV = 0 \tag{2}$$

  These balance equations are the basis of the well-known stoichiometric analysis of metabolic networks. We should also realise that, due to the pseudo-steady state property of the metabolic reaction network, these balances also apply to dynamic situations which allow one to formulate so-called black box stoichiometric/kinetic models which are reliable in a wide range of conditions (see also "model reduction").
- A final consequence is *the need for control mechanisms* on the production/consumption of each intracellular metabolite. Cells are, in their natural environment, continuously exposed to perturbations which change the rate of synthesis/consumption of metabolites. Given the low concentration of a metabolite, such a perturbation leads to very quick (second time scale) and drastic (up or down) changes in metabolite concentrations, which propagate through the network leading to potential damaging system responses. Control of metabolite levels is needed to limit these effects, and indeed such control mechanisms are widely found in biological systems and, most interesting, they operate at proper time scales. Most well known are *allosteric feed back inhibition* (e.g. in amino acid synthesis pathways), and *allosteric feed forward activation* (e.g. in glycolysis) mechanisms which operate within seconds. This is exactly the time scale expected from the t.o.t. for metabolite levels. The other mechanisms are slower. The *post translational modification* mechanisms (adenylation, (de)phosphatation, ...) take in the order of minutes, consume ATP and interconvert active/inactive enzyme, but do not change the total enzyme amount. The *genetic mechanisms* (induction, repression) take in the order of tens of minutes and change the amount of enzyme.

5 Thermodynamic Approach to Obtain Network Stoichiometry and Fluxes

To obtain the network fluxes (and therewith the stoichiometry of the network) requires one to solve (2). These balances put linear constraints on the reaction/uptake/secretion rates. The number of rates in a realistic network is typically several hundred; however the number of metabolite mass balances in (2) (= constraints) is also large. A general problem is the ATP-balance which contains uncertain ATP stoichiometric parameters (P/O ratio, unknown growth related ATP $(= K_X)$ and the unknown growth unrelated ATP (m_{ATP}). Van Gulik and Heijnen [7] and Van Gulik et al. [8] have shown how in vivo values for these ATP-parameters can be obtained using extensive experiments. For many organisms this ATP-information is not available. This means that the number of equations in (2) is always at least two lower than the number of rates (underdetermined). This means that solving all rates needs the specification of at least two experimental rates. Without this experimental information one cannot predict stoichiometry. Another approach which has received considerable attention in the past decade, and which aims to predict both rates and stoichiometry of networks, is constraint based modelling [43]. This approach uses an optimality criterion (e.g. maximal biomass yield) to obtain a solution of the underdetermined (2). However, close inspection reveals that this method still requires the above-mentioned experimentally based information:

- Specification of the uncertain ATP stoichiometric coefficients (P/O, growth related and unrelated maintenance values). This information is needed to make stoichiometry predictions!!
- Kinetic information, such as an experimentally determined substrate uptake rate or a maximum O_2-uptake rate. This is needed to calculate fluxes and, e.g. $\mu^{\max}$.

Together with this experimental information the optimality criterion forces ATP requiring processes such as futile cycles to zero and therewith one obtains a unique flux solution and therewith stoichiometry. When the above-mentioned (ATP and kinetic) experimental information is not available, constraint based modelling does not lead to a unique flux solution.

Thermodynamics offers an alternative, more widely applicable, approach to solve the network stoichiometry and fluxes for arbitrary organisms.

5.1 Thermodynamic Approach to Stoichiometry

Thermodynamics allows one, for a given specific growth rate μ (under substrate limited growth in absence of a non-catabolic product, hence only growth), to calculate all uptake/secretion rates. Herewith, all yields are also available (yield is ratio of rates). Heijnen and Van Dijken [9] and Heijnen [10] apply this approach

to any heterotrophic growth system under substrate limited condition. The only information needed is the nature of carbon source, electron donor, electron acceptor, N-source and temperature.

In this method, the ATP-balance in the network equation (2) is replaced by a Gibbs energy balance on all uptake and secretion rates q_i. We can define q_i as a biomass specific rate in mol of *i*/*h* per Cmol biomass. The quantity 1 Cmol of biomass (which has the average composition $C_1H_{1.8}O_{0.5}N_{0.2}$) is the amount (24.6 g organic dry matter) of biomass which contains 12 g of carbon (= 1 mol C-atom).

Compounds taken up have negative q_i values, secreted compounds have positive q_i-values. The Gibbs energy balance follows then as

$$\sum q_i \Delta G_{f_i}^{01} + q_G = 0. \tag{3a}$$

Here $\sum q_i \Delta G_{f_i}^{01}$ is a negative quantity (second law of thermodynamics) and is the total biomass specific rate of Gibbs energy of conversion. q_G is the Gibbs energy produced, which follows from (3a). The second law requires $q_G > 0$.

$\Delta G_{f_i}^{01}$ is the Gibbs energy of formation of compound i at standard condition (1 M, 298 K and at pH = 7.0). In principle, one needs to take actual concentrations into account, but this leads only in special cases to significant changes in q_G [10].

The key to the use of (3a) is to obtain a relation for q_G. Because cells require Gibbs energy for growth and maintenance, we can write a Herbert–Pirt type of relation for q_G, which expresses that (in absence of non-catabolic product) the cell needs Gibbs energy for growth and maintenance:

$$q_G = \frac{1}{Y_{GX}^{\max}} \mu + m_G. \tag{3b}$$

For m_G and $\frac{1}{Y_{GX}^{\max}}$, correlations have been established [9, 11].

5.1.1 Gibbs Energy for Maintenance

All living systems need to generate Gibbs energy for their maintenance (which represents all processes where energy is needed for example to repair degradation and export compounds that entered due to membrane leakage, etc.). Because living cells have similar membranes and composition, it can be assumed that different cells require a similar amount of energy expenditure for maintenance. It has indeed been found that Gibbs energy needed for maintenance is very similar for a large range of microorganisms and only depends on absolute temperature (*T*) [11].

The following *correlation* has been found for m_G $\left(\text{in } \frac{\text{kJ Gibbs energy}/h}{\text{CmolX}}\right)$:

$$m_G = 4.5 \exp\left(\frac{69{,}000}{R}\left(\frac{1}{298} - \frac{1}{T}\right)\right). \tag{4}$$

This correlation shows that at 25°C (298 K) 1 Cmol of cells (≈ 25 g dry matter) spends, and therefore needs to generate, 4.5 kJ of Gibbs energy per hour to cover the energy for maintenance. Also there is a steep temperature effect: for each 8°C temperature increase m_G doubles!!

This correlation applies to aerobic/anaerobic systems and arbitrary electron donors/acceptors. The Gibbs energy is generated from the catabolic reaction, so the maintenance reaction equals the catabolic reaction. For specific cases, one can always set up the catabolic reaction for 1 mol donor and obtain the catabolic Gibbs energy of reaction for 1 mol donor, called $\Delta G^{01}_{\text{cat},D}$, which is negative and is in kJ of Gibbs energy per 1 mol donor consumed in the catabolic reaction. It is then clear that we can write for the substrate (or donor) consumption $m_s\left(\frac{\text{in mol substrate}/h}{\text{CmolX}}\right)$ that must be catabolised for maintenance:

$$m_s = \frac{m_G}{\Delta G^{01}_{\text{cat},D}}. \tag{5}$$

Table 2 shows examples of catabolic reactions consuming 1 mol donor, and the Gibbs energy of catabolism per mol consumed donor $\left(\Delta G^{01}_{\text{cat},D}\right)$. It is obvious that $\Delta G^{01}_{\text{cat},D}$ can be two orders of magnitude different, dependent on the specific catabolism.

For example, consider *Saccharomyces cerevisiae* at 30°C. It follows from (4) that

$$m_G = \frac{7.1\,\text{kJ/h}}{\text{CmolX}}$$

Under *aerobic* conditions using glucose as substrate, catabolism $\Delta G^{01}_{\text{cat},D} = -2{,}843.1$ kJ per mol glucose. This gives

$$m_s = \frac{7.1}{(-2{,}843.1)} = -0.0025\,\frac{\text{mol glucose}/h}{\text{CmolX}}$$

Also $m_{O_2} = 6 \times m_s = -0.015\ \frac{\text{mol O}_2/h}{\text{CmolX}}$. Under *anaerobic* conditions the catabolic ethanol forming reaction shows $\Delta G^{01}_{\text{cat},D} = -225.4$ kJ per mol glucose. This gives $m_s = \frac{7.1}{-225.4} = -0.0315\,\frac{\text{mol glucose}/h}{\text{CmolX}}$ and $m_{\text{ethanol}} = 0.063\,\frac{\text{mol ethanol}/h}{\text{CmolX}}$.

Table 2 Catabolic reactions and their Gibbs energy of reaction, $\Delta G^{01}_{\text{cat,D}}$

Catabolic reactions and their Gibbs energy $\Delta G^{01}_{\text{cat,D}}$(kJ/mol donor)		
Donor	Catabolic reaction	$\Delta G^{01}_{\text{cat,D}}$
Glucose	$C_6H_{12}O_6 + 6O_2 \rightarrow 6HCO_3^- + 6H^+$	−2,843.1
Ethanol	$C_2H_6O + 3O_2 \rightarrow 2HCO_3^- + 2H^+ + 1H_2O$	−1,308.9
Glucose	$C_6H_{12}O_6 + 2H_2O \rightarrow 2C_2H_6O + 2HCO_3^- + 2H^+$	−225.4
Methanol	$CH_4O + 1.20NO_3^- + 0.20H^+ \rightarrow 0.60N_2 + HCO_3^- + 1.60H_2O$	−649.4
Iron (2+)	$Fe^{2+} + \frac{1}{4}O_2 \rightarrow Fe^{3+} + \frac{1}{2}H_2O$ (pH = 1.85)	−33.9
Acetate	$C_2H_3O_2^- + H_2O \rightarrow HCO_3^- + CH_4$	−31.0

So, the same organism has widely different m_s values due to different catabolism, but still has the same Gibbs energy need for maintenance!!

5.1.2 Gibbs Energy for Growth

$\frac{1}{Y_{GX}^{max}}$ is the amount of Gibbs energy needed to synthesise 1 CmolX (in kJ/CmolX). This amount has been found to depend only on two factors [9, 10]: first, the nature of the carbon source for *heterotrophic growth* and, second, the nature of electron donor for *autotrophic growth*.

Regarding the first factor, the nature of the carbon source for *heterotrophic growth*, more Gibbs energy is needed when the carbon source has a smaller number (C) of C-atoms and when its degree of reduction per C-atom (γ) is different from the degree of reduction of biomass (≈ 4.2). The explanation is straightforward that synthesis of biomass monomer molecules (which contain order 6 carbon atoms with $\gamma \approx 4$) requires more C–C-coupling and redox reactions for C-sources with a low number of C-atoms and which need reduction or oxidation because γ of the C-source differs from 4.2. These extra reactions lead to a higher Gibbs energy need. For heterotrophic growth this intuition is quantified in (6a), which is a *correlation*:

$$\frac{1}{Y_{GX}^{max}} = 200 + 18(6 - C)^{1.8} + \exp\left[|3.8 - \gamma|^{0.32} \times (3.6 + 0.4C)\right]. \qquad (6a)$$

This correlation shows that to synthesise 1 Cmol biomass one needs between $\approx$ 236 and 1,087 kJ Gibbs energy (dependent on C-source, e.g. 236 for glucose ($\gamma = 4$, $C = 6$) and 1,087 for CH_4 ($\gamma = 8$, $C = 1$).

Regarding the second factor, the nature of electron donor for *autotrophic growth*, in autotrophic growth CO_2 is the C-source which must be reduced to biomass using electrons from the electron donor. The Gibbs energy needed follows from (6b), which is also a *correlation* [9]:

$$\frac{1}{Y_{GX}^{max}} = 1,000(-\text{RET}) = 3,500(+\text{RET}). \qquad (6b)$$

The nature of the electron donor determines the absence (−RET) or need (+RET) of *R*everse *E*lectron *T*ransport. For several electron donors (e.g. NH_{4^+}, $NO_2{}^-$, Fe^{2+}) the reduction of CO_2 to biomass is not feasible thermodynamically. Therefore cells spend extra Gibbs energy to make the redox potential of electrons obtained from the available donor more negative (e.g. in the production of NADPH from the donor electrons, in a process called RET, NADPH is then used to reduce CO_2 to biomass). This extra Gibbs energy is very considerable (compare in (6b) 3,500 and 1,000). For H_2 as electron donor, (−RET), this problem does not exist (sufficient negative redox potential) and the Gibbs energy need is 1,000 kJ/CmolX (6b).

Summarising, the Gibbs energy needed to make 1 Cmol biomass ranges, dependent on C-source and electron donor, is between 200 and 3,500 kJ and does not depend on the type of electron acceptor. When the carbon source, electron donor and temperature are known, the correlations (4, 6a/6b) give the coefficients $\frac{1}{Y_{GX}^{\max}}$ and m_G in the Gibbs energy Herbert Pirt relation (3b), which completes the Gibbs energy balance (3a). This linear relation can be combined with all metabolite mass balances specified in (2), using the stoichiometric matrix S where the ATP-balance is also absent (due to unknown P/O, K_x, etc.). A constraint of minimal Gibbs energy dissipation will put futile cycles to zero and parallel pathways are also resolved. This set of linear balances gives, for any selected μ, all rates in the network (reaction/uptake/secretion)!! With these known rates all yields are known such as the biomass yield $Y_{SX} = \frac{\mu}{q_s}$ or yield of catabolic products.

This thermodynamic approach has been shown to predict biomass yield with 10–15% error for a wide variety (aerobic/anaerobic/hetero/autotrophic) of microbial systems where Y_{SX} spans a range of near *two orders of magnitude* [9].

5.2 Thermodynamic Approach for Maximal Growth Rate, $\mu^{\max}$

Microorganisms show a very large range (0.001–1 h^{-1}) in $\mu^{\max}$-values and it is relevant to understand why this is so!! A simple approach was proposed [10] which reproduces most of this range. The concept is that cells during evolution ultimately become limited in their *energy production capacity*. Most organisms generate energy by electron transport phosphorylation. This occurs by electron transport proteins embedded in membranes. Because membranes are space limited for protein embedding it is logical to propose that cells have evolved to a maximal electron transport capacity $q_{el}^{\max}\left(\text{in } \frac{\text{mol electrons}/h}{\text{CmolX}}\right)$ which depends mainly on temperature. Also it is known that smaller organisms have higher maximal growth rates, e.g. *E. coli* $\mu^{\max} = 2\ h^{-1}$, *Saccharomyces cerevisiae* $\mu^{\max} = 0.4\ h^{-1}$ and tissue cell cultures $\mu^{\max} = 0.04\ h^{-1}$. This is in line with the smaller surface/volume ratio ($\approx 6/d$, with cell diameter d) which leads to a membrane surface area, hence maximal electron capacity and $\mu^{\max}$ which is inverse to the cell diameter and which is indeed largely observed (e.g. *S. cerevisiae* has a 5× larger cell diameter compared to *E. coli*). This concept has in addition been inspired by the observation that in *E. coli*, for different substrates which lead to different $\mu^{\max}$, the $q_{O_2}^{\max}$-value is nearly constant [12].

The following correlation was proposed:

$$q_{el}^{\max} = 3\ \exp\left[\frac{69{,}000}{R}\left(\frac{1}{298} - \frac{1}{T}\right)\right]. \tag{7a}$$

This electron capacity determines the maximal production rate of Gibbs energy by catabolism ($q_G^{\max}$):

Table 3 Estimated μ^{max} values for different microbial catabolic classes at 25°C based on limiting Gibbs energy production

Catabolic classes	$\Delta G^{01}_{cat,\,D}$ (kJ/mol donor)	γ_D (mol electron/mol donor)	μ^{max} (h^{-1})
Aerobic/glucose	−2,843.1	24	1.5
Aerobic/acetate	−844.2	8	0.70
Anaerobic/(acetate→CH_4)	−31.0	8	0.015
Aerobic/Fe^{2+} oxidation	−38.6	1	0.030
Aerobic/nitrification	−274.8	6	0.040

$$q_G^{max} = q_{el}^{max}\frac{\left(-\Delta G^{01}_{cat,D}\right)}{\gamma_D}. \tag{7b}$$

Here, γ_D is the number of electrons released in catabolism of 1 mol donor and $\Delta G^{01}_{cat,D}$ is the catabolic Gibbs energy per mol donor. This maximal Gibbs energy sets μ^{max}according to (3b).

Combination of (3b) and (7b), (2) and using the correlations (4), (6a, b) and (7a) for $m_G, Y^{max}_{GX}, q^{max}_{el}$and the available value for $\Delta G^{01}_{cat,D}$ and γ_D immediately allows one to calculate μ^{max}-values for any growth system. These values agree reasonably with known values. Table 3 shows that this simple approach can explain a 100-fold difference in μ^{max}.

Some final remarks:

- This thermodynamic approach uses only three correlations (4, 6a/6b, 7), is simple and general and gives maximal q_i-rates and stoichiometry.
- Effect of temperature is included (in maintenance and μ^{max}).
- The three correlations are based on a wide range of experimental microbial growth systems and reflect that similar biochemical pathways are used (unity of biochemistry). When the predicted μ^{max} or/and stoichiometry are very different from experimental values, this indicates unusual anabolic and/or catabolic routes which might be novel. So this method can act as a filter for unusual behavior of biological systems.

6 Prediction of Gene Regulation of Enzymes Using Energy Optimality

Gene regulation seems at first glance highly complicated. For example, enzyme induction upon exposure to a new catabolic substrate involves many mechanisms between signal transduction, gene expression and production of enzyme for the new catabolic pathway. Model based prediction of enzyme induction therefore seems hopeless. However one could expect that evolution has fine tuned the available

regulation mechanisms such that growth yield is optimal. This optimality principle was tested using mixed substrates with *Saccharomyces cerevisiae*, which was chemostat (aerobic, substrate limited) cultivated at a dilution rate $D = 0.1\ h^{-1}$. Different feed mixtures of glucose and ethanol as substrates were applied between 100% glucose and 100% ethanol) [7, 13, 14].

Growth on glucose differs from that on ethanol. On 100% ethanol the cell has:

- Fully induced:
  - Glyoxylic acid pathway enzymes (isocitrate lyase (ICL) and malate synthase (MS))
  - Gluconeogenic enzymes PEP-carboxykinase (PEPCK) and F16 Bisphosphate (F16BP-ase)
- Fully repressed (or inactivated):
  - Pyruvate kinase/pyruvate carboxylase (PK/PYC)
  - F16 bis P-kinase (PFK)

It is obvious that, under substrate limited condition for mixed substrates, the induction/repression pattern of these enzymes is determined by the residual ethanol and glucose concentration, which would be the basis of a complicated gene regulation mechanism and model to predict the occurrence and concentration of these enzymes as function of the residual ethanol and glucose concentration.

A different approach, based on energy optimality of gene expression, was followed. Using a metabolic stoichiometric model endowed with experimentally obtained stoichiometric values for the ATP-balance (P/O, K_x, m_{ATP}), linear programming was applied with maximal biomass production (or the equivalent minimal energy consumption) as optimality criterion. It was possible to calculate, for each glucose/ethanol supply ratio (which is virtually equal to the ratio of their consumption rates due to the low residual ethanol and glucose concentration) the optimal rates of all reactions in central metabolism. It was observed that, for the above inducible enzymes, clear predictions were made on their need as function of increasing ethanol fraction. This approach predicted:

- The ethanol/glucose feed ratio where a particular enzyme started to be induced
- The enzyme amount then increased linear with increasing ethanol fraction

These predictions were qualitatively, but more surprising also quantitatively, validated using the wild type yeast [13, 14]. Later additional validation was performed with null-mutants in the above enzymes, leading to a predicted maximal ethanol uptake rate of each mutant.

This prediction was again quantitatively confirmed [15]. This example clearly indicates that gene regulation mechanisms might have evolved to provide maximal biomass yield (giving a competitive edge). This maximal biomass yield is the same as Gibbs energy optimality because enzyme induction is such that futile cycles are avoided (e.g. simultaneous activity of FBP-ase/FPK or pyruvate carboxylase/PEP

carboxy kinase). To the authors knowledge this is one of the earliest and most successful examples of experimentally demonstrated thermodynamic optimality of living cells.

7 Thermodynamics Based Model Reduction in System Biology

Mathematical models of biological systems are useful to design processes and/or to redesign organisms using the rec-DNA tool box. Model reduction is an important issue, given the complexity of biological systems. Model reduction aspects will be discussed for two categories of mathematical models.

7.1 Black Box Models for Design of Biotechnological Processes: From Complexity to Simplicity Due to Pseudo-Steady State Coupling

We have observed that the thermodynamically based osmotic limit leads to very low metabolite concentrations in intracellular metabolism. The immediate consequence is that at process time scales larger than about 10 min (as occurs in bath, fed batch processes) the pseudo-steady state condition for all intracellular metabolites holds. The consequence of this condition is that all uptake and secretion rates are directly coupled. This *pseudo-steady state coupling* can be evaluated by linear rearranging the metabolite mass balances (2). Usually, in the absence of by-products, there are only a few (two to three) degrees of freedom, meaning that all uptake/secretion rates can be written as linear combination of only two to three rates (usually growth rate, product formation rate, maintenance). A prime example of such a linear relation is the Herbert–Pirt relation for substrate distribution:

$$q_s = \frac{1}{Y_{SX}^{max}}\mu \; + \; \frac{1}{Y_{SP}^{max}} q_p + m_s. \tag{8a}$$

Furthermore, the substrate uptake relation is usually a hyperbolic relation in the limiting substrate concentration:

$$q_s = q_s^{max}\left(\frac{C_s}{K_s + C_s}\right). \tag{8b}$$

A final relation, which holds generally under *single substrate limited condition* is that there is a unique relation between q_p and μ:

$$q_p = f_p(\mu). \tag{8c}$$

The three equations (8a, b, c) have only *one degree of freedom* (e.g. C_s or μ). All the other uptake/secretion rates $(O_2, CO_2, NH_{4+}, \text{water}, H^+, \text{heat}, \text{etc.})$ can be obtained from q_s, μ and q_p and the conservation relations (elements and charge).

This black box approach for kinetic modeling of biological processes is only possible due to the general biological system property of low intracellular metabolite concentrations, which has its origin in a thermodynamic property (osmotic limit).

Examples of this black box approach can be widely found. A very nice example is the model for penicillin production [8, 16] and the model developed for biological P-removal using mixed cultures in a cyclic process [17]. These black box models show that highly complex biological systems, comprising thousands of reactions, can be effectively modeled with a reduced model containing only about 6–12 parameters. These models are the basis of process design. This simple behavior of complex biological systems (from complexity to simplicity!!) is the direct result of the metabolite pseudo-steady state property which results from an osmotic limit!

7.2 Metabolic Reaction Network Models to Redesign Organisms: Reducing Complexity

Genetic intervention in metabolic reaction networks is possible in many ways by, e.g. changing enzyme levels, changing enzyme kinetics (e.g. abolish feed back inhibition), introducing different reactions, etc. The problem is that, following such interventions, the prediction of changes in secretion or uptake rates is very difficult due to the complex nature of the network structure and of the highly non-linear kinetics of the enzymes and gene regulation mechanisms. It is generally recommended that a mathematical model of a metabolic (reaction) network can help to select rationally genetic engineering targets in the redesign of organisms.

Building such a model is essentially straightforward and based on (1). Usually matrix S is known with high confidence. Uncertainties in S (e.g. cycles, parallel reactions, etc.) can be addressed using ^{13}C approaches [18].

A much bigger problem is to obtain kinetic relations for each enzyme, meaning the function $v\,(e, x, p)$ in (1) for each enzyme.

Traditionally, in vitro obtained kinetic functions and parameters have been used. Here two problems arise. First, it was found that in vitro kinetics do not reflect in vivo kinetics [19]. Second, for many enzymes in vitro kinetics are not available. The only solution therefore is to perform experiments with whole cells to obtain in vivo kinetic behavior of all enzymes simultaneously. Here rapid pulse experiments as pioneered by the groups of Reuss [20, 21] and Heijnen [22] offer advantages, but the challenge is the parameter identifiability problem which calls for model reduction as shown by Nikerel et al. [23–25].

Model reduction in biological systems has a thermodynamic basis as will be outlined below.

7.2.1 Pseudo-Equilibrium Reactions

A first approach for model reduction is to replace a kinetic function by an equilibrium relation. This is possible when the enzyme catalysed reaction is so rapid that the reaction remains very close to equilibrium, even when the reaction rates increase.

A strong indication of this situation is that the so-called mass action ratio of a reaction remains nearly constant (independent from flux) and close to the reported equilibrium constant. Many examples have been found which show this behaviour in *Penicillium chrysogenum* [3, 4] and in *Saccharomyces cerevisiae* [26] such as for phospho glucose isomerase, enolase, phosphoglycerate mutase, phosphoglucomutase, fumarase.

Recently a near equilibrium reaction ($F6P + \mathrm{NADH} + H^+ \rightarrow \mathrm{mannitol} - 1 - P + \mathrm{NAD}^+$)was used as a heterologous sensor reaction to obtain the cytosolic NAD/NADH ratio in *Saccharomyces cerevisiae* [2], even under dynamic conditions. It was found that the cytosolic NAD/NADH ratio was about 100. In contrast, a ratio of 4 was obtained by analysing the total amounts of NADH and NAD in whole yeast cells. The large difference is due to the fact that nearly all NADH is present in the mitochondria. Even more important is that, using the cell average NAD/NADH ratio, the ΔG_R of glycolysis between F16BP and (2+3 PG) was > 0, which is impossible. Use of the cytosolic value of 100 leads to ΔG_Rwhich is slightly negative, as expected. It can be expected that metabolism shows much more near equilibrium reactions. However here we need much more accurate data on in vivo equilibrium constants on a genome wide scale!!

7.2.2 Pseudo-Steady State Lumping

We have seen that many metabolites have low concentration levels (due to an osmotic upper limit) which leads to very low turnover times ($<$1 s) of metabolite pools. This allows, even in pulse experiment at a 300-s time scale, lumping of the synthesis and consumption reactions of a fast metabolite pool into 1 "lumped reaction". This model reduction based on pseudo-steady state lumping leads to much less parameters and much less parameter identifiability problems (see [23–25]).

7.2.3 Thermodynamic Inspired Kinetics

Replacing kinetics of individual reactions by equilibrium constants (pseudo-equilibrium) and by lumped kinetics (pseudo-steady state) significantly reduces the number of required kinetic functions and the number of parameters which need to be indentified, with no loss of model performance!! (see [24, 25]).

Nevertheless, there remain a significant number of far from equilibrium reactions into and from metabolite pools with slower turnover times for which we need

to specify a kinetic function. Unfortunately, mechanism based enzyme kinetics provide highly non-linear rate functions which contain many parameters. Examples are bi–bi kinetics, Hill functions, etc.

Although the number of parameters has been significantly reduced (near-equilibrium, pseudo-steady state based reduction) the identification of the remaining parameters still poses enormous problems, due to the non-linear parameter characteristics. Non-linear parameter estimation algorithms need a decent initial set of parameter values, which is not available, and in addition they do not guarantee an optimal result; they often end in a local minimum of the error criterion. Finally, these identification problems scale very unfavourably with increasing network size (e.g. genome scale metabolic models contain in the order of 1,000 reactions!!).

A possible solution is the use of proper approximative kinetic functions [27]. Here lin-log kinetics has been developed recently [28], which has its roots in the concept that the rate of a process is related to the thermodynamic driving force!! [29].

8 Thermodynamics Inspired Kinetics: Lin-Log Kinetics

8.1 Introduction

Lin-log kinetics is an approximative kinetic format which is a generalisation of the driving force concept and has been compared recently to other approaches such as linear, powerlaw, loglin (for review see [27]) and it was concluded that the lin-log format has significant advantages. Therefore we will focus here on use of lin-log kinetics.

Consider (Fig. 2) an enzyme e_1 which is kinetically affected by its substrate x_1 and product x_2, and is also allosterically affected by a metabolite x_8. Moreover, we consider a reference steady state (superscript ^{0}).

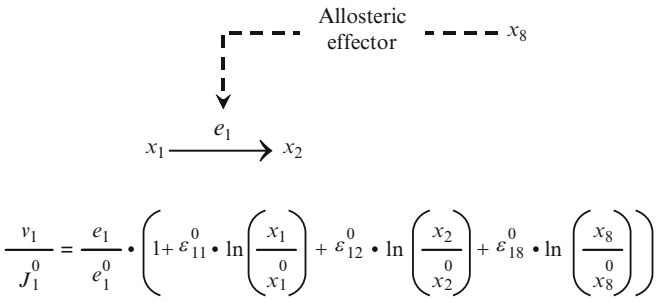


Fig. 2 Lin-log kinetics

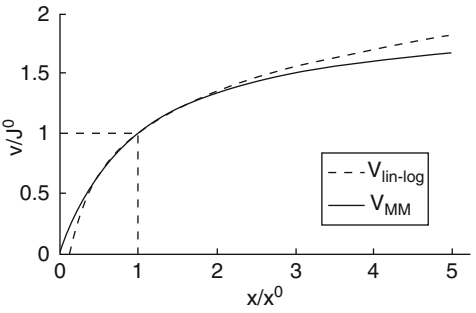


Fig. 3 Comparison between lin-log approximative and hyperbolic mechanistic kinetics

We can then formulate lin-log kinetics:

$$\frac{v_1}{J_1^o} = \frac{e_1}{e_1^o}\left(1 + \varepsilon_{11}^o \ln\frac{x_1}{x_i} + \varepsilon_{12}^o \ln\frac{x_2}{x_2^o} + \varepsilon_{18}^o \ln\frac{x_8}{x_8^o}\right). \tag{9}$$

We can compare approximative lin-log kinetics with traditional mechanistic hyperbolic kinetics and we see (Fig. 3) a very good approximation with respect to large changes (factor 5 up and down) in metabolite concentration.

Other approximative formats, such as power law as S- or GMA-systems and loglin, show similar quality of approximation for such changes in metabolite levels [27].

However metabolic reaction systems not only show *large changes in metabolite concentrations* but especially one faces (in the light of metabolic engineering ambitions) *large changes in enzyme levels*. For example, one can easily achieve enzyme concentration changes of factor 10 up or down due to genetic interventions in gene regulation (promoter libraries) or in gene dosis. Here, lin-log kinetics shows distinct advantages compared to the other approximative kinetic formats, as explained before [27]. A final favourable property of lin-log kinetics is that its parameters (elasticities) are linear in (1) which is significant in the light of parameter identification analysis and parameter estimation algorithms [23–25].

Lin-log kinetics has, after its conception, been successfully applied to kinetic modelling of metabolic reaction networks using in silico studies but has also been applied to experimental systems as will be discussed below.

General important aspects of lin-log kinetics are:

- The format is non-linear in metabolite concentrations.
- The format has a minimum number of kinetic parameters which helps minimise the identification problem.
- The parameters (elasticities) are linear in (1) which has significant advantages with respect to the parameter identifiability and parameter estimation [23–25].
- The elasticity parameters are bounded, e.g. Michaels–Menten kinetics $|\varepsilon| < 1$, for Hill kinetics $|\varepsilon| < n$ (with n subunits in the protein and signs (+ or −) are known. This constrains the ε-estimation problem.

- Lin-log kinetics cannot be applied for datasets where concentrations become zero [28], [30]. An analogous problem also occurs in powerlaw format where an inhibitor concentration cannot become zero.

8.2 *In Silico Studies with Lin-Log Kinetics*

Although Fig. 3 suggests good performance for lin-log kinetics for an individual reaction, a basic question remains how good does the lin-log approximation work in networks. *This has been studied in several simulation studies*. In a first study, where the lin-log concept was introduced [28], a small branched reaction network, including cofactors, was used as a test case. The strongly non-linear kinetic model was approximated with a lin-log model using the theoretical elasticity parameters. A rapid perturbation experiment, where metabolite concentrations did change several fold, was successfully reproduced. Moreover, use of lin-log kinetics leads to the so-called "design equation", which allows one to specify new fluxes/metabolite levels and to calculate analytically the required large changes in enzyme levels. In a follow up study [31] a non-linear model of glycolysis in *E. coli* was successfully approximated using lin-log kinetics and a connected product pathway was successfully redesigned in silico with respect to the required large changes in enzyme activities. Even more interesting was a recent study of Smallbone et al. [32] in which they showed that a lin-log kinetic model of glycolysis in *Saccharomyces cerevisiae*, in which elasticities were assumed equal to their reaction stoichiometric coefficients, gave surprising agreement with the mechanistic Teusink model!! More recently [30] it was shown that lin-log kinetics could also successfully simulate a genetic network with strongly non-linear kinetics. These simulation results show that lin-log kinetics provide a convenient and satisfying approximation of mechanistic kinetic functions for small and large models (metabolic, genetic), provided that metabolite concentrations do not become zero.

Having ascertained that lin-log kinetics provides a decent approximation to non-linear kinetics of networks, the next important problem is to identify the lin-log parameters (elasticities) from experimental data. The obvious experimental protocol is to perform rapid pulse experiments. The identification of lin-log parameters (elasticities) from such experiments was studied first in silico. Kresnowati et al. [33] showed, for a simple linear pathway, that elasticities can easily be obtained by using the integrated equation (1). The important aspect is that the resulting equations are linear in elasticities, so that linear regression can be used to obtain the elasticities. Subsequently a more complex glycolysis model was used for an in silico study of parameter identification aspects to obtain elasticities from dynamic pulse experiments [23–25]. It was found that, in such rapid pulse experiments, fundamental identification problems occur due to occurrence of pseudo-steady state metabolite pools (with turnover time <1 s) in the rapid pulse experiment. The identification problem could only be resolved by providing additional information (e.g. combined

steady state and dynamic perturbations. Also, a novel non-linear algorithm for elasticity parameter estimation was introduced where, due to properties of the lin-log format, one can obtain a reliable initial estimate of their values using linear regression of the integrated equation (1).

Finally, a completely different approach to the parameter identifiability problem in pulse experiments was shown to be possible due to the unique properties of lin-log kinetics [34]. It was shown that use of lin-log kinetics allows *a priori model* reduction by lumping pseudo-steady state pool coupled reactions. The reduced model reproduced the dynamic pulse experiment (*S. cer*, anaerobic glycolysis) accurately and allowed calculation of, e.g. flux control coefficients [34].

An interesting application of a lin-log kinetic model is to identify the function of so-called silent genes [35, 36] which shows how lin-log kinetic models could be used to resolve gene-annotation problems.

8.3 Application of Lin-Log Kinetics to Experimental Data

Lin-log kinetics allows a general steady state analytical solution of networks which gives fluxes as a non-linear function of enzyme activities with flux control coefficients C^J as parameters [28]. This equation was successfully used to obtain *C^J-values* from experimental data:

- A linear product pathway in Penicillin synthesis using fed batch data on flux and enzymes in the penicillin pathway [37]
- The lower glycolysis in a reconstituted linear three enzyme pathway [38]
- A branch point for lysine or glutamate synthesis [39]

Estimation of *elasticities* from experimental data using lin-log format requires experimental information of fluxes, enzyme levels and metabolite levels. Using an extensive steady state dataset for lower glycolysis [38] showed that elasticity values are easily obtained using linear regression. Also lin-log kinetics was used to set-up a kinetic model for leucine/valine synthesis [40], glycolysis in *Lactococcus lactis* [30] and a batch fermentation [41]. In the last two studies it was shown that lin-log kinetics should not be applied to datasets where metabolite concentrations become zero (which is obvious). Even more impressive was a recent study [42] in which a lin-log model was parameterised on a rapid pulse experiment. The model represents *E. coli* central metabolism and anabolism to all cell compounds, comprising 126 reactions and 130 metabolites (7 conserved moieties). The presence of allosteric mechanisms was taken from the MetaCyc database. In total, 921 elasticities were estimated using evolutionary algorithms and high performance computing. This work is a first, genome like scale, whole cell dynamic model and shows the power of the canonical lin-log format approach.

Parameter identification from experiments, however, remains an important issue. Here recently Nikerel et al. [34] showed the successful application of their

a priori model reduction approach (made possible due to lin-log format) on anaerobic glucose pulse experiments in *Saccharomyces cerevisiae*.

This short overview shows that the recently introduced lin-log kinetics, inspired by thermodynamic principles, shows considerable promise in achieving genome scale kinetic modelling of metabolic networks.

9 Conclusion

Thermodynamic principles do shape the properties of biological systems, with considerable and highly interesting consequences for their mathematical models needed in systems biology. Especially noteworthy are the far reaching consequences of the osmotic limit, such as pseudo-steady state, black box kinetics, just in time, control mechanisms, model reduction etc. Also of future importance is the principle of energy optimality for modelling of genetic mechanisms and of thermodynamic driving force (lin-log kinetics) for kinetic modelling. Lin-log kinetics seems to hold considerable promise to obtain realistic genome scale kinetic models (especially due to the lin-log based possibilities towards model reduction, a priori identifiability analysis and an initial estimate of the elasticity parameters).

Abbreviations and Symbols

X	Intracellular metabolite level	μ mol/g DM
S	Stoichiometry matrix	
e	Enzyme amount per cell mass	
p	Parameter	
μ	Specific growth rate	h^{-1}
q_i	Specific uptake/secretion rate	mol (or kJ/h)/CmolX
P_{osm}	Osmotic pressure	N/m^2
v	Rate of intracellular reaction	mol i/h/CmolX
ΔG^0_{fi}	Standard Gibbs energy of formation	kJ/mol
Y^{max}_{GX}	Maximal yield of biomass on Gibbs energy	CmolX/kJ
Y^{max}_{SX}	Maximal yield of biomass on substrate	CmolX/molS
Y^{max}_{SP}	Maximal yield of product on substrate	molP/molS
m_G	Maintenance Gibbs energy requirement	kJ/h/CmolX
C	Number of C-atoms in carbon source	
γ	Degree of reduction of C-source (per C-atom)	
R	Gas constant 8.314	J/molK
T	Absolute temperature	K
$\Delta G^{01}_{cat,D}$	Gibbs energy of the catabolic reaction per mol donor	kJ/mold
K_s	Affinity for substrate	$molS/m^3$
C_s	Substrate concentration	$molS/m^3$
t.o.t.	Turn over time	s
J	Flux	mol/h/C-molX
ε_{ij}	Elasticity coefficient $\left(\frac{x_j}{v_i}\frac{\partial v_i}{\partial x_j}\right)$ of metabolite j on enzyme i	
m_i	Maintenance coefficient for i (O_2, substrate, ethanol, etc.)	$\frac{\text{mol of } i/h}{\text{CmolX}}$

Subscripts

j	Metabolite j
i	Reaction i
p	Product
S	Substrate
x	Biomass
G	Gibbs energy
o	O_2
el	Electrons
D	Donor
Cat	Catabolic

Superscripts

^{1}	Biochemical standard (pH = 7)
^{0}	Standard condition (1 mol/L, 1 bar, 298 K) or reference condition, also reference steady state

References

1. Atkinson DE (1969) Limitation of metabolite concentrations and the conservation of solvent capacity in the living cell. Curr Top Cell Reg 1:29–43
2. Canelas A, van Gulik WM, Heijnen JJ (2008) Determination of the cytosolic free NAD/NADH ratio in *Saccharomyces cerevisiae* under steady-state and highly dynamic conditions. Biotechnol Bioeng 100(4):734–743
3. Nasution U, van Gulik WM, Kleijn RJ, van Winden WA, Proll A, Heijnen JJ (2006) Measurement of intracellular metabolites of primary metabolism and adenine nucleotides in chemostat cultivated *Penicillium chrysogenum*. Biotechnol Bioeng 94(1):159–166
4. Nasution U, van Gulik WM, Proell A, van Winden WA, Heijnen JJ (2006) Generating short-term kinetic responses of primary metabolism of *Penicillium chrysogenum* through glucose perturbation in the bioscope mini reactor. Metab Eng 5(5):395–405
5. Taymaz-Nikerel H, de Mey M, Ras C, ten Pierick A, Seifar RM, van Dam JC, Heijnen JJ, van Gulik WM (2009) Development and application of a differential method for reliable metabolome analysis in *Escherichia coli*. Anal Biochem 386(11):9–19
6. Maris AJA, Konings WN, Dijken JP, Pronk JT (2004) Microbial export of lactic and 3-hydroxypropanoic acid: implications for industrial fermentation processes. Metab Eng 6 (4):245–255
7. Van Gulik WM, Heijnen JJ (1995) A metabolic network stoichiometry analysis of microbial growth and product formation. Biotechnol Bioeng 48(6):681–98
8. Van Gulik WM, De Laat WTAM, Vinke JL, Heijnen JJ (2000) Application of metabolic flux analysis for the identification of metabolic bottlenecks in the biosynthesis of penicillin-G. Biotechnol Bioeng 68(6):602–618
9. Heijnen JJ, Van Dijken JP (1992) In search of a thermodynamic description of biomass yields for the chemotrophic growth of microorganisms. Biotechnol Bioeng 39(8):833–58
10. Heijnen JJ (1999) Bioenergetics of microbial growth. In: Flickinger MC, Drew SW (eds) Encyclopedia of bioprocess technology: fermentation, biocatalysis and bioseparation. ISBN: 0-471-13822-3. Wiley, New York, pp 267–291

11. Tijhuis L, van Loosdrecht MCM, Heijnen JJ (1993) A thermodynamically based correlation for maintenance Gibbs energy requirements in aerobic and anaerobic chemotrophic growth. Biotechnol Bioeng 42(4):509–19
12. Anderson KB, Meyenburg K von (1980) J Bacteriol 144:114–123
13. De Jong-Gubbels P, Vanrolleghem P, Heijnen S, Van Dijken JP, Pronk JT (1995) Regulation of carbon metabolism in chemostat cultures of Saccharomyces cerevisiae grown on mixtures of glucose and ethanol. Yeast 11(5):407–418
14. Vanrolleghem PA, de Jong-Gubbels P, van Gulik WM, Pronk JT, van Dijken JP, Heijnen S (1996) Validation of a metabolic network for *Saccharomyces cerevisiae* using mixed substrate studies. Biotechnol Progr 12(4):434–448
15. Stuckrath I, Lange HC, Kotter P, Van Gulik WM, Entian KD, Heijnen JJ (2002) Characterization of null mutants of the glyoxylate cycle and gluconeogenic enzymes in *S. cerevisiae* through metabolic network modeling verified by chemostat cultivation. Biotechnol Bioeng 77(1):61–72
16. Van Gulik WM, Antoniewicz MR, DeLaat WTAM, Vinke JL, Heijnen JJ (2001) Energetics of growth and penicillin production in a high-producing strain of *Penicillium chrysogenum*. Biotechnol Bioeng 72(2):185–193
17. Smolders GJF, van der Meij J, van Loosdrecht MCM, Heijnen JJ (1995) A structured metabolic model for anaerobic and aerobic stoichiometry and kinetics of the biological phosphorus removal process. Biotechnol Bioeng 47(3):277–87
18. Wiechert W (2002) Modeling and simulation: tools for metabolic engineering. J Biotechnol 94:37–63
19. Teusink B, Passarge J, Reijenga CA, Esgalhado E, Van der Weijden CC, Schepper M et al. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. Eur J Biochem 267:5313–5329
20. Rizzi M et al. (1997) In vivo analysis of metabolic dynamics in *Saccharomyces cerevisiae*: II. Mathematical model. Biotechnol Bioeng 55:592–608
21. Theobald U et al. (1997) In vivo analysis of metabolic dynamics in *Saccharomyces cerevisiae*: I. Experimental observations. Biotechnol Bioeng 55:305–316
22. Visser D et al. (2002) Rapid sampling for analysis of in vivo kinetics using the BioScope: a system for continuous-pulse experiments. Biotechnol Bioeng 79:674–681
23. Nikerel IE, Van Winden WA, Van Gulik WM, Heijnen JJ (2006) A method for estimation of in-vivo elasticities in metabolic networks using data from steady-state and rapid sampling experiments with lin-log kinetics. BMC Bioinformatics 7:540
24. Nikerel IE, Van Winden WA, Van Gulik WM, Heijnen JJ (2007) Linear-logarithmic kinetics; a framework for modelling kinetics of metabolic reaction networks. Simulation News Europe 17(1):19–26
25. Nikerel IE, Van Winden WA, Verheijen PJT, Heijnen JJ (2009) Model reduction and a priori kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with lin-log kinetics. Metabolic engineering 11:20–30
26. Wu L, Van Dam JC, Schipper D, Kresnowati MTAP, Pröll A, Ras C, Van Winden WA, Van Gulik WM, Heijnen JJ (2006) Short term metabolome dynamics and carbon, electron and ATP balances in chemostat-grown *Saccharomyces cerevisiae* CEN-PK.113-7D following a glucose pulse. Appl Environ Microbiol 72(5):3566–3577
27. Heijnen JJ (2005) Approximative kinetic formats used in metabolic network modelling – review. Biotechnol Bioeng 91(5):534–545
28. Visser D, Heijnen JJ (2003) Dynamic simulation and metabolic re-design of a branched pathway using lin-log kinetics. Metab Eng 5:164–176
29. Westerhoff HV, ven Dam K (1987) Mosaic non-equilibrium thermodynamics and the control of biological free energy transduction. Elsevier, Amsterdam
30. Del Rosario RCH, Mendoza E, Voit EO (2008) Challenges in lin-log modelling of glycolysis in *Lactococcus lactis*. IET Syst Biol 2(3):136–149
31. Visser D, Schmid JW, Mauch K, Reuss M, Heijnen JJ (2004) Optimal redesign of primary metabolism in *Escheria coli* using lin-log kinetics. Metab Eng 6:378–390

32. Smallbone K, Simeonides E, Broomhead D, Kell D (2007) Something from nothing – bridging the gap between constraint – based and kinetic modelling. FEBS J 274:5576–85
33. Kresnowati MTAP, van Winden WA, Heijnen JJ (2005) Determination of elasticities, concentration and flux control coefficients from transient metabolite data using linlog kinetics. Metab Eng 7(2):142–153
34. Nikerel IE, Canelas AB, Jol SJ, Verheijen PJT, Heijnen JJ (2009) Construction of kinetic models for metabolic reaction networks: lessons learned in analysing short-term stimulus response data. Proceedings MATHMOD 09, 805–814. Presented at MATHMOD 09, Vienna, February 11–13, 2009
35. Raamsdonk LM, Teusink B, Broadhurst D, Zhang NS, Hayes A, Walsh MC et al (2001) A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. Nat Biotechnol 19:45–50
36. Wu L, Van Winden WA, Van Gulik WM, Heijnen JJ (2005) Application of metabolome data in functional genomics: a conceptual strategy. Metab Eng 7:302–310
37. Van Gulik WM, van Winden WA, Heijnen JJ (2002) Metabolic flux analysis, modeling and engineering solutions. In Vinci VA, Parekj SR (eds) Handbook of industrial cell culture. Mammalian, microbial, and plant cells. Humana Press, Towota, pp 351–395
38. Wu L, Wang W, Van Winden WA, Van Gulik WM, Heijnen JJ (2004) A new framework for the estimation of control parameters in metabolic pathways using lin-log kinetics. Eur J Biochem 271:3348–3359
39. Heijnen JJ, van Gulik WM, Shimizu H, Stephanopoulus G (2004) Metabolic flux control analysis of branch points: an improved approach to obtain flux control coefficients from large perturbation data. Metab Eng 6:391–400
40. Magnus JB, Hollwedel D, Oldiger M, Takors R (2006) Monitoring and modelling of the reaction dynamics in the valise/leucine synthesis pathway in *Corynebacterium glutamicum*. Biotechnol Progr 22:1071–1083
41. Feng-Sheng Wang, Chih-Lung Ko, Voit EO (2007) Kinetic modelling using S-systems and lin-log approaches. Biochem Eng J 33:238–247
42. Reuss M, Aquilera-Vazquez L, Mauch K (2007) Reconstruction of dynamic network models from metabolite measurements. Top Curr Genet 18:97–127. Nielsen J, Jewett MC (eds) Metabolomics. Springer, Heidelberg
43. Reed JL, Palsson BO (2003) Thirteen years of building constraint-based in-silico models of *Escherichia coli*. J Bacteriol 185(9):2692–2699
44. Heinrich R, Schuster S (1996) The regulation of cellular systems. Science 372 p

Index