Christoph Wittmann · Sang Yup Lee
Editors

Systems Metabolic Engineering

Springer

Systems Metabolic Engineering

Christoph Wittmann • Sang Yup Lee
Editors

Systems Metabolic Engineering

Springer

Editors

Christoph Wittmann
Institute of Biochemical Engineering
Braunschweig Integrated Center for Systems Biology (BRICS)
Center for Pharmaceutical Engineering
Technische Universität Braunschweig
Braunschweig, Germany

Sang Yup Lee
Metabolic and Biomolecular Engineering National Research Laboratory
Department of Chemical and Biomolecular Engineering (BK21 Program)
Center for Systems and Synthetic Biotechnology
Institute for the BioCentury KAIST
Daejeon, Republic of Korea

ISBN 978-94-007-4533-9 ISBN 978-94-007-4534-6 (eBook)
DOI 10.1007/978-94-007-4534-6
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2012941228

Printed on acid-free paper

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

To Heike, Isabelle, Felix and Florian from Christoph and to Hyejean and Gina from Sang Yup for their love, support and inspiration

Preface

The integration of systems-wide omics approaches, genome-scale modeling and simulation, synthetic biological approaches, and even evolutionary engineering is opening a new era of industrial strain engineering – the design-based creation of tailor-made overproducers that are optimized at the global level. This integrated approach of metabolic engineering is now called systems metabolic engineering. At the entry into a new millennium, facing strong needs for a novel bio-economy due to global warming and shortage of fossil fuels, this seems one of the most relevant and promising areas of research and industrial application. The present book is devoted to this fascinating area of systems-wide analysis and engineering of cellular metabolism. Through a series of exciting chapters, world leading experts provide us with up-front approaches to analyze, model and re-design biological systems towards desired properties and enrich this by real-case applications for the most relevant workhorses in industrial bio-production.

The book starts with computational and experimental methods on the systems-wide analysis of biological systems, the entry and basis to create understanding and enable knowledge-based systems metabolic engineering. In Chap. 1, Professors Lee and Palsson together with their teams combine their pioneering expertise on computational modeling of genome-scale networks, the tin-opener in many of the successful projects on systems metabolic engineering reported today. They provide the full picture, touching metabolic networks, transcriptional networks and cell-signaling networks and include interesting case studies on the biological systems picked up in the later application chapters of the book. Chapter 2 by Professor Palsson and colleagues extends the stoichiometric modeling of Chap. 1 to kinetic models of metabolism, crucial to describe systems dynamics. Their contribution gives valuable hands-on advice for the creation of kinetic models, provides the fundamental mathematics and closes with practical application examples. Chapter 3 by Professor Shimizu's group complements the dry lab analysis of networks via wet lab approaches, covering state-of-art omics technologies. They describe how thoroughly designed experimental studies deliver deep understanding of industrial microorganisms and how metabolic engineers can efficiently exploit this towards optimized strains.

The above approaches on systems-wide computational and experimental analysis of biological systems, which form the initial part of the book, provide design strategies for superior cell factories that have to be implemented on the DNA level. The more we want to shape and create, the larger the genetic changes necessary. In this regard, Chap. 4 by Professor Panke and co-workers discusses how to efficiently translate design concepts into synthetic DNA sequences. The authors especially focus on novel large-scale synthetic engineering approaches and provide us with an interesting view on completely design-based synthetic systems.

Chapters 5, 6, 7, 8, 9 and 10 pickup most relevant industrial workhorses from the groups of bacteria, yeasts and fungi and illustrate how systems metabolic engineering is used today for next-level strain and bioprocess development. Chapter 5 by the group of Professor Lee focuses on *Escherichia coli*, probably the most deeply studied microorganism on the systems level. Their set of examples on a wide set of products underline how advanced we are in creating tailor-made *E. coli* cell factories and what is needed to go even further. In Chap. 6, Professor Wittmann and his team review systems metabolic engineering of *Corynebacterium glutamicum*. They explain how this gram-positive soil bacterium can be tailored to convert a broad spectrum of renewable raw materials into various chemicals, fuels, materials or therapeutics and thus, similarly, to *E*. coli, is becoming a successful bio-production platform. Chapter 7 by Professor Papoutsakis and colleagues deals with *Clostridium acetobutylicum*, a famous bacterium for production of solvents since the very beginning of industrial biotechnology almost a 100 years ago. Their contribution focuses on improved tolerance from a systems view point, a key target of superior strains, and especially valid towards high-level production of the often unnatural chemicals toxic for the cell. Chapters 8 and 9 highlight eukaryotic production systems. In Chap. 8, Professor Heijnen's group describe systems-level design of *Penicillium chrysogenum*, well-known for its high relevance for antibiotics production and a model system for the rich set of industrial processes with other filamentous fungi. Their contribution nicely recruits modeling to elucidate function and control of metabolism for strain design. Chapter 9 by Professor Nielsen and co-workers deals with yeast and illustrates how omics technologies can be integrated with synthetic biology for rational DNA modification into a knowledge-based framework for systems metabolic engineering. They complement this by two case studies from biofuel production, which are among the most relevant bioprocesses in yeast industrial biotechnology. In Chap. 10, Professor Kondo and his team provide us with interesting examples on systems metabolic engineering of cellular properties that are crucial to successfully integrate cell factories into the rising concept of biorefinery. In this regard their review discusses improved utilization of renewable raw materials – direct conversion of the mainly mixed, polymeric substrates as well as improved tolerance to toxic ingredients.

Chapter 11 closes the book by opening a new door. Professor Stephanopoulos and colleagues illustrate how we can exploit ideas and tools of systems metabolic engineering and systems biology to address key questions in medicine. Their contribution on cancer as a metabolic disease might stimulate to further extend the application of the engineering concepts described throughout the book towards a new field of research.

As compiled in this book, we are now reaching the level of global analysis, design and engineering of biological systems. This provides a cornucopia of novel possibilities – sustainable supply of chemicals, materials and fuels in a new era of bio-production as well as tailor-made therapies of threatening diseases. We hope that the book is interesting and valuable for researchers and engineers from the various disciplines that are all integrated into the field. Thanks to the worldwide experts and their excellent contributions, which are greatly appreciated, this book hopefully sets a milestone with perpetual value. We would like to deeply thank the members of our labs, led by Dr. Judith Becker, for their great efforts in editing and formatting the book. Finally, we would like to thank the people at Springer for their assistance in the production. Admittedly, we are still away from the immaculate cell factory, but the way towards it has become visible – and it is a privilege to walk on and share it with you.

TU Braunschweig, Germany — Christoph Wittmann
KAIST, Daejeon, Republic of Korea — Sang Yup Lee

Contents

Chapter 1
Genome-Scale Network Modeling

Sang Yup Lee, Seung Bum Sohn, Hyun Uk Kim, Jong Myoung Park, Tae Yong Kim, Jeffrey D. Orth, and Bernhard Ø. Palsson

Contents

S.Y. Lee (✉)
Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 Program), Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, KAIST, Daejeon, Republic of Korea
e-mail: leesy@kaist.ac.kr

S.B. Sohn • J.M. Park
Metabolic and Biomolecular Engineering National Research Laboratory, Department of Chemical and Biomolecular Engineering (BK21 Program), Center for Systems and Synthetic Biotechnology, Institute for the BioCentury, KAIST, Daejeon, Republic of Korea

Bioinformatics Research Center, KAIST, Daejeon, Republic of Korea

H.U. Kim •
T.Y. Kim
Bioinformatics Research Center, KAIST, Daejeon, Republic of Korea

J.D. Orth • B.Ø. Palsson
Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA

C. Wittmann and S.Y. Lee (eds.), *Systems Metabolic Engineering*,
DOI 10.1007/978-94-007-4534-6_1, © Springer Science+Business Media Dordrecht 2012

Abstract Genome-scale models have garnered considerable interest for their ability to elucidate cellular characteristics and lead to a better understanding of biological systems. Metabolic models in particular have been widely used to study complex metabolic pathways in order to better understand microbial systems and to design strategies for engineering various biotechnological applications. Similar to metabolic networks, transcriptional and signaling network models have also been reconstructed to elucidate regulatory interactions and to further understand the response of systems to various environmental stimuli. However, a true genome-scale model that integrates all these characteristics into one comprehensive model has not yet been constructed. For the time being, the existing network models have individually contributed to the knowledge of their respective fields and to our understanding of biological systems. In selected cases they have provided design strategies for systems wide engineering of metabolism. There have been several attempts to integrate these networks to realize the full potential of a complete cellular network model, although there is still room for further development. Here, we review the different network types and highlight their contributions to biotechnological applications via illustrative examples.

Keywords Genome-scale model • Metabolic network • Transcriptional network • Signaling network • *Escherichia coli* • Network reconstruction • Genome annotation • *In silico* • Stoichiometric matrix • Steady-state • Automatic reconstruction • Single-input module • Top-down • Bottom-up • Boolean formalism • Flux analysis • Stoichiometric formalism • Kinetic formalism

1.1 Introduction

Understanding and visualizing of biological networks has become an important aspect of systems biology as more information and knowledge are being generated. The availability of a network describing a particular aspect of the biological system, whether it is metabolism or transcriptional regulation, allows the user to better understand how the system can respond to the ever-changing external environment. With the advent of the full genome sequence, the reconstruction of a full genome-scale model of a cellular system has become feasible. Currently, there are genome-scale metabolic models [1, 2] and recently genome-scale transcriptional networks have begun to appear [3, 4]. However, a true genome-scale model which integrates the metabolism, the transcriptional regulatory network, and all other networks that are found in biological systems into one comprehensive model is still being developed.

Current models of biological systems have provided researchers with a wealth of knowledge regarding their respective scopes. Metabolic network models have aided in the design of new strategies for systems metabolic engineering of host strains for the production of high-value compounds in cell factories of *Escherichia coli* [5] or *Corynebacterium glutamicum* [6], as outlined later for the respective microorganisms

Table 1.1 Overview of the different networks

Type	Metabolic networks	Transcriptional networks	Signaling networks
Definition	Network of biochemical reactions which reflect the metabolic state of the cell	Network reflecting the expression state of the genome	Network of proteins that transduce information that changes the transcriptional state of the cell
Components	Metabolites Metabolic reactions	Promoters Transcription factors Metabolites	Proteins Protein-protein interactions
Source of data	Genome annotation ^{13}C flux data Enzyme analysis Biochemical databases Other curated databases Literature	Gene expression data Location analysis Predictive algorithms for promoters Curated databases Literature	Signaling databases Protein-protein interactome Gene expression data RNAi knockdown of signaling pathways Proteome Fluorescent localization data

throughout this book. Moreover, transcriptional network models have aided in the identification of a number of new transcription factors or binding sites [7]. Despite the incompleteness of these models, they continue to provide valuable knowledge in filling in gaps in our understanding of these biological networks [8].

Here, we discuss the characteristics of three biological networks that have been extensively studied in recent years (Table 1.1). Metabolic networks have been the most studied of the three, and have been utilized in a large number of applications from drug discovery to industrial production of high-valued biochemicals opening a new era of design-based metabolic engineering as illustrated throughout this book. Transcriptional networks and signaling networks are still in their infancy with regard to large-scale network reconstructions. However, there have been major advances in each field, which will be discussed in their respective sections.

1.2 Metabolic Networks

A metabolic network model describes the metabolic state of the cell. It is composed of biochemical reactions that are constrained by the laws of thermodynamics and mass action. The metabolic network can be modeled on different levels of complexity based on what is being examined. The uppermost level is the cellular level, where only the cellular inputs and outputs are of concern and the mechanics within the cell are not. Below that is the functional level of the metabolic network where the network is divided based on the functions performed by a particular section of the network, e.g., catabolic or anabolic. The next level examines the pathways in functional groups, such as glycolysis or amino acid biosynthesis. Finally, at the foundation of the metabolic network are the individual biochemical reactions.

In the rest of this chapter, we will concern ourselves with this level of the metabolic network as most current genome-scale metabolic networks are reconstructed as lists of biochemical reactions.

To rebuild the metabolic network of a particular organism, a draft reconstruction is first assembled based on the organism's genome annotation. All known metabolic reactions in the organism of interest must be collected and incorporated into the network reconstruction. The reactions associated with each metabolic gene can come from sources such as annotated gene names, EC numbers [9], and GO terms [10]. Multispecies metabolic databases such as KEGG [11] can also be used to match genes with their reactions. The assembly of a draft metabolic reconstruction can be performed manually, or it can be automated [12, 13]. The draft metabolic reconstruction will certainly contain errors and missing information, particularly if the draft reconstruction was automated, and so it must be manually curated.

Confidence levels for the presence of each biochemical reaction in the network should be determined. These confidence levels are based on evidence of the existence of the reaction in the organism according to literature and experimental studies. For instance, biochemical data indicating a reaction's presence in the organism, such as an assay of a purified enzyme, would have the highest confidence level and would be included in the metabolic reconstruction. Unfortunately, specific biochemical data for every biochemical reaction for every species does not exist. Therefore other data sources are utilized. With the availability of full genome sequences of many other organisms, the function of genes can be determined through the use of genetic data from previously characterized organisms indicating the corresponding metabolic reactions. Data on gene knockouts and their effect on metabolism and homology with genes with known functions from other species are some examples of genetic data, and account for most of the information in metabolic networks of less well-characterized species.

Physiological data, such as secreted metabolite concentrations or glucose and oxygen uptake rates, are also useful. When *in vivo* physiology does not match the *in silico* results, it is usually due to an incomplete annotation or insufficient characterization of the organism. For example, if a species is known to produce a certain metabolite and yet the genome does not include known genes encoding the required metabolic enzymes, those reactions are included in the metabolic network with a lower confidence level to ensure consistency with known physiology. Finally there are *in silico* simulation data, which gives the lowest confidence level for those biochemical reactions included in the metabolic network. Reactions of this sort are usually included to ensure that the target objective is generated, such as biomass synthesis reactions.

A completely curated metabolic network reconstruction must be converted to a mathematical model to make computational predictions. Constraint-based modeling is most often used to represent metabolic networks [14–17]. The reactions are encoded in a stoichiometric matrix in which each row represents a metabolite and each column represents a reaction. The elements of the matrix are the stoichiometric coefficients of each metabolite in the reaction. Upper and lower bound constraints on each reaction can be imposed. This model can then be used in different mathematical analyses, the most common being flux balance analysis (FBA) [15]. Here, the steady-state metabolic flux distribution of the cell is determined using linear programming where an objective function, such as maximum cell growth rate, is selected, giving a particular state of the metabolic network. The assumption of steady-state is made possible by the time scale difference between the cellular level growth and the reaction level fluxes. Once the mathematical model of the metabolic network is constructed, the accuracy of the network can be validated through comparison with experimental data. Phenotypic data for growth on different substrates [18] or with gene knockouts [19] can be directly compared to FBA predicted phenotypes. These comparisons continue in an iterative fashion, where the network is tested and updated based on the results. Other analyses that can be performed on the metabolic model include pathway analysis [20], elementary flux mode analysis [21], gene expression analysis [22], and adaptive evolution analysis [23].

1.2.1 Metabolic Network Case Studies

The metabolic network model of *Haemophilus influenza* was the first reconstructed genome-scale metabolic network, published in 1999 [24]. Since then, more than 70 genome-scale metabolic networks have been reconstructed and published (Fig. 1.1). Although the majority of published metabolic networks are of bacterial systems, metabolic networks for archaea and eukaryotes, including *Homo sapiens* [25, 26], have been reconstructed and studied. These reconstructed metabolic networks have been utilized to analyze and characterize their respective organisms. With these metabolic networks, researchers have been able to investigate physiological characteristics of the organism of interest or suggest engineering strategies for improving target organisms for the overproduction of value-added substances.

The *E. coli* metabolic network has been at the center of metabolic network reconstruction due to the important role it plays in microbiology. It is the best-characterized microorganism with a wealth of literature support, and its veteran status in microbial studies has created well established tools for genetic manipulation. Because of these resources, the *E. coli* metabolic model is perhaps the most easily validated metabolic model available. It has undergone continuous updates since its initial publication in 2000 [1, 27, 28]. Metabolic models of *E. coli* have been utilized in various biotechnological applications, particularly in metabolic engineering, where the *E. coli* metabolic network has been engineered to produce high quantities of value-added substances. Examples include the use of *E. coli* to produce lycopene [29, 30], L-valine [5], and biopolymers such as poly-lactate [31]. In all of these cases, the metabolic network was analyzed through the use of algorithms, such as MOMA (which uses quadratic programming to identify gene knockout targets) [19] or FSEOF (which identifies gene amplification targets) [30], to identify the best target for modification in the metabolic network such that production of the target compound increases.

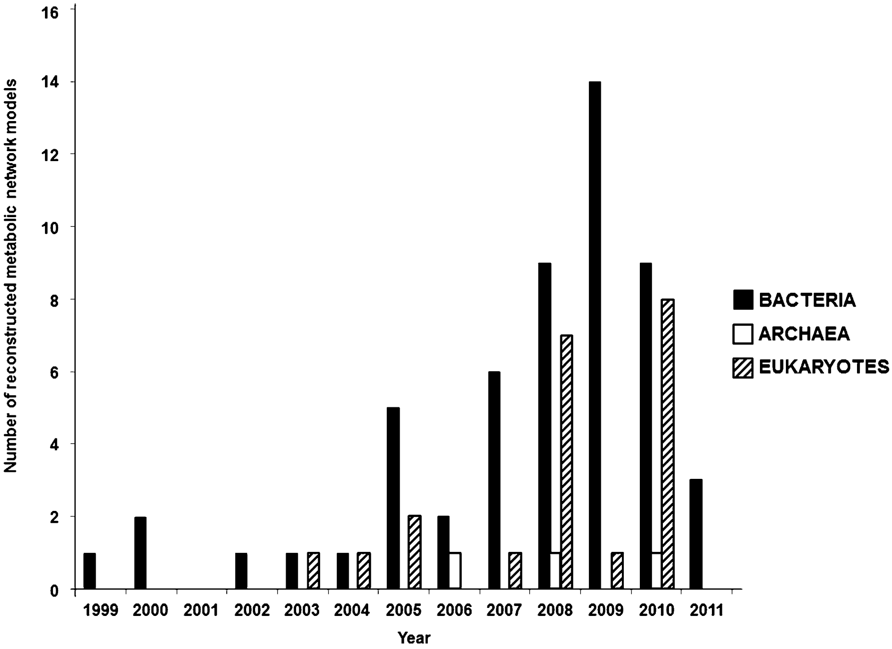


Fig. 1.1 Number of publications of metabolic networks over the last 10 years broken down for the number of publications for bacteria, archaea, and eukarya

Several algorithms now exist for the developing recombinant strains with improved production of high-valued compounds using metabolic network models. OptKnock [32] is an algorithm that uses bi-level linear programming to identify the optimal set of gene knockouts to couple production of a target compound to growth. OptGene [33] is a related method that uses a genetic algorithm to identify gene targets. The algorithm OptStrain [34] predicts new reactions to add to the metabolic network in conjunction with gene knockouts to improve production of the target compound. In addition to gene knockouts, the approach Flux Design [35] identifies targets for amplification as well as down regulation through the use of elementary flux mode analysis. Other strategies have been predicted for *E. coli* [36], and designs for the production of L-lactate have been constructed and optimized by adaptive evolution [37] and through media design for increased L-methionine production [38].

Other metabolic networks have been utilized in a similar capacity for the metabolic engineering of new strains with improved performance. *Mannheimia succiniciproducens* [39] has been studied for the production of succinic acid, *C. glutamicum* [40] for the production of various amino acids, and *Streptomyces coelicolor* A3(2) [41] for the production of antibiotics. Other industrial species with available metabolic network reconstructions include *Pseudomonas putida* [42], *Clostridium acetobutylicum* [43], and *Zymomonas mobilis* [44] to name a few. As many of these species are not as extensively characterized as *E. coli*, their genome-scale metabolic networks have had limited use in metabolic engineering. However, they have been helpful in furthering our understanding of the species' metabolism and have provided a platform for further experimental studies of their metabolism.

Pathogenic organisms such as *Helicobacter pylori* [45], *Vibrio vulnificus* [46], *Pseudomonas aeruginosa* PAO1 [47], and *Acinetobacter baumannii* [48] have had their metabolic networks reconstructed so that drug targets could be identified for the treatment of infections. These targets are usually at points of fragility in the metabolic network of the pathogen [48, 49]. Robustness analysis of metabolism identifies these fragile points, but also identifies possible alternate routes the organism can use to counteract treatments that are suggested [50, 51]. On a related note, the human metabolic network [25] is also utilized when searching for possible drug targets against pathogens. To prevent possible side-effects of a newly introduced drug in the human host, the human metabolic network is analyzed to determine if human metabolism will be affected. The drug must target components of the pathogenic metabolism that are not found in the human metabolic network.

The field of metabolic network reconstruction has mainly focused on prokaryotes due to the relative simplicity of prokaryotic cellular systems. However, there have been a number of eukaryotic organisms for which metabolic networks have been reconstructed. The most extensively studied eukaryotic metabolic network is that of the budding yeast *Saccharomyces cerevisiae* [52]. Like the prokaryote *E. coli, S. cerevisiae* is the best-characterized eukaryotic biological system with a wide array of tools available for in depth studies. The *S. cerevisiae* metabolic network has undergone multiple upgrades and revisions to incorporate new information and data into the metabolic network [52, 53]. However, because of this network's complexity, including compartmentalization of the metabolic network to represent organelles, there have been limited practical applications. Several recent updates to the metabolic network have improved the accuracy of the *S. cerevisiae* metabolic model [54, 55]. Zomorrodi and Maranas improved upon the previously reconstructed metabolic network of *S. cerevisiae*, *i*MM904 [52], and Dobson and coworkers improved upon the consensus reconstruction Yeast 1.0 [56]. The difference between the two reconstructions lies in the level of confidence the authors are willing to attribute to the metabolic reactions included in the network. These slightly different reconstructions of the same organism give different results to an identical problem, allowing researchers to perform comparative analyses, which in turn lead to the identification of previously unknown characteristics of the metabolic network and thereby improve our knowledge of the network. Other eukaryotic systems for which metabolic networks have been reconstructed include the methylotrophic yeast *Pichia pastoris* [2], several fungi from the genus *Aspergillus* [57], the mouse, *Mus musculus* [58], the plant model organism, *Arabidopsis thaliana* [59], and human [25, 26].

The human metabolic network has many complexities that prevent it from being analyzed in the same way as metabolic networks of unicellular organisms. The human biological system is composed of many different tissues with different cell types, each having their own unique metabolic profiles. As a result, use of the metabolic network typically requires additional information regarding the metabolism of the cell type of interest. Studies have been performed using the human metabolic network by mapping gene expression data from specific tissues types, such as brain cells, to the metabolic network [60, 61]. From these studies, the authors were able to elucidate specific aspects of the metabolism of a particular cell type under the given conditions from which the expression data was taken.

Recently, another emerging trend related to metabolic networks is the process of automatic reconstruction [12, 62]. Generally, the process of reconstructing genome-scale metabolic networks is performed manually in more than 90 steps [63]. This process is extremely complex and explains the slow pace of the construction of new metabolic networks. To speed up the process, Henry et al. created the Model SEED, a web-based platform that can, in a high-throughput manner, generate, optimize, and analyze reconstructed metabolic networks [12]. The Model SEED integrates existing methods; draft reconstruction of metabolic networks, gap-filling, analysis of metabolic networks, comparison of metabolic networks with phenotypic data, and manual curation. It introduces techniques to automate the steps of the reconstruction process, taking less than 2 days to reconstruct a metabolic network from an assembled genome sequence. Another automatic platform, MEMOSys (MEtabolic MOdel research and development System) supports the development of a new metabolic network by providing a version control system that can show the complete developmental history. Utilizing MEMOSys, existing models can be researched through the use of search systems, references to external databases, and feature-rich comparison mechanism to verify and refine pre-reconstructed metabolic networks [62]. While the tools for the reconstruction of metabolic networks still have limitations, particularly pertaining to poorly characterized organisms and complex metabolic networks, the automatic reconstruction pipeline is yet another advancement in the field of genome-scale metabolic networks.

1.3 Transcriptional Networks

A transcriptional network reconstruction is a representation of the network that controls the gene expression state of the cell. Not all genes in the genome are expressed at the same time in the cell, and the transcriptional network provides a blueprint for how the cell controls the timing of which genes are expressed under specific conditions. Biochemically, transcription is only partially understood, but it is known that the interactions in the transcriptional network include protein-protein interactions and protein-DNA interactions.

There are two fundamental building blocks needed to reconstruct the transcriptional network: the promoters of the genes and the transcription factors (TFs) that bind to each promoter. Identifying promoter regions of the genome is relatively straightforward, but identifying which TF binds to which promoter is more complex. In the case of higher organisms, the promoters possess sites for multiple TFs, resulting in an increasing number of combinations that can bind to each promoter. This greater complexity, and therefore a greater number of possibilities in transcriptional states, allows for more possibilities in functional states of the cell.

Information on promoters, TFs, and DNA binding proteins is usually taken from different types of experimental data which are classified by their increasing complexity: component data, interaction data, and network state data. Component data contains details about the individual components, such as TFs, in the transcription network system. Interaction data can show which TFs are active and with which promoters the TFs are interacting. Network state data shows the transcriptional state of the entire system at a specific time, which can be utilized to determine connectivity in the transcriptional network.

These components are then assembled into an interactive network representing the transcriptional state of the genome. With the construction of the transcriptional network, it was found that several basic motifs commonly occur, such as the feed-forward loop, in which components activate sections of the network downstream to speed up the response towards the input; the single input module (SIM), where a single input leads to the activation of multiple outputs; and dense overlapping regulons (DOR), which are composed of regions of complex interactions and TFs that are involved in multiple interactions in the network [64]. The classification of the different motifs simplifies the complexities of the transcriptional network and allows the user to better visualize the roles of the various components of the network.

There are two main approaches to reconstruct the transcriptional network: Top-down and Bottom-up (Fig. 1.2). The top-down approach usually utilizes high-throughput data that simultaneously measure large numbers of data points to identify the individual components of the transcriptional network. Some examples of top-down approaches include identification of the expression status of the genome, identification of all promoter sites computationally, and the experimental identification of all protein binding sites on the DNA. As with all high-throughput data, each type requires detailed curation before being used in the reconstruction. From the opposite end, the bottom-up approach involves the individual components, which are studied, characterized, and connected to the network. This method attributes high confidence to the data being used in the reconstruction of the transcriptional network. However, the process of obtaining all the necessary data for each individual component is time intensive. Therefore, a combination of the top-down approach and the bottom-up approach is usually preferred for transcriptional network reconstruction.

With the reconstruction of the transcriptional network, one can examine various properties of the system's transcriptional intricacies. First, the user can utilize the transcriptional network to obtain a better understanding of the transcription patterns through analysis of the network to reveal new information or explain observed effects. Second, causal relationships can be better understood and new relationships between previously unrelated components can be identified. Third, a reaction mechanism can be suggested based on the analysis of the transcriptional model. Finally, kinetic constants can be better estimated with the help of the transcriptional model through tuning and refinement [65].

Various methods have been developed to reconstruct transcriptional regulatory networks based on gene expression profiling data, literature data, and databases. These methods yield directed graph networks [64], Boolean networks [66–68], Boolean networks in a matrix format [69], dynamic modeling of the network [70], and probabilistic modeling of the network using Bayesian network analysis [71].

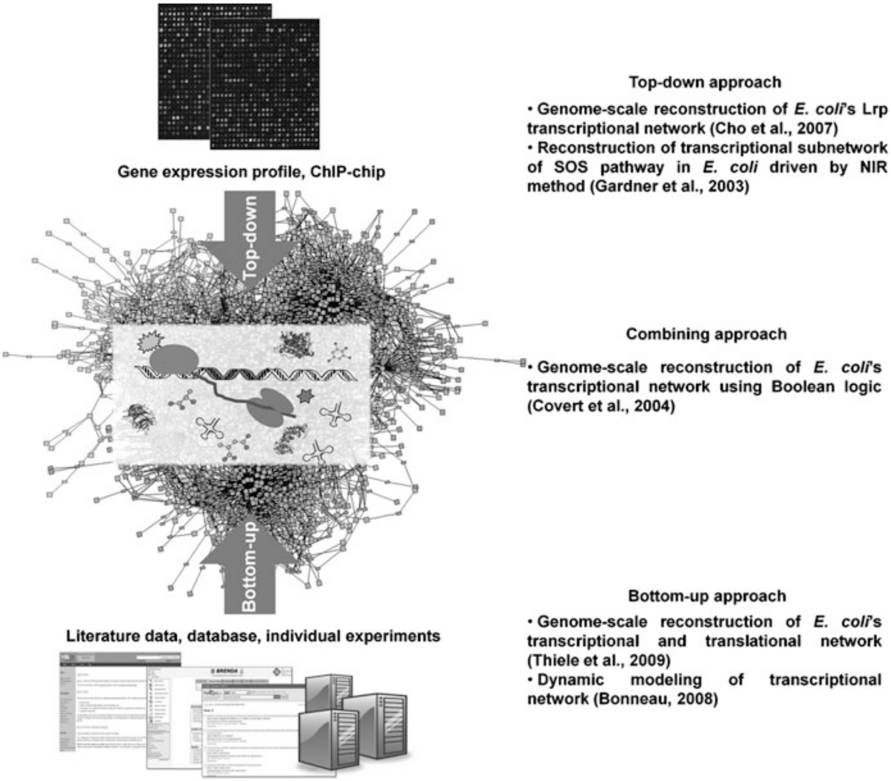

Gene expression profile, ChIP-chip
Top-down
Bottom-up
Literature data, database, individual experiments
Top-down approach
• Genome-scale reconstruction of E. coli's Lrp transcriptional network (Cho et al., 2007)
• Reconstruction of transcriptional subnetwork of SOS pathway in E. coli driven by NIR method (Gardner et al., 2003)
Combining approach
• Genome-scale reconstruction of E. coli's transcriptional network using Boolean logic (Covert et al., 2004)
Bottom-up approach
• Genome-scale reconstruction of E. coli's transcriptional and translational network (Thiele et al., 2009)
• Dynamic modeling of transcriptional network (Bonneau, 2008)

Fig. 1.2 Breakdown of the different approaches towards transcriptional network reconstruction. These two approaches are also utilized together in a combinatorial approach seen in Covert et al. [66]

Furthermore, the reconstructed transcriptional networks can be integrated with other network types, including metabolic [66, 68, 72, 73] and signaling networks [74, 75]. These integrated models allow the accurate prediction on the effects the transcriptional regulatory perturbations has for a given condition on the metabolic network. The various networks can be used to integrate omics data to analyze cellular phenotypes and more accurately predict the phenotypes of a cell for multiple conditions, and therefore can be useful for systems metabolic engineering.

1.3.1 Transcriptional Network Case Studies

Transcriptional network reconstruction can utilize high-throughput experimental data for large-scale measurement of transcriptional interactions and components, such as genome-wide expression profiling and chromatin immunoprecipitation followed by microarray hybridization (ChIP-chip) [76]. Based on combinations of high-throughput experimental data, several top-down approaches for the reconstruction of transcriptional networks have been developed [77, 78]. The reconstruction of the genome-scale transcriptional network of the leucine-responsive protein (Lrp) TF in *E. coli* K-12 MG1655 is one example of the top-down approach [77]. Lrp is a global transcriptional regulator and its regulon includes genes involved in pili synthesis, amino acid biosynthesis and degradation, among other cellular functions [79, 80]. To reconstruct the network, a systems approach integrating genome-wide data from ChIP-chip for Lrp and RNA polymerase and from gene expression profiling was employed. A four-step process to reconstruct the Lrp transcriptional network was performed. First, high-resolution ChIP-chip data and expression profiles were obtained to determine the Lrp-binding regions of the genome and to measure the changes in RNA polymerase occupancies of promoters. mRNA transcript levels were used to classify the binding states under multiple environmental conditions. Second, six distinct regulatory modes were determined, including independent, concerted, and reciprocal mode, all controlled by Lrp. Third, regulatory network motifs for metabolites that are affected by the corresponding gene products were identified. Fourth, the amino acids and metabolites with the same regulatory motifs were classified, and it was determined how leucine was able to affect the regulatory motifs for the metabolites. The physiological role of the Lrp regulon was thus understood comprehensively through the reconstruction of this transcriptional network.

Another example of the top-down approach is the transcriptional network reconstruction strategy called Network Identification by multiple Regression (NIR) [78]. In this strategy, genes in a nine transcript subnetwork of the SOS pathway in *E. coli* were perturbed for down- or up-regulation, and the resulting expression profiles for all genes were measured. Then, the NIR method, using a first-order model, was applied to infer a model of the perturbed network using the expression profiles. As a result of this analysis, a first-order model of regulatory interactions in this nine transcript subnetwork of the SOS pathway was reconstructed. The inferred network provides values between genes, called connection strengths, that indicate transcriptional relationships and interactions between genes.

Bottom-up approaches have been performed to reconstruct transcriptional networks from individual experiments, databases, and literature data [65, 70]. The genome-scale network of *E. coli*'s transcriptional and translational machinery was reconstructed using information from databases, literature, and the revised *E. coli* K-12 MG1655 genome annotation [65]. The mathematical representation of the reconstruction was designated the Expression-matrix (E-matrix), representing the expression of mRNA and proteins. By implementing the stoichiometric E-matrix from the transcriptional and the translational machinery, the quantitative integration of omics data into the transcriptional and translational network is possible. This reconstructed network can also be used to compute functional states of the network. For example, the network model accurately predicted the ribosome production in *E. coli* without any parameterization, as well as the effects of the deletion of single or multiple rRNA operons [65]. To understand transcriptional regulatory networks, transcriptional interactions and dynamics can also be modeled by differential equations and stochastic models based on individual experiments and analysis of subnetworks, which provide detailed descriptions of regulatory systems and require accurate measurement of a large number of parameters for each condition [70].

Accordingly, achieving full genome-scale analysis with dynamic modeling of transcriptional networks has significant limitations [70].

The combination of top-down and bottom-up approaches has also been utilized in the reconstruction of transcriptional networks [66, 68]. An integrated genome-scale model of transcriptional regulatory and metabolic networks in *E. coli* was reconstructed based on information from literature and databases. Gene expression profiling data was also used to reconstruct the transcriptional network [66, 68]. The model was then validated and upgraded by comparing computational predictions with experimental data from growth phenotypes for multiple gene knockouts and growth on different substrates, and with gene expression data from microarray experiments [66, 68]. To incorporate a metabolic network with a transcriptional regulatory network, Boolean logic was used to represent the availability (ON) or unavailability (OFF) of genes, proteins, and reactions as binary values [69, 81]. The transcriptional regulatory network was then combined with the genome-scale metabolic network of *E. coli* in order to determine which open reading frames (ORFs) are transcribed under given conditions and aid the accurate predictions of cellular physiology and model-driven discovery [1, 22]. Methods for the prediction of gene expression, metabolic fluxes, and steady-state regulatory flux balance analysis (SR-FBA) were developed [72]. In addition, other methods, including iFBA [74] and idFBA [75], integrating metabolic, transcriptional regulatory, and signal transduction were developed.

1.4 Cell Signaling Networks

A cell signaling network is a communication network that transduces information regarding the external conditions of the cell, allowing the cell to adjust its transcriptional state accordingly. When a cell receives a signal at the external membrane, it activates a cascade of events and information flow through the cell that ultimately ends at the nucleus or the genome. It is here that the information is integrated to affect transcription. Signals from the environment are received by the cells through various means, such as chemical (e.g. chemotaxis) and physical (e.g. pressure) stimuli. Radiation can also serve as a signal to cells, as in the case of phototaxis. In multicellular organisms, cells in different parts of the body require means to send signals to each other to ensure proper function of the body. This can be accomplished through chemicals, such as hormones, dissolved in blood or other circulatory media, physical changes to the extracellular matrix, or through direct cell to cell communication. Input from these extracellular stimuli is one of the three main components of the signaling network. The other two components are the reactions that make up the signaling network from the membrane to the nucleus and the events in the nucleus affecting transcription. Through these steps, information is transduced through the cell to the nucleus so that it can be processed and allow the cell to appropriately respond to external stimuli from the environment.

Detailed mechanisms of the complete signaling network are not fully known. However, advancements in the reconstruction of signaling networks are being

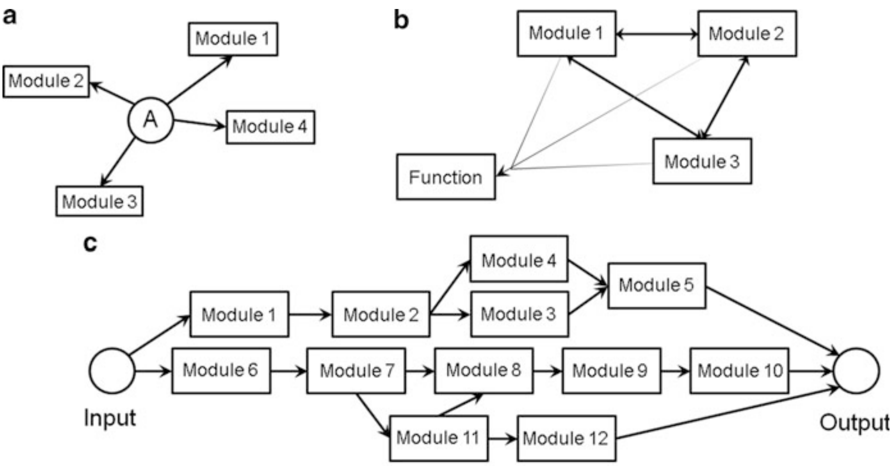
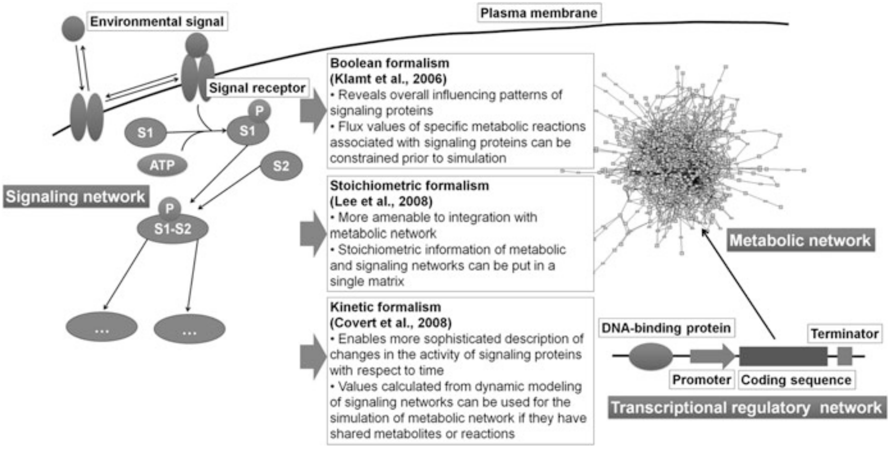


Fig. 1.3 Different strategies in the approach towards signaling network modeling. (**a**) Single signaling molecule represented by the sole node linked to various modules in the signaling network. (**b**) Different modules working together towards a common function. (**c**) Single input initiating a cascade of events that lead to an output

made. Studies have found that signaling network structure is similar to that of the metabolic network with respect to the interconnectivity and interactions between the various nodes. For instance, the degree of interconnectivity of metabolic networks is similar to that seen in the *S. cerevisiae* signaling network where there is an average of more than five protein interactions for a given protein [82–84]. One significant property of the signaling network is the combinatorial control of the components, where a few proteins can form combinations with other proteins to create receptors that can respond to a wide range of environmental stimuli [20, 85]. Therefore, while detailed studies characterizing the components of the network are being performed and have allowed for a limited level of construction of the signaling network, the full potential of the signaling network has yet to be realized.

Due to our limited knowledge of the complete signaling network, there are three different strategies for modeling signaling networks (Fig. 1.3). The first strategy involves the reconstruction of a network centered on a specific node, for example, all pathways in which the neurotransmitter acetylcholine is involved. This would encompass all paths regardless of functionality and include all roles that the node could play in the signaling network. The second strategy is the grouping of different cellular signaling components that function together under certain conditions [86, 87]. This would incorporate the interactions between different components under the specified conditions and usually includes kinetic parameters. The third strategy is the reconstruction of a signaling network consisting of a single given input and output [88].

The levels of detail in the network can also be incorporated based on the available information. The connectivity of the nodes can be either simple or complex, depending on the level of information on the mechanisms of the reactions in the signaling network (e.g. $A \rightarrow B$ as opposed to $A \rightarrow C \rightarrow B$). The reactions in



Boolean formalism (Klamt et al., 2006)
- Reveals overall influencing patterns of signaling proteins
- Flux values of specific metabolic reactions associated with signaling proteins can be constrained prior to simulation

Stoichiometric formalism (Lee et al., 2008)
- More amenable to integration with metabolic network
- Stoichiometric information of metabolic and signaling networks can be put in a single matrix

Kinetic formalism (Covert et al., 2008)
- Enables more sophisticated description of changes in the activity of signaling proteins with respect to time
- Values calculated from dynamic modeling of signaling networks can be used for the simulation of metabolic network if they have shared metabolites or reactions

Fig. 1.4 Overview of the signaling network and the different approaches utilized in the signaling network reconstruction. There is the Boolean formalism, the stoichiometric formalism and the kinetic formalism

the network can also be further detailed by the inclusion of kinetic information. With the incorporation of kinetic parameters, an additional time dimension is added to the network allowing for a better dynamic representation of the transcriptional network. Without the kinetic information, the network would consist of reactions represented by simple causal relationships. Finally, mechanistic information on the signaling reactions can be incorporated in the form of stoichiometric coefficients (e.g. 2 A + 2 B → 1 AB_BA).

High-throughput techniques for the characterization of signaling components allowing their incorporation into a signaling network reconstruction fall under two categories: (1) biochemical techniques used to characterize protein-protein interactions, and (2) assays that elucidate functional characteristics. Some examples of protein-protein interaction studies include two-hybrid systems and mass spectrometry [89, 90]. Assays include perturbation analysis [91], RNAi knockdown [92, 93], proteome analysis [94], and fluorescence labeling [95]. These methods all have their advantages and disadvantages. Therefore, combining methods to compensate for disadvantages is suggested. Success has been achieved by integrating various data types to generate a systems-level hypotheses on the nature of the interactions between several essential proteins [96].

Modeling methodologies for signaling networks deserve further discussion (Fig. 1.4). Thus far, signaling networks have been constructed based on stoichiometric and Boolean formalisms, as they do not necessitate intricate kinetic parameters, and can be easily scaled to a large size [67, 96]. In addition to these methods, dynamic or kinetic modeling and network inference using machine learning algorithms can be applied. The question then becomes, what is the best option for modeling the signaling network? Our belief is that there is no 'one-size-fits-all' solution. Each approach has its unique strengths such that they should be considered simultaneously in order to complement one another. This would help assemble separate pieces and reveal the whole picture of the signaling network as well as its interaction with other layers of biological networks, namely metabolic and transcriptional networks.

1.4.1 Signaling Network Case Studies

Despite the relative lack of detailed information on signaling networks, there have been several attempts to model them. Most of the existing signaling network models are focused on mammalian and human cells because of their sophisticated sensing systems and cell-to-cell communication. Palsson and colleagues reconstructed the largest signaling network so far for toll-like receptors, comprised of 909 reactions and 752 components [97]. Similar to procedures used in metabolic network modeling, this signaling network was reconstructed based on a stoichiometric formalism, such that flux balance analysis (FBA) could be used for simulations. Distinct input–output pathways were calculated and control points that are specific for the target pathways while not affecting other parts of the signaling network were identified.

The two-component regulatory system for signal transduction has been modeled for bacterial systems wherein sensor proteins embedded in the cell membrane sense an external signal from the environment and are phosphorylated, transmitting the information to the response regulator proteins [98]. The response regulator protein ultimately binds DNA to accordingly regulate transcription. Because of the dynamic behavior of signal transduction, stoichiometric network modeling of this system under pseudo-steady state has not been reported to our knowledge. Instead, most mathematical modeling of the two-component system resorts to kinetic modeling. Examples include phototaxis and chemotaxis of the archaeon *Halobacterium salinarum* [99], chemotaxis of *E. coli* with emphasis on the phosphatase CheZ [100], bacterial chemotaxis focused on the histidine kinase CheA [101], and the KdpD/KdpE system of *E. coli* that regulates expression of the high affinity K^+ uptake system [102]. Successful descriptions of such signaling pathways in production hosts will be of great importance in systems metabolic engineering because they may contribute to identification and optimization of unnoticed bioprocess parameters.

Aside from these studies, relatively few studies have been conducted on signaling network modeling for organisms appropriate for systems metabolic engineering when only the stoichiometric formalism and optimization-based simulations are considered. One reason would be that bacterial signaling networks are still not fully understood, despite the relative simplicity of their intracellular networks compared to eukaryotic signaling networks. Many bacterial signaling networks include signaling between other organisms in their environment, and thus can be just as complex as eukaryotes. Furthermore, the links between metabolic networks and signaling networks are not fully established. Current research has limited the integration between the two networks to specific regions, and not full networks.

Therefore, in the case of microorganisms, the focus has been on the development of integrative network models rather than independent signaling networks.

Examples of integrative modeling include the models of *E. coli* [74] and *S. cerevisiae* [75]. Both models simultaneously account for metabolic, transcriptional regulatory, and signal transduction information. In the study of *E. coli*, integrated FBA (iFBA) was developed [74], in which the stoichiometric metabolic network model of *E. coli* was integrated with a Boolean regulatory model [81] and an ordinary differential equation (ODE)-based kinetic model of *E. coli* describing the phosphotransferase (PTS) catabolite repression mechanism [103]. In this algorithm, a Boolean model of transcriptional regulation is used to constrain reactions to be active or inactive, under given condition. Then an ODE model of PTS catabolite repression is used to calculate numerical values which are passed to the model through common metabolites. iFBA was demonstrated for wild-type *E. coli* and single gene mutants for diauxic growth on glucose/lactose and glucose/glucose-6-phosphate. A significant improvement in predictive capability was found compared to individual FBA and ODE models.

Likewise, integrated dynamic FBA (idFBA), was developed and applied to the high-osmolarity glycerol response (HOG) pathway in *S. cerevisiae*, a crucial signaling pathway for adaptation to high external osmolarity [75]. In contrast to prokaryotic hosts such as *E. coli*, the consideration of signaling networks becomes more important in yeast as it has a more developed and complex signaling system. Here, unlike iFBA, signaling information was incorporated into the metabolic network through a stoichiometric formalism, thereby enabling simultaneous simulation via optimization. Another important distinction is the use of the incidence matrix with binary parameters that indicate activation or inactivation of reactions, 1 or 0 respectively, at each discretized time point. This matrix is updated progressively, producing a time-dependent dynamic simulation.

1.5 Concluding Remarks

Biological networks are complex systems that are not fully understood at our current level of knowledge. However, as compartmentalized networks such as metabolic, transcriptional, or signaling networks, become more sophisticated, we move one step closer to achieving a fully reconstructed genome-scale cellular network. While transcriptional and signaling networks do not encompass the same level of information as metabolic networks, recent studies have elucidated many characteristics of these networks that were then reincorporated into the metabolic network to aid in further understanding. Networks incorporating limited information from other networks have also been valuable in the study of biological systems. They also provide hypotheses for designing strategies to fully incorporate the different types of networks into a single network representing a complete biological system.

There should be consistent feedback between experimental and network modeling to gain better insight into biological systems. Experimental data no doubt lays the foundation for reconstructing initial versions of biological networks, which in turn generate new hypotheses that must be experimentally validated. Once validated, these hypotheses would then contribute to updating the biological network. Although we did not discuss experimental techniques in detail in this chapter, various high-throughput techniques at different biological levels, including genome, transcriptome, proteome, metabolome, and fluxome levels, deserve close attention. Pieces of information from this modeling effort, in combination with experimental data, should help to elucidate the big picture of biological systems.

Acknowledgments This work was supported by the Intelligent Synthetic Biology Center (2011-0031963) through the Global Frontier Project of Ministry of Education, Science and Technology.

References

1. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BØ (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol 3:121
2. Sohn SB, Graf AB, Kim TY, Gasser B, Maurer M, Ferrer P, Mattanovich D, Lee SY (2010) Genome-scale metabolic model of methylotrophic yeast *Pichia pastoris* and its use for *in silico* analysis of heterologous protein production. Biotechnol J 5(7):705–715
3. Gianchandani EP, Joyce AR, Palsson BØ, Papin JA (2009) Functional states of the genome-scale *Escherichia coli* transcriptional regulatory system. PLoS Comput Biol 5(6):e1000403
4. Thiele I, Fleming RM, Bordbar A, Schellenberger J, Palsson BØ (2010) Functional characterization of alternate optimal solutions of *Escherichia coli*'s transcriptional and translational machinery. Biophys J 98(10):2072–2081
5. Park JH, Lee KH, Kim TY, Lee SY (2007) Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. Proc Natl Acad Sci USA 104(19):7797–7802
6. Becker J, Zelder O, Häfner S, Schröder H, Wittmann C (2011) From zero to hero–design-based systems metabolic engineering of *Corynebacterium glutamicum* for L-lysine production. Metab Eng 13(2):159–168
7. Cho BK, Zengler K, Qiu Y, Park YS, Knight EM, Barrett CL, Gao Y, Palsson BØ (2009) The transcription unit architecture of the *Escherichia coli* genome. Nat Biotechnol 27(11):1043–1049
8. Orth JD, Palsson BØ (2010) Systematizing the generation of missing metabolic knowledge. Biotechnol Bioeng 107(3):403–412
9. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28(1):304–305
10. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. Nat Genet 25(1):25–29
11. Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38(Database issue):D355–D360