2

Bringing Genomes to Life: The Use of Genome-Scale *In Silico* Models

Ines Thiele and Bernhard Ø. Palsson

Summary

Metabolic network reconstruction has become an established procedure that allows the integration of different data types and provides a framework to analyze and map high-throughput data, such as gene expression, metabolomics, and fluxomics data. In this chapter, we discuss how to reconstruct a metabolic network starting from a genome annotation. Further experimental data, such as biochemical and physiological data, are incorporated into the reconstruction, leading to a comprehensive, accurate representation of the reconstructed organism, cell, or organelle. Furthermore, we introduce the philosophy of constraint-based modeling, which can be used to investigate network properties and metabolic capabilities of the reconstructed system. Finally, we present two recent studies that combine *in silico* analysis of an *Eschirichia coli* metabolic reconstruction with experimental data. While the first study leads to novel insight into *E. coli*'s metabolic and regulatory networks, the second presents a computational approach to metabolic engineering.

Key Words: Metabolism; reconstruction; constraint-based modeling; *in silico* model; systems biology.

1. Introduction

Over the past two decades, advances in molecular biology, DNA sequencing, and other high-throughput methods have dramatically increased the amount of information available for various model organisms. Subsequently, there is a need for tools that enable the integration of this steadily increasing amount of data into comprehensive frameworks to generate new knowledge and formulate hypotheses about organisms and cells. Network reconstructions of biological systems provide such frameworks by defining links between the network components in a bottom-to-top approach. Various types of "omics" data can be used to identify the list of network components and their interactions. These network reconstructions represent biochemically, genetically, and genomically

Table 1. Organisms and network properties for which genome-scale metabolic reconstructions have been generated.

	ORFs	SKI	N_G	N_M	N_R	Ref
BACTERIA						
Bacillus subtilis	4,225	4.8	614	637	754	19
Escherichia coli	4,405	55.1	904	625	931	20
			720	438	627	21
Geobacter sulfurreducens	3,530		588	541	523	22
Haemophilus influenzae	1,775	8.9	296	343	488	23
			400	451	461	24
Helicobacter pylori	1,632	13	341	485	476	25
			291	340	388	26
Lactococcus lactis	2,310		358	422	621	27
Mannheimia succiniproducens	2,463		335	352	373	28
Staphylococcus aureus	2,702	16	619	571	641	29
Streptomyces coelicolor	8,042	0.13	700	500	700	30
ARCHAEA						
Methanosacrcina barkerii	5,072		692	558	619	31
EUKARYA						
Mus musculus	28,287	15.6	1,156[b]	872	1,220	32
Saccharomyces cerevisiae	6,183	10.6	750	646	1,149	33
			708	584	1,175	34

Listed is the number of open reading frames (ORF) of each organism, the number of genes included in the reconstruction (N_G), as well as the number of metabolites (N_M) and reactions (N_R) in the metabolic network. The Species Knowledge Index (SKI) (1) is a measure of the amount of scientific literature available for an organism. Adapted from Reed (18).

(BIGG) structured databases that simultaneously integrate all component data, and can be used to visualize and analyze further high-throughput data, such as gene expression, metabolomics, and fluxomics data.

There are at least three ways to represent BIGG databases: (i) textual representation, which allows querying of its content; (ii) graphical representation, which allows the visualization of the network interactions and their components; and (iii) mathematical representation, which enables the usage of a growing number of analytical tools to characterize and study the network properties. Several metabolic reconstructions have been published recently, spanning all domains of life (Table 1), and most of them are publicly available.

In this chapter, we will first define the general properties of a biological system, and then learn to how to reconstruct metabolic networks. The second part of the chapter will introduce the philosophy of constraint-based modeling and highlight two recent research efforts that combined experimental and computational methods. Although this chapter concentrates on metabolic reconstructions, networks of protein–protein interactions, protein–DNA interactions, gene regulation, and cell signaling can be reconstructed using similar rules and techniques. The general scope of this chapter is illustrated in Figure 1, which represents the main process of "bringing genomes to life."

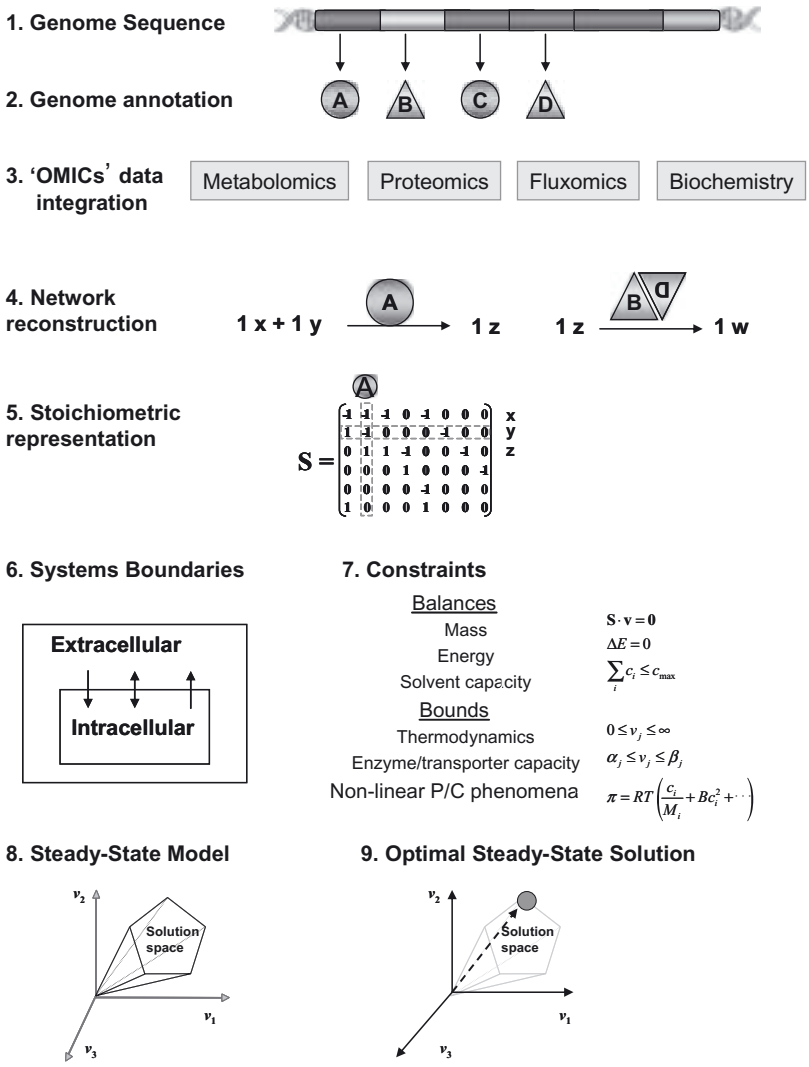


Figure 1. Bringing genomes to life. This figure illustrates the main outline of the chapter and the general approach to network reconstruction and analysis. Starting from the genome sequence, an initial component list of the network is obtained. Using additional data such as biochemical and other omics data the initial component list is refined as well as information about the links between the network components. Once the network links, or reactions, are formulated, the stoichiometric matrix can be constructed using the stoichiometric coefficients that link the network components. The definition of the system boundaries transforms a network reconstruction into a model of a biological system. Every network reaction is elementary balanced and may obey further constraints (e.g., enzyme capacity). These constraints allow the identification of candidate network solutions, which lie within the set of constraints. Different mathematical tools can be used to study these allowable steady-state network states under various aspects such as optimal growth, byproduct secretion and others.

2. Properties of Biological Networks

In this section, we will discuss general properties of biological systems and how these can be used to define a general scheme that describes biological systems in the terms of the components and links of the network.

2.1. General Properties of Biological Systems

The philosophy of network reconstruction and constraint-based modeling is based on the fact that there are general principles any biological system has to obey. Because the interactions, or links, between network components are chemical transformations, they are based on principles derived from basic chemistry. First, in living systems, the prototypical transformation is bilinear at the molecular level. This association involves two compounds coming together to either be chemically transformed through i) the breakage or formation of covalent bonds, as is typical for metabolic reactions and reactions of the macromolecular synthesis,

$$X + Y \leftrightarrow X - Y \quad \text{covalent bonds}$$

or ii) two molecules associate together to form a complex that may be held together by hydrogen bonds and/or other physical association forces to form a complex, which has a different functionality from the individual components:

$$X + Y \leftrightarrow X{:}Y \quad \text{association of molecules.}$$

An example of the latter association is the binding of a transcription factor to DNA to form an activated transcription site that enables the binding of the RNA polymerase.

Second, the reaction stoichiometry is fixed and described by integer numbers counting the molecules that react and that are formed as a consequence of the chemical reaction. Chemical transformations are constrained by elemental and charge balancing, as well as other features. The stoichiometry is invariant between organisms for the same reactions, and it does not change with pressure, temperature, or other conditions. Therefore, stoichiometry gives the primary topological properties of a biochemical reaction network.

Third, all reactions inside a cell are governed by thermodynamics. The relative rate of reactions, forward and backward, is therefore fixed by basic thermodynamic properties. Unlike stoichiometry, thermodynamic properties do change with physicochemical conditions, such as pressure and temperature. In addition, the thermodynamic properties of association between macromolecules can be changed, for example, by altering the sequence of a protein or the base-pair sequence of a DNA-binding site.

Fourth, in contrast to stoichiometry and thermodynamics, the absolute rates of chemical reactions inside cells are evolutionarily malleable. Cells can thus extensively manipulate the rates of reactions through changes in their DNA sequence. Highly evolved enzymes are very specific in catalyzing particular chemical transformations.

These rules dictate that cells cannot form new links at will, and candidate links are constrained by the nature of covalent bonds and by the thermodynamic nature of interacting macromolecular surfaces. All of these are subject to the basic rules of chemistry and thermodynamics. Furthermore, intracellular conditions restrict the activity of systems, such as physicochemical conditions, spatiotemporal organization of cellular components, and the quasicrystalline state of the cell.

2.2. Steady-State Networks

Biological systems exist in a steady state, rather than in equilibrium. In a steady-state system, flow into a node is equal to flow out of a node. Consequently, depletion or accumulation in a steady-state network is not allowed, which means that a produced compound has to be consumed by another reaction. If this is not the case, the corresponding compound represents a network gap (or dead end), and its producing reaction is called a blocked reaction because no flux through this reaction is possible.

3. Reconstruction of Metabolic Networks

The genome annotation, or 1D annotation, provides the most comprehensive list of components in a biological network. In metabolic network reconstructions, the genome annotation is used to identify all potential gene products involved in the metabolism of an organism. By using more types of information, such as biochemical, physiological, and phenotype data, the interaction of these components will be defined. Subsequently, we will refer to network reconstructions as 2D genome annotation because the network links defined in the network reconstruction represent a second dimension to the 1D genome annotation.

3.1. Sources of Information

1D genome annotations are one of the most important information sources for reconstructions because they provide the most comprehensive list of network components. However, one has to keep in mind that without biochemical or physiological verification, the 1D annotation is merely a hypothesis.

The links in metabolic networks are the reactions carried out by metabolic gene products. To assign cellular components with the metabolic reactions, different information is required and provided by various sources. Organism-specific and non–organism-specific databases contain a vast amount of data regarding gene function and associated metabolic activities. Especially valuable are organism-specific literature providing information on the physiological and pathogenic properties of the organism, along with biochemical characterization of enzymes, gene essentiality, minimal medium requirements, and favorable growth environments. Although biochemical data are used during the initial reconstruction effort to define metabolic reactions, organism-specific information such

as medium requirements and growth environment can be used to derive transport reactions when not provided by the 1D genome annotations. In addition, gene essentiality data can be used during the network evaluation process to compare and validate the reconstruction. Physiological data, such as medium composition, secretion products, and growth performance, are also needed for the evaluation of the reconstruction and can be found in primary literature or can be generated experimentally. Phylogenetic data can substitute organism-specific information when a particular organism is not well studied, but has a close relative that is. In addition, cellular localization of enzymes can be found in studies that use immunofluorescence or GFP-tagging for individual proteins to identify their place of action. Alternatively, there are several algorithms predicting a protein's compartmentalization based on localization signal sequences.

Because some of these information sources are more reliable than others, a confidence scoring system may be used to distinguish them.

3.2. How to Choose an Organism to Reconstruct

The amount of information available differs significantly from organism to organism; therefore, the choice of organism to reconstruct is critical for the quality of the final reconstruction. Because the genome annotation serves as a first parts list in most reconstruction efforts, its availability and high quality are primary criteria. Furthermore, the quantity of primary and review publications available for metabolism should be considered. A good estimate of legacy data available for an organism can be obtained with the Species Knowledge Index (SKI) (1). This SKI value is a measure of the amount of scientific literature available for an organism, calculated as the number of abstracts per species in PubMed (National Center for Biotechnology Information) divided by the number of genes in the genome (see Table 1 for some SKI values of reconstructed organisms). Finally, organism-specific databases maintained by experts can be very valuable sources of information during the reconstruction process.

3.3. Formulation of Model

The translation of a 1D genome annotation into a metabolic network reconstruction can be done in a step-wise fashion by incorporating different types of data. First, relevant metabolic genes have to be identified from the 1D annotation. The gene functions have to be translated in elementary and charged balanced reactions. Next, the network is assembled by considering each metabolic pathway separately and by filling in missing reactions as necessary. When this first version of the network reconstruction is finished, the reconstruction will be tested *in silico* and compared with physiological data to ensure that it has the same metabolic capabilities as the cell *in vivo*. This latter step might identify further reactions that need to be included, whereas other ones will be replaced or their directionality might be changed. It is important to remember that the sequence-derived list of metabolic enzymes cannot be assumed to be complete because of the large numbers of open reading

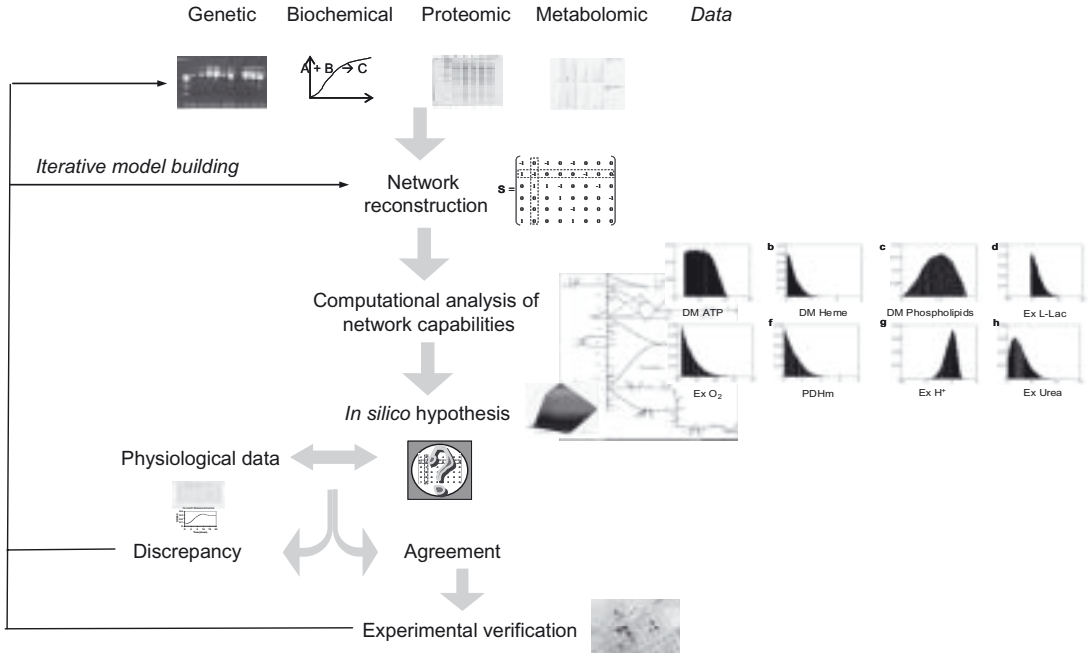


Figure 2. The iterative process of network reconstruction. Normally, several iterations of reconstruction are necessary to ensure quality and accuracy of the reconstructed network. After an initial reconstruction, accounting for the main components identified by the different sources of information, is obtained, the reconstruction will be tested for its ability to produce certain metabolites such as biomass precursors. Comparison with experimental data, like phenotypical and physiological data, will help to identify any discrepancy between *in silico* and *in vivo* properties. The iterative re-evaluation of legacy data and network properties will eventually lead to a refined reconstruction.

frames (ORFs) still having unassigned functions. The iterative process of network reconstruction and evaluation will lead to further refinement of reconstruction (Figure 2).

3.3.1. Defining Biochemical Reactions

The biochemical reaction carried out by a gene product can be determined in five steps (Figure 3). First, the substrate specificity has to be determined because it can differ significantly between organisms. In general, one can distinguish between two groups of enzymes based on their substrate specificity. The first group of enzymes can only act on a few highly similar substrates, whereas the second group recognizes a class of compounds with similar functional groups; thus, the enzymes have a broader substrate specificity. The substrate specificity of either type of these enzymes may differ across organisms for primary metabolites, as well as for coenzymes (such as NADH vs. NADPH and ATP vs. GTP). Often, it is very difficult to derive this information solely from the gene sequence because substrate- and coenzyme-binding sites might be similar for related compounds.

Once the metabolites and coenzymes of an enzyme are identified, the charged molecular formula at a physiologically relevant pH has to be calculated, as a second step. In general, a pH of 7.2 is used in the reconstruction. However, the pH in some organelles can differ from the rest of the cell, as is the case for peroxisomes, where the pH has been reported to be between 6 and 8 (2,3). The pK_a value for a given compound can be used to determine its degree of protonation.

Third, the stoichiometry of the reaction needs to be specified. As in basic chemistry, reactions need to be charge and mass balanced, which may lead to the addition of protons and water.

The fourth step adds basic thermodynamic considerations to the reaction, defining its reversibility. Biochemical characterization studies will sometimes test the reversibility of enzyme reactions, but the directionality can differ between *in vitro* and *in vivo* environments because of differences in temperature, pH, ionic strength, and metabolite concentrations.

The fifth step requires reactions and proteins to be assigned to specific cellular compartments. This task is relatively straightforward for

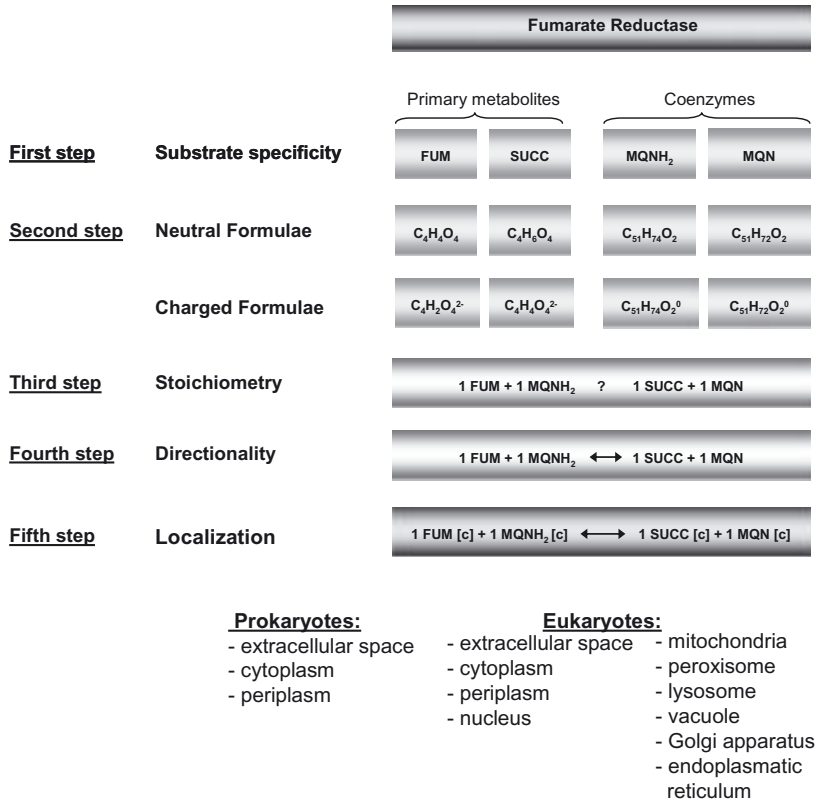


Figure 3. The five steps to formulate a biochemical reaction. The reaction carried out by a metabolic gene product can be determined by the five depicted steps. Here, we show the example of the fumarate reductase of *E. coli*, which converts fumarate (FUM) into succinate (SUCC) using menaquinone (MQN) as electron donor.

prokaryotes, which do not exhibit compartmentalization, but becomes challenging for eukaryotes, which may have up to 11 subcellular compartments (Figure 3). Incorrect assignment of the location of a reaction can lead to additional gaps in the metabolic network and misrepresentation of the network properties. In the absence of experimental data, proteins should be assumed to reside in the cytosol to reduce the number of intracellular transport reactions, which are also often hypothetical and therefore have a low confidence score.

3.3.2. Assembly of Metabolic Network Reconstruction

Once the network reactions are defined, the metabolic network can be assembled in a step-wise fashion by starting with central metabolism, which contains the fueling reactions for the cell, and moving on to the biosynthesis of individual macromolecular building blocks (e.g., amino acids, nucleotides, and lipids). The step-wise assembly of the network facilitates the identification of missing steps within the pathway that were not defined by the 1D annotation. Once well-defined metabolic pathways are assembled, reactions can be added that do not fit into these pathways, but are supported by the 1D annotation or biochemical studies. Such enzymes might be involved in the utilization of other carbon sources or connect different pathways.

3.3.3. Gap Analysis

Even genomes of well-studied organisms harbor genes of unknown functions (e.g., 20% for *E. coli*). Subsequently, metabolic networks constructed solely on genomic evidence often contain many network gaps, so-called blocked reactions. Physiological data may help to determine whether a pathway is functional in the organism, and thus may provide evidence of the missing reactions. This procedure is called gap filling, and it is a crucial step in network reconstruction. For example, if proline is a nonessential amino acid for an organism, then the metabolic network should contain a complete proline biosynthesis pathway, even if some of the enzymes are not in the current 1D annotation. In contrast, if another amino acid, let's say methionine, is known to be required in the medium, then the network gap should not be closed, even if only one gene is missing. In this case, filling the gap would significantly change the phenotypical *in silico* behavior of the reconstruction.

These examples show that physiological data of an organism provide important evidence for improving, refining, and expanding the quality and content of reconstructed networks. Reactions added to the network at this stage should be assigned low confidence scores if there are no genetic or biochemical data available to confirm them. Subsequently, for each added reaction, putative genes can be identified using homology-based and context-based computational techniques. Such added reactions and putative assignments form a set of testable hypotheses that are subject to further experimental investigation. Because the reconstructed network integrates many different types of data available for an organism, its completeness also reflects the knowledge about the organism's metabolism. Remaining unsolved network gaps involving blocked reactions or dead-end metabolites reflect these knowledge gaps.

3.3.4. Evaluation of a Network Reconstruction

Network evaluation is a sequential process (Figure 3). First, the network is examined to see if it can generate the precursor metabolites, such as biomass components, and metabolites the organism is known to produce or degrade. Second, network gaps have to be identified and metabolic pathways may need to be completed based on physiological information. Finally, the comparison of the network behavior with various experimental observations, such as secretion products and gene essentiality, will ensure similar properties and capabilities of the *in silico* metabolic network and the biological system. This sequential, iterative process of network evaluation is labor intensive, but it will ensure high accuracy and quality by network adjustments, refinements, and expansions.

3.4. Automating Network Reconstruction

The manual reconstruction process is laborious and can take up to a year for a typical bacterial genome, depending on the amount of literature available. Hence, efforts have been undertaken to automate the reconstruction process. Like most manually assembled reconstructions, most automatic reconstruction efforts start from the annotation. For example, Pathway Tools (4) is a program that can automate a network reconstruction using metabolic reactions associated with Enzyme Commision numbers (5) and/or enzyme names from a 1D genome annotation. To overcome missing annotations, Pathway Tools has the option to include missing gene products and their reactions in a pathway if a significant fraction of the other enzymes are functionally assigned to this pathway in the genome annotation. As for the manually curated reconstruction, the automated gap filling procedure has to be done with caution, as the inclusion of reactions without confidence may alter the phenotypical outcome of the reconstruction.

Although the automation of reconstruction is necessary on a larger scale, the results of these informatics approaches are limited by the quality of the information on which they operate. Therefore, automated reconstructions need detailed evaluation to assure their accuracy and quality. Frequent problems with these automated reconstructions involve incorrect substrate specificity, reaction reversibility, cofactor usage, treatment of enzyme subunits as separate enzymes, and missing reactions with no assigned ORF. Although an initial list of genes and reactions can be easily obtained by using the automated methods, a good reconstruction of biological networks demands the understanding of properties and characteristics of the organism or the cell. Because the number of experimentally verified gene products and reactions is limited for most organisms, knowledge about the metabolic capabilities of the organism is crucial.

4. Mathematical Characterization of Network Capabilities

In this section, we briefly illustrate the general philosophy of the constraint-based modeling approach that resulted in a growing number of mathematical tools to interrogate a reconstructed network. The method

relies primarily on network stoichiometry, and thus it is not necessary to define kinetic rate constants and other parameters, which are difficult or impossible to determine accurately in the laboratory. A more comprehensive description of the different tools can be found in Palsson's work (6) and in a recently published review (7).

4.1. Stoichiometric Representation of Network

The stoichiometric matrix, denoted as S, is formed by the stoichiometric coefficients of the reactions that comprise a reaction network (Figure 1 and Figure 4). This matrix is organized such that every column corresponds to a reaction, and every row corresponds to a compound. The matrix entries are integers that correspond to the stoichiometric coefficients of the network reactions. Each column describes a reaction, which is constrained by the rules of chemistry, such as elementary balancing. Every row describes the reactions in which a compound participates, and therefore how the reactions are interconnected.

Mathematically, the stoichiometric matrix, S, transforms the flux vector v, which contains the reaction rates, into a vector that contains the time derivatives of the concentrations. The stoichiometric matrix, thus contains chemical and network information. Mathematically spoken, the stoichiometric matrix S is a linear transformation of the flux vector,

$$v = (v_1, v_2, \ldots, v_n),$$

to a vector of time derivatives of the concentration vector,

$$x = (x_1, x_2, \ldots x_n),$$

as

$$dx/dt = S.v.$$

At steady state, there is no accumulation or depletion of metabolites in a metabolic network, so the rate of production of each metabolite in the network must equal its rate of consumption. This balance of fluxes can be represented mathematically as

$$S.v = 0.$$

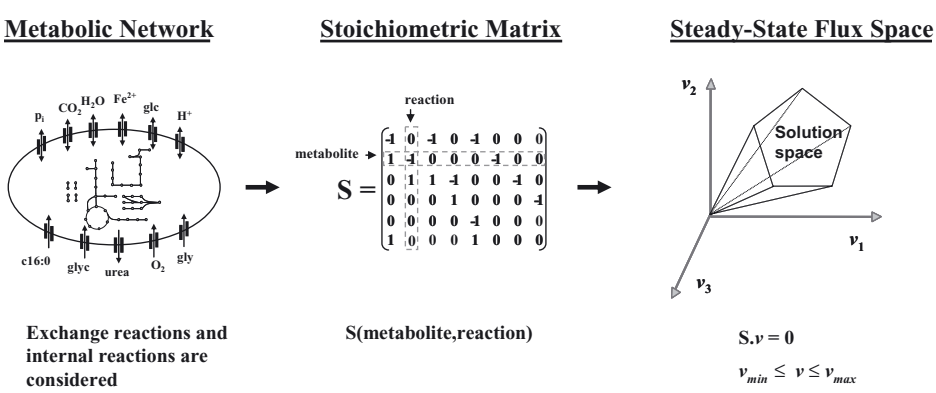


Figure 4. Matrix representation of metabolic network.

Bounds that further constrain the values of individual variables can be identified, such as fluxes, concentrations, and kinetic constants. Upper and lower limits can be applied to individual fluxes, such that

$$v_{i,min} \leq v_i \leq v_{i,max}.$$

For elementary (and irreversible) reactions, the lower bound is defined as $v_{min} = 0$. Specific upper limits (v_{max}) that are based on enzyme capacity measurements are generally imposed on reactions.

4.2. Reconstruction Versus Model

The network reconstruction represents the framework for a biological model. The definition of systems boundaries provides the transition from a network reconstruction to a model. These systems boundaries can be drawn in various ways (Figure 5). Typically, the systems boundaries are drawn around the cell, which is consistent with a physical entity, and the resulting model can be used to investigate properties and capabilities of the biological system. However, it might be useful to draw "virtual" boundaries to segment the network into subsystems (e.g., nucleic acid synthesis or fatty acid synthesis).

The "physical" systems boundaries are drawn to distinguish between the inside metabolites of the cell to the outside metabolites and thus, correspond to the cell membrane. Reactions that connect the cell and its environment are called exchange reactions. These exchange reactions allow the exchange of metabolites in and out of the cell boundaries.

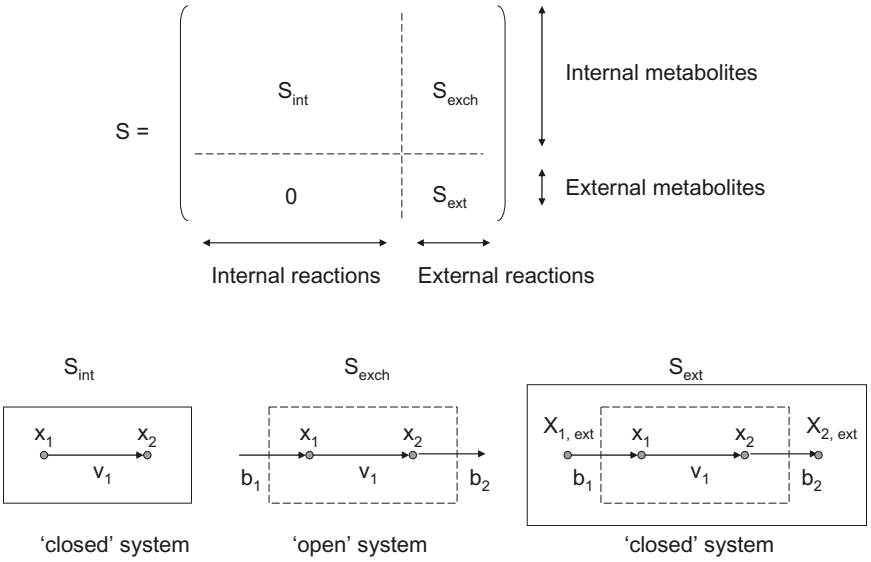


Figure 5. Systems Boundaries. The network reactions are partitioned in internal (int) and external (ext) reactions. The exchange fluxes are denoted by b_i and internal fluxes by v_i.

The stoichiometric matrix S (or S_{tot}) can be partitioned such that there are three fundamental subforms of S_{tot}: i) the exchange stoichiometric matrix (S_{exch}), which does not consider external metabolites and only contains the internal fluxes and the exchange fluxes with the environment; ii) the internal stoichiometric matrix (S_{int}), which considers the cell a closed system; and iii) the external stoichiometric matrix (S_{ext}), which only contains external metabolites and exchange fluxes (Figure 5). These different forms of S can be used to study topological properties of the network. For example, S_{exch} is frequently used in pathway analysis (extreme pathway analysis), whereas S_{int} is useful to define pools of compounds that are conserved within the network (e.g., currency or secondary metabolites such as ATP, NADH, and others).

4.3. Identification of Constraints

Cellular functions are limited by different types of constraints, which can be grouped in four general categories: fundamental physicochemical, spatial or topological, condition-dependent environmental, and regulatory or self-imposed constraints. Although the first two categories of constraints are assumed to be independent from the environment, the latter two may vary in the simulation.

4.3.1. Physicochemical Constraints

Many physicochemical constraints are found in a cell. These constraints are inviolable and provide "hard" constraints on cell functions because mass, energy, and momentum must be conserved. For example, the diffusion rates of macromolecules inside a cell are generally slow because the contents of a cell are densely packed and form a highly viscous environment. Reaction rates are determined by local concentrations inside the cell and are limited by mass transport beside their catalytic rates. Furthermore, biochemical reactions can only proceed in the direction of a negative free-energy change. Reactions with large negative free-energy changes are generally irreversible. These physicochemical constraints are normally considered when formulating the network reactions and their directions.

4.3.2. Spatial Constraints

The cell content is highly crowded, which leads to topological, or spatial, constraints that affect both the form and the function of biological systems. For example, bacterial DNA is about 1,000 times longer than the length of a cell. Thus, on one hand, the DNA must be tightly packed in a cell without becoming entangled; however, on the other hand, the DNA must also be accessible for transcription, which results in spatial-temporary pattern. Therefore, two competing needs, which are the packaging and the accessibility of the DNA, constrain the physical arrangement of DNA in the cell. Incorporating these constraints is a significant challenge for systems biology.

4.3.3. Environmental Constraints

Environmental constraints on cells are time and condition dependent. Nutrient availability, pH, temperature, osmolarity, and the availability of electron acceptors are examples of such environmental constraints. This

group of constraints is of fundamental importance for the quantitative analysis of the capabilities and properties of organisms because it allows determining their fitness, or phenotypical properties, under various environmental settings. Because the performance of an organism varies under different environmental conditions, data from various laboratories can only be compared and integrated when the experimental conditions, such as medium composition, are well documented. In contrast, laboratory experiments with undefined media composition are often of limited use for quantitative *in silico* modeling.

4.3.4. Regulatory Constraints

Regulatory constraints differ from the three categories discussed above, as they are self-imposed and subject to evolutionary change. For this reason, these constraints may be referred to as regulatory constraints, in contrast to hard physicochemical constraints and time-dependent environmental constraints. On the basis of environmental conditions, regulatory constraints allow the cell to eliminate suboptimal phenotypic states. Regulatory constraints are implemented by the cell in various ways, including the amount of gene products made (transcriptional and translational regulation) and their activity (enzyme regulation).

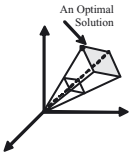
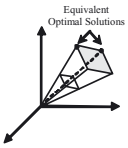
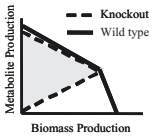
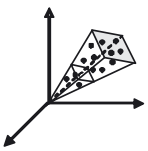
4.4. Tools For Analyzing Network States

The analysis of an organism's phenotypic functions on a genome scale using constraint-based modeling has developed rapidly in recent years. A plethora of steady-state flux analysis methods can be broadly classified into the following categories: i) finding best or optimal states in the allowable range; ii) investigating flux dependencies; iii) studying all allowable states; iv) altering possible phenotypes as a consequence of genetic variations; and v) defining and imposing further constraints. In this section, we will discuss some of the numerous methods that have been developed (Table 2). A more comprehensive list of methods can be found in Price's work (7).

4.4.1. Optimal or Best States

Mathematical tools, such as linear optimization, can be used to identify metabolic network states that maximize a particular network function, such as biomass, ATP production, or the production of a desired secretion product. The objective function can be either a linear or nonlinear function. For linear functions, linear optimization or linear programming (LP) can be used to calculate one optimal reaction network state under the given set of constraints. Growth performance of an organism can be assessed by calculating the optimal (growth) solution under different medium conditions. Using visual tools, such as metabolic maps, the optimal network state can be easily accessed and compared. This mathematical tool has been widely used for the identification of optimal network states for the objective function of interest. Interestingly, for genome-scale networks in particular, there can be multiple network states or flux distributions with the same optimal value of the objective function; therefore the need for enumerating alternate optima arises.

Table 2. List of constraint-based modeling methods.

Analysis Method	Illustration	Applied metabolic networks	References
Optimal solutions		***Escherichia coli***	**35**
Alternate Optima		***Escherichia coli*, human cardiac myocyte mitochondrion**	**8, 36, 37**
OptKnock		***Escherichia coli***	**12, 38**
Sampling		**Red blood cell, *Helicobacter pylori*, human cardiac myocyte mitochondrion**	**39, 42**

A myriad of analytical methods have arisen over recent years. The methods discussed in this chapter are depicted in this table along with some metabolic networks that have been applied to study network properties. Redrawn from Price (7).

4.4.2. Alternate Optima

Alternate optima are a set of flux distributions that represent equally optimal network states given any particular objective function. The number of such alternate optima varies depending on the size of the metabolic network, the chosen objective function, and the environmental conditions. In general, the larger and more interconnected the network, the higher the number of alternate optimal phenotypes. A recursive mixed-integer LP algorithm has been developed to exhaustively enumerate all alternate optima (8). Genome-scale metabolic networks contain several redundant pathways, which makes the enumeration of all optima computationally challenging.

4.4.3. OptKnock

OptKnock is a bilevel optimization algorithm to computationally predict gene deletion strategies for byproduct overproduction, such as succinate, lactate, and amino acids. The OptKnock algorithm calculates solutions that simultaneously optimize two objective functions, which are biomass formation and secretion of a target metabolite. Multiple gene deletions can be introduced in the metabolic network, such that the fluxes through reactions of the target metabolite are optimally used, while reactions leading to other byproducts from common precursors are deleted from the network. The premise underlying this bilevel optimization algorithm

is that overproduction of target metabolites can be achieved by altering the structure of the metabolic network through gene deletions. With this direct stoichiometric coupling of target metabolite production to biomass, it is hypothesized that an increase in growth rate should concurrently result in an increase in the target metabolite production rate.

4.4.4. Unbiased Modeling

In addition to the above listed examples of optimization-based methods, non–optimization-based techniques have also been developed to study the full range of achievable metabolic network states that are provided by the solution spaces. These methods enable the user to determine not only the solutions selected by the statement of an objective, but all the solutions in the space. The results are therefore not biased by a statement of an objective, but indicate properties of the genome-scale network as a whole. Uniform random sampling is one example of an unbiased method. Here, the solution space is sampled by calculating uniform, random points within the space. The content of a solution space can be studied by the set of uniform random sampling of points within the space. The sampling points describe candidate metabolic states that are in agreement with the imposed constraints. The projection of the sampling point into a 2D diagram results in a flux distribution for every reaction in the network that can be understood as a probability distribution of flux values for every reaction.

The methods described in this section have been successfully used to characterize and investigate the network capabilities of numerous genome-scale metabolic networks. Until recently, it has focused on the steady-state flux distributions through a reconstructed network, but is now being used to study all allowable concentration and kinetic states (9).

5. Two Sample Studies

In this section, we will highlight two studies that combined *in silico* analysis and experimental data to gain new insight into the metabolism of *E. coli.*

5.1. "Integrating High-throughput and Computational Data Elucidates Bacterial Networks" (10) (Figure 6)

Regulatory constraints are used by cells to control the expression state of genes, leading to distinct sets of expressed genes under different environmental conditions. Assuming the expression state of a gene can be only on or off (expressed or depressed), the regulation of genes can be represented in the form of Boolean rules (on or off, 1 or 0).

For the purpose of this study, the regulatory rules for the metabolic genes included in *i*JR904 (11) were created and incorporated based on literature and databases. The resulting reconstruction, MC1010v1, was the first integrated genome-scale *in silico* reconstruction of a transcriptional regulatory and metabolic network. MC1010v1 accounted for 1,010 genes in *E. coli*, including 104 regulatory genes whose products, together with other stimuli, regulate the expression of 479 of the 904 genes in the reconstructed metabolic network.

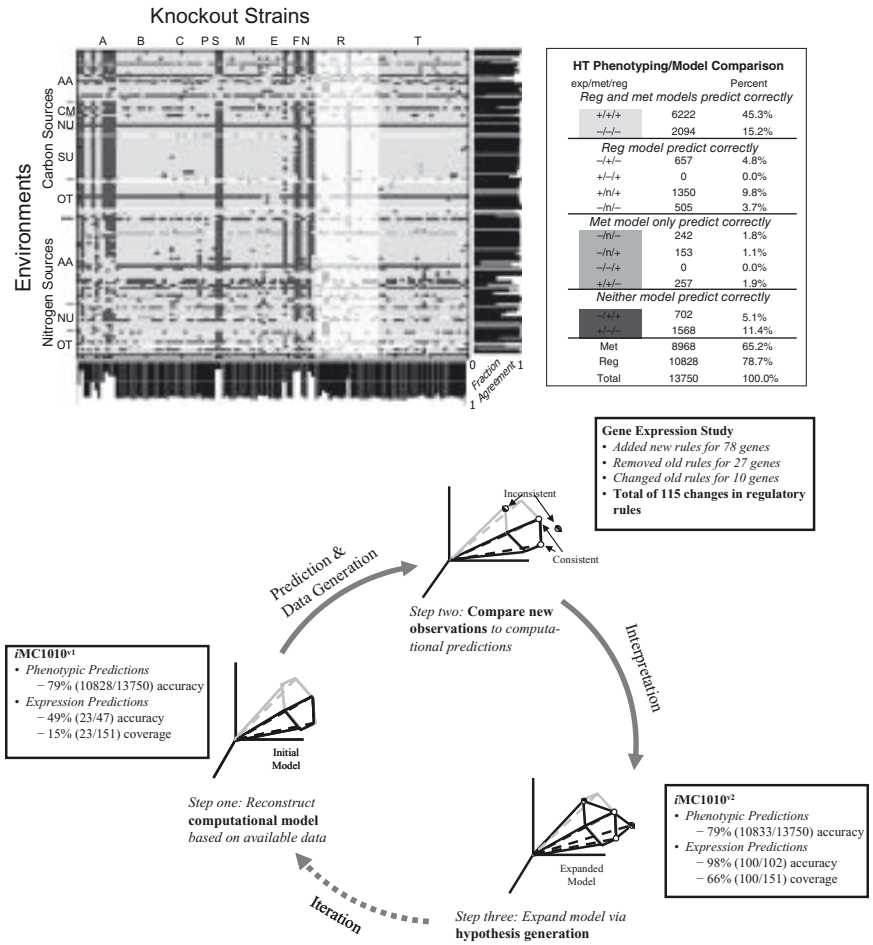


Figure 6. "Integrating high-throughput and computational data elucidates bacterial networks." Top Panel: Comparison of high-throughput phenotyping array data with the *in silico* predictions for the *E. coli* network, with (Reg) and without (Met) regulatory constraints. Each case lists the results of the experimental data (exp), metabolic model (met) and regulatory metabolic model (reg). "+": predicted or observed growth, "–": no growth, and 'n': for cases involving a regulatory gene knockout not predictable by the metabolic model.
Bottom Panel: Metabolic and regulatory networks may be expanded by using high-throughput phenotyping and gene expression data coupled with the predictions of a computational model. The accuracy refers to the percentage of model predictions that agreed with experimental data; the coverage indicates the percentage of experimental changes predicted correctly by the model. Redrawn from (10).

To determine the importance of regulatory rules on the predictive potential of the metabolic reconstruction, both reconstructions, *i*JR904 (unregulated metabolic network) and MC1010v1 (regulated metabolic network), were used to calculate *in silico* growth performance under different medium conditions and to assess the outcome of gene deletion to the growth performance. The *in silico* results were compared with the outcomes of high-throughput growth phenotyping and gene expression

experiments (Figure 6). Based on these results, several substrates and knockout strains were found whose growth behavior did not match predictions. Further investigation of these conditions and strains led to the identification of five environmental conditions in which dominant, yet uncharacterized, regulatory interactions actively contributed to the observed growth phenotype. In addition, five environmental conditions and eight knockout strains were identified that highlight uncharacterized enzymes or noncanonical pathways and that are predicted to be used by this study. Furthermore, the results indicated that some transcription factors that were involved in the regulation differed from previously reported data. These new rules were incorporated in the reconstruction leading to a second version, MC1010v2, which could successfully predict the outcome of high-throughput growth phenotyping and gene expression experiments.

The results of this study, and the iterative modification of the regulatory rules, led to two main observations. First, some of the results of the knockout perturbation analysis are complex enough to make Boolean rule formulation difficult. Second, many of these gene expression changes involve complex interactions and indirect effects. Transcription factors may be affected, for example, by the presence of fermentation byproducts or the buildup of internal metabolites. Such effects would be extremely difficult to identify or account for without a computational model.

This study showed that the reconciliation of high-throughput data sets with genome-scale computational model predictions enables systematic and effective identification of new components and interactions in microbial biological networks.

5.2. "In Silico Design and Adaptive Evolution of Escherichia coli for Production of Lactic Acid" (12) (Figure 7)

In this study, OptKnock was used to design candidate knockout mutations *in silico*, which were subsequently analyzed and verified experimentally. The overall goal was to create an *E. coli* mutant that could overproduce lactic acid in minimal medium supplemented with glucose. In contrast, *E. coli* wild type produces only traces of lactate under this medium condition. Other studies already engineered lactate-overproducing *E. coli* mutants; however, in this study it was shown how to use metabolic reconstructions to successfully engineer stable mutants.

The most recent reconstruction of *E. coli*'s metabolism, *i*JR904 (11), was used by the OptKnock algorithm to identify the possible solutions that induce *E. coli* to secrete lactic acid as a byproduct during optimal cellular growth. For this purpose single, double, triple, and quadruple gene deletions were designed *in silico* and tested for bioptimal production of lactic acid and growth yield. Based on these calculations, three different designs for production of lactate were selected for experimental verification: (i) pta-adhE double-deletion strain, (ii) pta-pfk double-deletion strain, and (iii) pta-adhE-pfk-glk quadruple deletion strains (pta, phosphate acetyltransferase; adhE, acetaldehyde dehydrogenase; pfk, 6-phosphofructokinase; glk, glucokinase) (Figure 7).

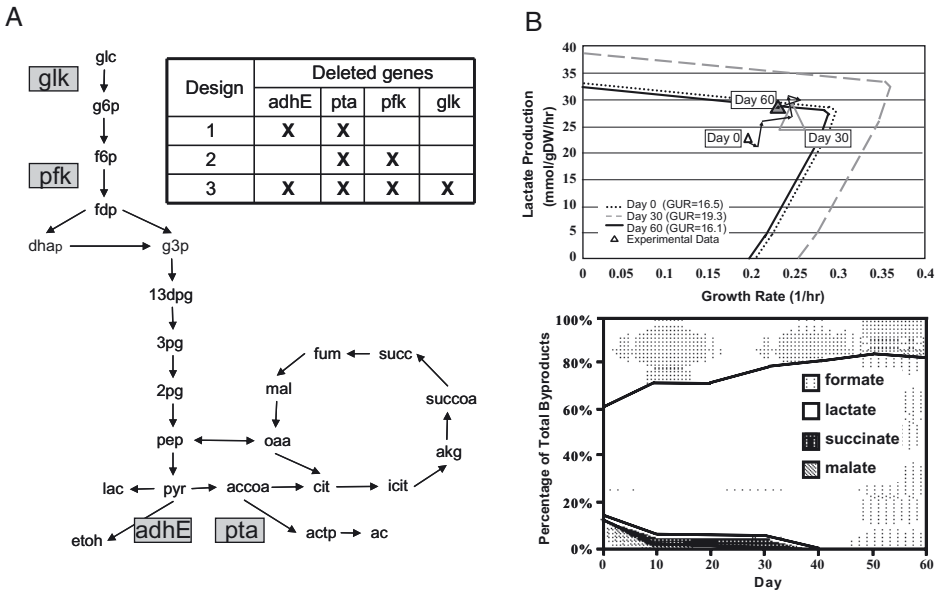


Figure 7. "In Silico Design and Adaptive Evolution of *Escherichia coli* for Production of Lactic Acid." A: schematic picture of the pathways in which the gene deletions are involved. B: Strain design 1 ($adhE^{-}$,pta^{-}). **Top panel:** growth performance and lactate secretion at the beginning of adaptive evolution (day 0), in the middle (day 30) and at the end (day 60). The computational predictions (lines) were done based on the glucose uptake rate (GUR) measured in the deletion strain at the different time points. The **bottom panel** shows the byproduct secretion rates for the mutant strain during the course of adaptive evolution. It is easily visible that lactate becomes the main fraction of byproduct. Redrawn from (12).

Predicted strain designs were constructed *in vivo* and evolved over 60 d. Over this time period, the growth rates of constructed strains and the byproduct secretion rates were monitored. By measuring these growth rates and lactic acid secretion rates, as well as the glucose uptake rates, the experimental phenotypes could be directly compared to the computationally predicted possible solutions for each design. Both the pta-adhE strains and the pta-pfk strains showed good agreement with the computationally determined solution spaces. In all cases, the byproduct secretion profiles stabilized after approximately 20 d of adaptive evolution, with all strains showing sustained elevated lactic acid titers throughout the course of adaptive evolution over the wild-type strain.

The goal of this study was to experimentally test computationally predicted strain designs calculated from a genome-scale metabolic model using the OptKnock algorithm. For the generated designs, it was shown that this combination of computational approaches can prospectively and effectively calculate strain designs for lactic acid overproduction. The long-term adaptive evolution experiments showed that: i) the computationally predicted phenotypes are experimentally reproducible and consistent; ii) the process of adaptive evolution leads to increased secretion rates of a target metabolite and can lead to improved product titers;

and iii) the generation of stable production strains can be achieved through this method. Overall, all evolved strains exhibited secretion profiles that supported the OptKnock hypotheses, in which the metabolite overproduction was stoichiometrically coupled to biomass generation.

6. Further Levels of Annotation

The majority of this chapter focused on the second dimension of genome annotation that defines the network links between the components given by the 1D annotation. In this section, we will briefly look at the remaining two dimensions, i.e., space and time. Although no reconstruction exists to date that considers these two additional dimensions, further research will provide the basis, and thus enable such reconstructions.

6.1. 3D Annotation: Spatial Position and Orientation

In the previous sections, we saw that the 1D annotation delivers a list of genes and their functions, which can be translated into a table of gene products and their known interactions (2D annotation). These interaction networks must operate within the three dimensional structure of a cell. A growing number of studies indicate that both the genomic location (i.e., the linear allelic address), as well as the spatial location (i.e., the position of a gene within the cell) of a gene is important in genome function (13). In addition, the growth phase of a cell influences the geometrical, and therefore topological, organization of a genome. An explicit link between the geometrical organization of the genome and the expression level of individual genes has yet to be established. However, log phase growth clearly requires many genes to be expressed contemporaneously, which cannot be achieved with a condensed chromosome.

6.1.1. 4D Annotation: Evolutionary Changes

Genomes can undergo short-term adaptive changes; thus, one can think of a fourth dimension to the genome—time. Such changes can be caused epigenetically or genetically, leading to modification in genome function over time. Mechanisms and how they function during adaptation have been studied for individual loci (such as arcB [14], mglD [15], mglO [15], and glpR [16] in *E. coli*), but have not yet been elucidated on a genome scale, with the exception of genome rearrangements. It is becoming appreciated that the genome sequence we have are "snap-shots" of a genome that is continually evolving. Thus, a more detailed understanding of the plasticity and adaptation of genomes on a genome scale is needed. The genetic basis for adaptation of genomes may emerge from full genome resequencing, enabling us to fully determine all the sequence changes that occur in genomes. Furthermore, resequencing may have the potential to provide insights into the mechanisms and functions of these adaptive evolutionary changes of an entire genome.

7. Future Directions

The four dimensions of genome annotation are important for describing and capturing the functional capabilities of a cell. A detailed, quality-controlled, and quality-assessed process for genome-scale reconstruction of metabolic networks (as an example of a 2D annotation) has developed over the past 5–10 years (17,18). It is a laborious and detailed process that involves the manual curation of a wide range of data types. Somewhat similar to sequence assembly and 1D genome annotation, this process of 2D annotation is iterative, involving the successive addition of more and more detailed data as they become available for a particular organism. These high-quality reconstructions can be used as the basis for computation of phenotypic traits, and they represent a key step in the development of the burgeoning field of systems biology (6). The number of organisms with publicly available genome-scale reconstructions continues to grow (Table 1).

Although the focus of this chapter was on metabolic networks, other networks, such as protein interaction, signaling, and regulatory networks, can be reconstructed in a similar manner. The nature of these networks is often qualitative in nature; the description of its components and their interactions may lack the biochemical details of metabolic reconstructions. However, these networks abide by the same chemical laws governing metabolic networks, such as conservation of mass and energy. Thus, many of the reconstruction details presented in this chapter are transferable to these networks if the details, such as stoichiometry, are known.

References

1. Janssen P, Goldovsky L, Kunin V, et al. Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications. *EMBO Rep* 2005;6(5):397–399.
2. Dansen TB, Wirtz KW, Wanders RJ, Pap EH. Peroxisomes in human fibroblasts have a basic pH. *Nat Cell Biol* 2000;2(1):51–53.
3. Nicolay K, Veenhuis M, Douma AC, et al. A 31P NMR study of the internal pH of yeast peroxisomes. *Arch Microbiol* 1987;147(1):37–41.
4. Karp PD, Riley M, Saier M, et al. The EcoCyc and MetaCyc databases. *Nucleic Acids Res* 2000;28(1):56–59.
5. (NC-IUBMB) NCotIUoBaMB. Enzyme Nomenclature. 6th ed. San Diego: Academic Press; 1992.
6. Palsson BO. Systems Biology: Properties of Reconstructed Networks. New York: Cambridge University Press; 2006.
7. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2(11):886–897.
8. Reed JL, Palsson BO. Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res* 2004;14(9):1797–1805.
9. Famili I, Mahadevan R, Palsson BO. k-Cone analysis: determining all candidate values for kinetic parameters on a network scale. *Biophys J* 2005; 88(3):1616–1625.

10. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004;429(6987):92–96.
11. Reed JL, Vo TD, Schilling CH, et al. An expanded genome-scale model of *Escherichia coli* K-12 (*i*JR904 GSM/GPR). *Genome Biol* 2003;4(9):R54.1–R.12.
12. Fong SS, Burgard AP, Herring CD, et al. In silico design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 2005;91(5):643–648.
13. Thanbichler M, Viollier PH, Shapiro L. The structure and function of the bacterial chromosome. *Curr Opin Genet Dev* 2005;15(2):153–162.
14. Flores N, Flores S, Escalante A, et al. 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* 2005;7(2):70–87.
15. Notley-McRobb L, Ferenci T. Adaptive mgl-regulatory mutations and genetic diversity evolving in glucose-limited *Escherichia coli* populations. *Environ Microbiol* 1999;1(1):33–43.
16. Raghunathan A, Palsson B. Scalable method to determine mutations that occur during adaptive evolution of *Escherichia coli*. *Biotechnol Lett* 2003; 25:435–441.
17. Reed JL, Palsson BO. Thirteen years of building constraint-based in silico models of *Escherichia coli*. *J Bacteriol* 2003;185(9):2692–2699.
18. Reed JL, Famili I, Thiele I, Palsson BO. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;7(2):130–141.
19. Park SM, Schilling CH, Palsson BO. Compositions and methods for modeling *Bacillus subtilis* metabolism. USA. 2003. Available at http://www.freepatentsonline.com/20030224363.html.
20. Reed JL, Vo TD, Schilling CH, et al. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* 2003;4(9): R54.
21. Edwards JS, Palsson BO. The *Escherichia coli* MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA* 2000;97(10):5528–5533.
22. Mahadevan R, Bond DR, Butler JE, et al. Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. *Appl Environ Microbiol* 2006;72(2):1558–1568.
23. Schilling CH, Palsson BO. Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J Theor Biol* 2000;203(3):249–283.
24. Edwards JS, Palsson BO. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. J Biol Chem 1999;274(25):17410–17416.
25. Thiele I, Vo TD, Price ND, et al. An expanded metabolic reconstruction of *Helicobacter pylori* (*i*IT341 GSM/GPR): An in silico genome-scale characterization of single and double deletion mutants. *J Bacteriol* 2005; 187:5818–5830.
26. Schilling CH, Covert MW, Famili I, Church GM, Edwards JS, Palsson BO. Genome-scale metabolic model of *Helicobacter pylori* 26695. *J Bacteriol* 2002;184(16):4582–4593.
27. Oliveira AP, Nielsen J, Forster J. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol* 2005;5(1):39.
28. Hong SH, Kim JS, Lee SY, et al. The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat Biotechnol* 2004; 22(10):1275–1281.

29. Becker SA, Palsson BO. Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol* 2005;5(1):8.
30. Borodina I, Krabben P, Nielsen J. Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. Genome Res 2005;15(6):820–829.
31. Feist AM, Scholten JCM, Palsson BO, et al. Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. 2006; 2(1):msb4100046-E1-msb-E14.
32. Sheikh K, Forster J, Nielsen LK. Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. *Biotechnol Prog* 2005;21(1):112–121.
33. Duarte NC, Herrgard MJ, Palsson BO. Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res* 2004;14(7):1298–1309.
34. Forster J, Famili I, Fu P, et al. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* 2003;13(2):244–253.
35. Majewski RA, Domach MM. Simple constrained optimization view of acetate overflow in *E. coli*. *Biotechnol Bioeng* 1990;35:732–738.
36. Lee S, Phalakornkule C, Domach MM, et al. Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comp Chem Eng* 2000;24:711–716.
37. Vo TD, Greenberg HJ, Palsson BO. Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J Biol Chem* 2004;279(38):39532–39540.
38. Burgard AP, Pharkya P, Maranas CD. OptKnock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 2003;84(6):647–657.
39. Thiele I, Price ND, Vo TD, Palsson BO. Candidate metabolic network states in human mitochondria: Impact of diabetes, ischemia, and diet. *J Biol Chem* 2005;280(12):11683–11695.
40. Price ND, Thiele I, Palsson BO. Candidate states of *Helicobacter pylori's* genome-scale metabolic network upon application of "loop law" thermodynamic constraints. *Biophys J* 2006;90(11):3919–3928.
41. Price ND, Schellenberger J, Palsson BO. Uniform sampling of steady state flux spaces: means to design experiments and to interpret enzymopathies. *Biophysl J* 2004;87(4):2172–2186.
42. Wiback SJ, Famili I, Greenberg HJ, et al. Monte Carlo sampling can be used to determine the size and shape of the steady state flux space. *J Theor Biol* 2004;228(4):437–447.

http://www.springer.com/978-1-58829-706-8

Introduction to Systems Biology
Choi, S. (Ed.)
2007, XVI, 542 p. 163 illus., 2 illus. in color., Hardcover
ISBN: 978-1-58829-706-8
A product of Humana Press