6

# *IN SILICO* GENOME-SCALE METABOLIC MODELS: THE CONSTRAINT-BASED APPROACH AND ITS APPLICATIONS

Andrew R. Joyce[1] and Bernhard Ø. Palsson[2]

*[1]Bioinformatics Program, University of California, San Diego,*
*9500 Gilman Drive, La Jolla, California 92093*
*[2]Department of Bioengineering, University of California, San Diego,*
*9500 Gilman Drive, La Jolla, California 92093*

## 6.1 INTRODUCTION TO MODELING USING THE CONSTRAINT-BASED APPROACH

The development of high-throughput experimental techniques in recent years has led to an explosion of genome-scale data sets for a variety of organisms. Considerable efforts have yielded complete genomic sequences for dozens of organisms [1] from which gene annotation provides a list of individual cellular components. Microarray technology affords researchers the ability to probe gene expression patterns of cells and tissues on a genome scale. Genome-wide location analysis, also known as ChIP-chip [2], provides transcription factor binding site information for the entire cell. Furthermore, advances in the fields of fluxomics [3] and proteomics add to the vast quantity of data currently available to researchers. Integration of these data sets to extract the most relevant information to formulate a comprehensive view of

*Systems Biology and Synthetic Biology* Edited by Pengcheng Fu and Sven Panke

biological systems is a major challenge the biological research community [4] currently facing. Achieving this task will require comprehensive models of cellular processes.

A prudent approach to gain biological understanding from these complex data sets involves the development of mathematical modeling, simulation, and analysis techniques [5]. For many years, researchers have developed and analyzed models of biological systems via simulation, but these efforts often have been hampered by lack of complete or reliable data. Some examples of the modeling philosophies and approaches that have been pursued include deterministic kinetic modeling [6,7], stochastic modeling [8,9], and Boolean modeling [10]. Many of these approaches are hindered by requiring knowledge of unknown parameters that are difficult to determine experimentally. Furthermore, the above approaches typically require substantial computational power, thus, limiting the scale of the models that can be developed.

In recent years, however, great strides have been made in developing and using genome-scale metabolic models of a number of organisms using another modeling technique that is not subject to the above limitations. This approach, known as constraint-based modeling [11–15], has been employed to generate genome scale for organisms from all three major branches of the tree of life. While bacterial models dominate this growing collection, a model from archaea has recently appeared, and several eukaryotic models are also available (see Table 6-1 for an overview of existing constraint-based models).

In complimentary efforts, many analytical tools have been developed to use these models in computational investigations of model organisms (reviewed in Ref. [12]). One method in particular, known as flux balance analysis (FBA) [16,17], is a powerful mathematical approach that uses optimization by linear programming (LP) to study the properties of metabolic networks under various conditions. When using FBA, the investigator chooses a property to optimize, such as biomass production in microbial models, and then calculates the optimal flux distribution(s) that lead to this result. Therefore, FBA is useful for computationally assessing the ability of an organism to grow on a particular substrate or in a particular environment and can also be used to assess the effect of metabolic gene deletions under various growth conditions. Given that these types of analyses rely on computer simulation, computational results must be confirmed at the bench through experimental means. However, by first investigating these situations at the computer work station, researchers can be directed to the most interesting and scientifically meaningful experiments to perform, thus limiting the amount of time spent conducting experiments of less scientific value.

In this chapter, we provide an introduction to the principles that underlie constraint-based modeling and FBA of biological systems. We give a brief, but practical example to introduce the method and concepts directly. Furthermore, we discuss both the utility and potential shortcomings of these models by reviewing several published studies that use these models to assess gene essentiality, which is simply defined as the study of organism viability despite harboring single or multiple gene knockouts. Finally, we briefly discuss additional analytical techniques and interesting applications of constraint-based modeling as well as their future implications.

**Table 6-1 Currently available constraint-based models**

| Organism | Total Genes | Model Genes | Model Metabolites | Model Reactions | Reference |
|---|---|---|---|---|---|
| Bacteria | | | | | |
| *Bacillus subtilis* | 4225 | 614 | 637 | 754 | [114] |
| *E. coli* | 4405 | 904 | 625 | 931 | [61] |
| | | 720 | 438 | 627 | [63] |
| *Geobacter sulfurreducens* | 3530 | 588 | 541 | 523 | [67] |
| *Haemophilus influenzae* | 1775 | 296 | 343 | 488 | [79] |
| | | 400 | 451 | 461 | [82] |
| *Heliobacter pylori* | 1632 | 341 | 485 | 476 | [65] |
| | | 291 | 340 | 388 | [64] |
| *Lactococcus lactis* | 2310 | 358 | 422 | 621 | [115] |
| *Mannheimia succinciproducens* | 2463 | 335 | 352 | 373 | [116] |
| *Staphylococcus aureus* | 2702 | 619 | 571 | 641 | [66] |
| *Streptomyces coelicolor* | 8042 | 700 | 500 | 700 | [68] |
| Archaea | | | | | |
| *Methanosarcina barkeri* | 5072 | 692 | 558 | 619 | [69] |
| Eukarya | | | | | |
| *Mus musculus* | 28,287 | 1156 | 872 | 1220 | [75] |
| *S. cerevisiae* | 6183 | 750 | 646 | 1149 | [71] |
| | | 672 | 636 | 1038 | [72] |
| | | 708 | 584 | 1175 | [70] |
| Human cardiac mitochondria | 615[a] | 298 | 230 | 189 | [56] |
| Human red blood cell | NA | NA | 39 | 32 | [76] |

This table summarizes model statistics for the models developed and published to date. *E. coli*, *Escherichia coli*; *S. cerevisiae, Saccharomyces cerevisiae*; NA, not applicable.

[a] This number is based on the protein species identified in a proteomics study of the human cardiac mitochondria from which the components of the reconstruction were derived [117].

## 6.2 BUILDING A CONSTRAINT-BASED MODEL

This section outlines the general procedure (Fig 6-1) followed in constructing a constraint-based model with a slant toward metabolic network. Furthermore, we introduce FBA as an example of a useful analytical method that can be used in conjunction with these models. This model building and analysis approach can be divided into approximately four successive steps:

1. Network reconstruction
2. Stoichiometric ($S$) matrix compilation
3. Identification and assignment of appropriate constraints to molecular components

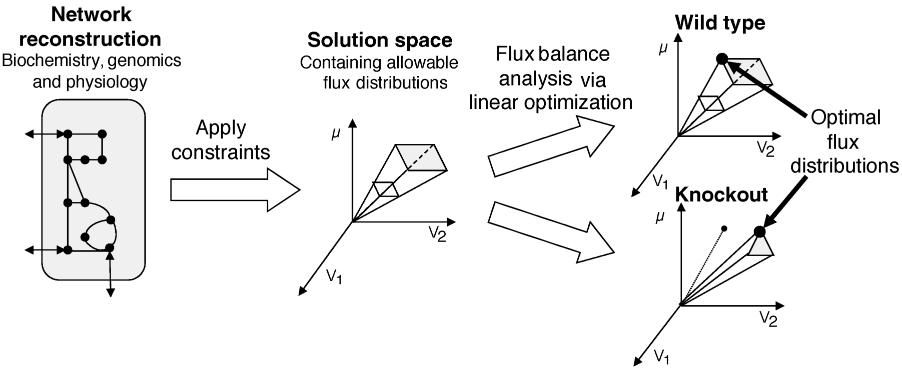

Network reconstruction
Biochemistry, genomics and physiology
Apply constraints
Solution space
Containing allowable flux distributions
μ
V2
V1
Flux balance analysis via linear optimization
Wild type
Optimal flux distributions
Knockout

**Figure 6-1** Constraint-based modeling. Application of constraints to a reconstructed metabolic network leads to a defined solution space that specifies a cell's allowable metabolic phenotypes. Flux balance analysis uses linear optimization to find solutions in the space that maximize or minimize a given objective. The effects of gene knockouts on the solution space and metabolic capabilities can be assessed by simulating a gene knockout and comparing its ability to grow *in silico* relative to wild type.

4. Optimal flux distribution determination and assessment of gene essentiality via flux balance analysis

We will consider each of the above components in turn. In addition, a simple example will be provided in Section 6.2.5 to illustrate directly the concepts described herein.

### 6.2.1 Network Reconstruction

The first step in constraint-based modeling, known as network reconstruction, involves generating a model that describes the system of interest. This process can be decomposed into three parts typically performed simultaneously during model construction. These components, known as data collection, metabolic reaction list generation, and gene–protein reaction relationship (GPR) determination, are detailed in this section.

***6.2.1.1 Data Collection*** Perhaps the most critical component of the constraint-based modeling approach involves data collection relevant to the system of interest. Not long ago, this was among the most challenging steps as researchers had very limited access to amounts of biochemical data. However, the success of recent genome sequencing and annotation projects, advances in high-throughput technologies, and the extensive development of online database resources have improved matters dramatically.

After identifying the system or organism of interest, relevant data sources must be identified to begin the compiling of appropriate metabolites, biochemical reactions, and associated genes to be included in the model. The three primary types of resources are the biochemical literature, high-throughput data, and integrative database resources.

*Biochemical Literature* Direct biochemical information found in the primary literature usually contains the highest quality data for use in reconstructing biochemical networks. Important details, such as precise reaction stoichiometry and reaction reversibility, are often directly available. Given that scrutinizing each study individually is time consuming and tedious task, biochemical textbooks and review articles should be utilized when available, relying on the primary literature used to resolve conflicts. Furthermore, many volumes devoted to individual organisms and organelles, such as *Escherichia coli* [18] and the mitochondria [19], are increasingly available and are typically excellent resources.

*High-Throughput Data* Genomic and proteomic data are useful sources of information for identifying relevant metabolic network components. In recent years, the complete genome sequence of hundreds of organisms has been determined and many more sequencing projects are underway [20]. This collection is dominated by microbial and viral sequences, but several highly publicized higher eukaryotic sequences are also available [21–24]. Furthermore, extensive bioinformatics-based annotation efforts continue to make great strides toward automatically identifying all coding regions contained within the sequence [25–27]. To illustrate a common approach to gene functional annotation, consider the case in which a biochemical reaction is known to occur in the organism, but whose corresponding gene(s) are unknown. Sequence alignment tools such as BLAST and FASTA [28] can be utilized to assign putative functions based on similarity to orthologous genes and proteins of known function in other sequenced organisms. However, it should be noted that putative assignments represent functional hypotheses and are subject to revision upon direct biochemical characterization. As one final note on genome annotation, interesting efforts are also underway to automatically reconstruct networks based on annotated sequence information alone [29]. However, these automated approaches are limited in that they can be only as good as the genome annotation from which they are derived. Therefore, considerable quality control efforts should be conducted prior to extensive use of these networks.

The proteome of a biological system defines the full complement, localization, and abundance of proteins. Although these data are generally difficult to obtain, data for some subcellular components and bacteria are available [30,31]. Proteomic data are of particular importance in eukaryotic systems modeling in which care must be taken to assign reactions to their appropriate subcellular compartment or organelle. Similarly, when modeling a system under a single condition, these data are important in identifying active components.

In addition to the primary literature, genomic and proteomic data repositories can be accessed via the Internet as can the additional resources discussed in the next section. Some popular resources are provided in Table 6-2.

*Integrative Database Resources* In recent years, significant efforts have been devoted to developing comprehensive databases that integrate many information sources including those data types previously described. Of particular interest are resources that have incorporated these disparate data sources into metabolic

Table 6-2 Online data resources

| Data Type | Resource | Description | URL |
|---|---|---|---|
| Genomic | Genomes OnLine Database (GOLD) | Repository of completed and ongoing genome projects | http://www.genomesonline.org |
| | The Institute for Genomic Research (TIGR) | Curated databases for microbial, plant, and human genome projects | http://www.tigr.org |
| | National Center for Biotechnology Information (NCBI) | Curated databases of DNA sequences as well as other data | http://www.ncbi.nlm.nih.gov |
| Transcriptomic | Gene Expression Omnibus (GEO) | Microarray and SAGE-based genome-wide expression profiles | http://www.ncbi.nlm.nih.gov/geo |
| | Stanford Microarray Database (SMD) | Microarray-based genome-wide expression data | http://genome-www5.stanford.edu/ |
| Proteomic | Expert Protein Analysis System (ExPASy) | Protein sequence, structure, and 2D-PAGE data. | http://au.expasy.org |
| | BRENDA | Enzyme functional data. | http://www.brenda.uni-koeln.de/ |
| | Open Proteomics Database (OPD) | Mass-spectrometry-based proteomics data | http://bioinformatics.icmb.utexas.edu/OPD |
| Protein–DNA Interaction | Biomolecular Network Database (BIND) | Published protein–DNA interactions | http://www.bind.ca/Action/ |
| | Encyclopedia of DNA Elements (ENCODE) | Database of functional elements in human DNA | http://genome.ucsc.edu/encode/ |
| Protein–protein interaction | Munich Information Center for Protein Sequences (MIPS) | Links to protein–protein interaction data and resources | http://mips.gsf.de/proj/ppi |
| | Database of Interacting Proteins (DIP) | Published protein–protein interactions | http://dip.doe-mbi.ucla.edu |

| | | | |
|---|---|---|---|
| Subcellular location | Yeast GFP-fusion Localization Database | Genome-scale protein localization data for yeast | http://yeastgfp.ucsf.edu |
| Phenotype | A Systematic Annotation Package (ASAP) for community analysis of genomes | Single-gene-deletion phenotype microarray data for *E. coli* | http://www.genome.wisc.edu/tools/asap.htm |
| | General Repository for Interaction Datasets (GRID) | Synthetic lethal interactions in yeast | http://biodata.mshri.on.ca/grid |
| Pathway | Kyoto Encyclopedia of Genes and Genomes (KEGG) | Pathway maps for many biological processes | http://www.genome.ad.jp/kegg/ |
| | BioCarta | Interactive graphic models of molecular and cellular pathways | http://www.biocarta.com/genes/index.asp |
| Organism specific | EcoCyc | Encyclopedia of *E. coli* K12 genes and metabolism | http://www.ecocyc.org |
| | *Saccharomyces* Genome Database (SGD) | Scientific database of the molecular biology and genetics of *S. cerevisiae* | http://www.yeastgenome.org |
| | BioCyc | A collection of 205 pathway/genome databases for individual organisms. | http://www.biocyc.org |

This table details some of the databases that store and distribute genome-scale data, gene ontological information, and organism specific data. It should also be noted that this table is by no means comprehensive in its content, but rather provides a reasonably broad sample of the data and resources that are readily accessible to researchers today. 2D-PAGE, two-dimensional polyacrylamide-gel electrophoresis; GFP, green fluorescent protein; SAGE, serial analysis of gene expression.

pathway maps. Among these resource types, Kyoto Encyclopedia of Genes and Genomes (KEGG) [32] is perhaps the most extensive and well known. Pathway maps for numerous metabolic processes are available. KEGG also provides information regarding orthologous genes for a variety of organisms, thus greatly enhancing the power of this resource. Additional organism-specific database resources are also available. EcoCyc [33] incorporates gene and regulatory information as well as enzyme-reaction pathways particular to *E. coli*. The Comprehensive Yeast Genome Database (CYGD) [34] and *Saccharomyces* Genome Database (SGD) [35] are other examples of *Saccharomyces cerevisiae*-specific comprehensive resources. Finally, the BioCyc resource [36,37] contains automated annotation-derived pathway/genome databases for 205 individual organisms.

An additional important wealth of information can be found in resources that provide functional information for individual genes and gene products. These ontology-based tools strive to describe how gene products behave in a cellular context. The most well-known resource is Gene Ontology Consortium (GO) [38,39] that contains information for a variety of organisms. In recent years, organism-specific ontologies, such as GenProtEC [40] for *E. coli*, also have appeared. In sum, these online resources are valuable that they typically integrate information regarding individual genes and proteins as well as information regarding their regulation and participation in enzymatic reactions in a single location.

***6.2.1.2 Metabolic Reaction List Generation*** The next step in defining a constraint-based model requires clearly specifying the reactions to be included based on the metabolite and enzyme information collected in the previous step. A metabolic reaction can be viewed simply as substrate(s) conversion to product(s), often by enzyme-mediated catalysis. In light of this notion, each reaction in a metabolic network must adhere to the fundamental laws of physics and chemistry; therefore, reactions must be balanced in terms of charge and elemental composition. For example, the depiction of the first step of glycolysis in Figure 6-2a is neither elementally nor charge balanced. However, inclusion of hydrogen in Figure 6-2b balances the reaction in both regards.

Biological boundaries also must be considered when defining reaction lists. Metabolic networks are comprised of both intracellular and extracellular reactions. For example, the reactions of glycolysis and the tricarboxylic acid (TCA) cycle take

(*a*)

$$C_6H_{12}O_6 + ATP^{3-} \xrightarrow{\text{hexokinase}} C_6H_{11}O_6PO_3^{2-} + ADP^{2-}$$

(*b*)

$$C_6H_{12}O_6 + ATP^{3-} \xrightarrow{\text{hexokinase}} C_6H_{11}O_6PO_3^{2-} + ADP^{2-} + H^{+}$$

**Figure 6-2** Charge and elementally balanced reactions. (a) This depiction of the hexokinase-mediated conversion of glucose to glucose-6-phosphate is neither elementally, nor charge balanced. (b) Inclusion of hydrogen both elementally and charge balances the reaction.

place intracellularly in the cytosol. However, glucose must be transported into the cell via an extracellular reaction in which a glucose transporter takes up extracellular glucose. An additional boundary consideration must be recognized particularly when modeling eukaryotic cells. Given that certain metabolic reactions take place in the cytosol and others take place in various organelles, reactions must be compartmentalized properly. Data is now being generated in which proteins are tagged, for example, with green fluorescent protein (GFP) or recognized by antibodies and localized to subcellular compartments or organelles [41–43]. Furthermore, computational tools have also been developed to predict subcellular location of proteins in eukaryotes [44].

Finally, reaction reversibility must be defined. Certain metabolic reactions can proceed in both directions. Thermodynamically, this permits reaction fluxes to take on both positive and negative values. The KEGG and BRENDA online resources (Table 6-2) are two useful resources that catalog enzyme reversibility.

***6.2.1.3 Determining Gene–Protein Reaction Relationships*** Upon completing the reaction list, the protein or protein complexes that facilitate each metabolite substrate to product conversion must be determined. Each subunit protein from a complex must be assigned to the same reaction. Additionally, some reactions can be catalyzed by different enzymes. Collectively, each enzyme that fits this criterion is known as an isozyme for a particular reaction. Accordingly, isozymes must all be assigned to the same appropriate reaction. Biochemical textbooks often provide the general name of the enzyme(s) responsible; however, the precise gene and associated gene product specific for the model organism of interest must be identified. The database resources detailed in Section 6.2.1.1 and Table 6-2 assist this process. In particular, KEGG and GO provide considerable enzyme-reaction information for a variety of organisms. Furthermore, protein–protein interaction data sets, for example, those derived from yeast two-hybrid experiments [45], may be useful resource for defining enzymatic complexes in less defined situations. One must take care in using these data because of their high false-positive rate and questionable reproducibility [46,47].

### 6.2.2 Defining the Stoichiometric Matrix

The compiled reaction list can be represented mathematically in the form of a $S$ matrix. The $S$ matrix is formed from the stoichiometric coefficients of the reactions that participate in the defined reaction network. It has $m \times n$ dimensions, where $m$ is the number of metabolites and $n$ is the number of reactions. Therefore, the $S$ matrix is organized such that every column corresponds to a reaction and every row corresponds to a metabolite. The $S$ matrix describes how many reactions a compound participates in, and thus, how reactions are interconnected. Accordingly, each network that is reconstructed in this way effectively represents a two-dimensional annotation of the genome [11,48].

Figure 6-3 shows how a simple two-reaction system can be represented as an $S$ matrix. In this example, $v_1$ and $v_2$ denote reaction fluxes and are associated with

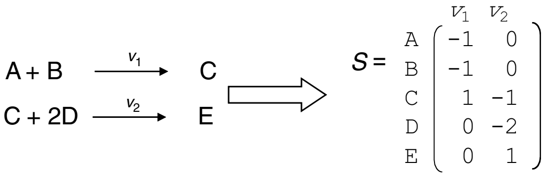


**Figure 6-3** Generating the *S* matrix. The reaction list on the left is mathematically represented by the *S* matrix on the right. As a convention, each row represents a metabolite and each column represents a reaction in the network. Additionally, input or reactant metabolites have negative stoichiometric coefficients and outputs or products have positive stoichiometric coefficients. Metabolites that do not participate in a given reaction are assigned a zero value.

individual proteins or protein complexes that catalyze the reactions. In the *S* matrix representation, each row denotes an individual metabolite while each column corresponds to an individual reaction. Element $S_{ij}$ represents the stoichiometric coefficient of the metabolite associated with row $i$ in the reaction associated with column $j$. Furthermore, notice that substrates are assigned negative coefficients and products are given positive coefficients. Also, for those reactions in which a metabolite does not participate, the corresponding *S* matrix element is assigned a zero value.

### 6.2.3 Identifying and Applying Constraints

Having developed a mathematical representation of a metabolic network in the form of the *S* matrix, the next step requires that any constraints be identified and imposed on the model. Cells are subject to a variety of constraints from environmental, physiochemical, evolutionary, and regulatory sources [12,14]. In and of itself, the *S* matrix is a constraint in that it defines the mass and charge balance requirements for all possible metabolic reactions available to the cell. These stoichiometric constraints establish a geometric solution space that, in principle, contains all possible metabolic behaviors.

Additional constraints can be identified and imposed on the model, which has the effect of further limiting the metabolic behavior solution space. Maximum enzyme capacity, $V_{\mathrm{max}}$, which can be determined experimentally for some reactions is one example and can be imposed by limiting the flux through any associated reactions to that maximum value. Furthermore, the uptake rates of certain metabolites can be determined experimentally and used to restrict metabolite uptake to the appropriate levels when mathematically analyzing the metabolic model. Additional types of constraints have also been applied including thermodynamic limitations [49], internal metabolic flux determinations [13], and transcriptional regulation [50–53]. This latter topic will receive considerable detailed treatment in Section 6.4.3.

With respect to computationally assessing gene essentiality, a similar strategy to setting the maximum enzyme capacity can be utilized. By simply restricting the flux through reactions associated with the protein of interest to zero, a gene knockout can be simulated. Flux balance analysis then can be used to examine the simulated knockout properties relative to wild type, as outlined in the next section.

### 6.2.4 Assessing the Model Using Flux Balance Analysis

Flux balance analysis is a powerful computational method that relies on optimization techniques by linear programming [54] to investigate the production capabilities and systemic properties of a metabolic network. By defining an objective, such as biomass production, ATP production, or by-product secretion, linear optimization may be used to find an optimal flux distribution for the network model that maximizes the stated objective. This section briefly introduces some main concepts that underlie FBA, with an emphasis on how FBA can be utilized to assess gene essentiality in a metabolic network.

***6.2.4.1 Linear Optimization*** As stated previously, the solution space defined by constraint-based models can be explored via optimization by linear programming. The LP problem corresponding to the search for the optimal flux distribution determination through a metabolic network can be formulated as follows:

$$\begin{array}{ll} \text{Maximize} & Z = c^{\mathbf{T}} v \\ \text{Subject to} & S \cdot v = 0 \\ \alpha_i \leq v_i \leq \beta_i & \text{for all reactions } i \end{array}$$

In the above representation, $Z$ represents the objective function, and $c$ is a vector of weights on the fluxes $v$. The weights are used to define the properties of the particular solution that is sought. The latter statements represent the flux constraints for the metabolic network. $S$ is the matrix defined in the previous section and contains the mass and charge balanced representation of the system. Furthermore, each reaction flux $v_i$ in the system is subject to lower and upper bound constraints, represented in $\alpha_i$ and $\beta_i$, respectively.

The solution to this problem yields not only a value for $Z$ but also results in an optimal flux distribution ($v$) that allows the highest flux through the chosen objective function, $Z$. Furthermore, computational assessment of gene essentiality is performed easily within this framework. By setting the upper and lower flux bound constraints to zero for the reaction(s) corresponding to the gene(s) of interest, a simulated gene deletion strain may be created. Examining the results of simulations run before and after knocking out a gene lead to gene essentiality predictions.

Problems of this type can be formulated and solved readily by commercial software packages,such as Matlab (The MathWorks, Inc., Natick, MA), Mathematica (Wolfram Research, Inc., Champaign, IL), LINDO (LINDO Systems, Inc., Chicago, IL), and tools available through the General Algebraic Modeling System (GAMS Development Corporation, Washington, DC). Section 6.2.5 presents a simple, hypothetical example solved using Matlab. It should also be noted that these types of analyses yield a single answer; however, it is possible that multiple equivalent flux distributions that yield a maximal biomass function value for a given network and simulation conditions. This topic has been explored using mixed integer linear programming (MILP) techniques with genome-scale metabolic models [55,56], but is beyond the scope of this chapter and will not be discussed further.

***6.2.4.2 Constraints*** As previously stated, the *S* matrix constrains the system by defining the mass and charge balance constraints for all possible metabolic reactions within the system. In mathematical terms, the *S* matrix is a linear transformation of the reaction flux vector,

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

to a vector of time derivatives of metabolic concentrations

$$\mathbf{x} = (x_1, x_2, \ldots, x_n)$$

such that

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{S} \cdot v$$

Therefore, a particular flux distribution $v$ represents the flux levels through each reaction in the network. Since the time constants that describe metabolic transients are fast (on the order of tens of seconds or less), whereas the time constants for cell growth are comparatively long (on the order of hours to days) the behavior of cellular components can be considered as existing in a quasi steady state. This assumption leads to the reduction of the previous equation to

$$\mathbf{S} \cdot v = 0$$

By focusing only on the steady-state condition, assumptions regarding reaction kinetics are not needed. Furthermore, based on this premise, it is possible to determine all chemically balanced metabolic routes through the metabolic network.

The second constraint set is imposed on the individual reaction flux values. The constraints defined by

$$\alpha_i \leq v_i \leq \beta_i \quad \text{for all reactions } i$$

specify lower and upper flux bounds for each reaction. If all model reactions are irreversible, α equals to 0. Similarly, if the enzyme capacity, $V_{\max}$, is experimentally defined, setting β to the known experimental value limits the allowable reaction flux through the enzyme. In contrast, a gene knockout is simulated by setting $\beta_i = 0$ for gene *i* (see Section 6.2.5 and Box 6-1). If no constraints on flux values through reaction $v_i$ can be identified, then $\alpha_i$ and $\beta_i$ are set to -∞ and $+\infty$, respectively, to allow for all possible flux values. In practice, ∞ is typically represented as an arbitrarily large number that will exceed any feasible internal flux (for an example, see Section 6.2.5).

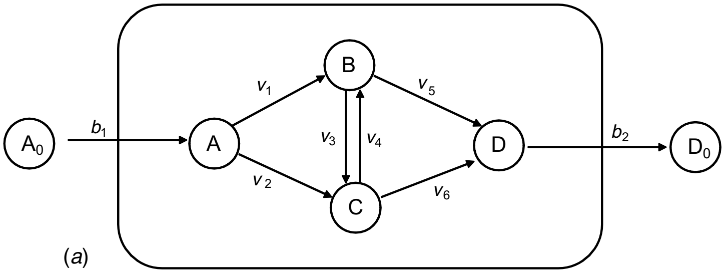
A brief consideration should also be given to specifying input and output constraints on the system. When analyzing metabolic models in the context of assessing cellular growth capabilities, input constraints effectively define the environmental conditions being considered. For example, organisms have various elemental requirements that must be provided in the environment in order to support growth.

Some organisms that lack certain biosynthetic processes are auxotrophic for certain biomolecules, such as amino acids, and these compounds also must be provided in the environment. From an FBA standpoint, these issues mean that input sources must be specified in the form of input flux constraints specified in $v$. For example, if one desires to simulate rich medium conditions, flux constraints are specified such that all biomolecules that can be served as inputs to the system, in other words all compounds that are available extracellularly, are left unconstrained and can flow freely into the system. In contrast, when modeling minimal medium conditions (for an example of a large-scale analysis performed of *E. coli* growth simulations on minimal media, see Ref. [57]) only those inputs required for cell growth, or biomass formation in the formalism being considered here, are allowed to flow into the system with all other input fluxes constrained to zero. It should also be noted that certain output flux constraints may need to be set appropriately in order to allow for the simulated secretion of biomolecules that may "accumulate" in the process of forming biomass. A simple example of this is allowing for lactate and acetate secretion when modeling fermentative growth of microbes.

***6.2.4.3 The Objective Function*** Given that multiple possible flux distributions exist for any given network, linear optimization is used to identify a particular solution that maximizes or minimizes a defined objective function. Commonly used objective functions include production of ATP or production of a secreted by-product. When assessing the growth capabilities of a microbe using its associated metabolic model, growth rate, as defined by the weighted consumption of metabolites needed to make biomass, is maximized. The general analysis strategy asks the question "is the metabolic reaction network able to support growth under the specified growth conditions?" Therefore, biomass generation in this modeling framework is represented as a reaction flux that drains intermediate metabolites, such as ATP, NADPH, pyruvate, and amino acids, in appropriate ratios (defined in the vector $c$ of the biomass function $Z$) to support growth. As a convention, the biomass function is typically written to reflect the needs of the cell in order to make 1 g of cellular dry weight, and has been experimentally determined for *E. coli* [58]. In sum, the choice of biomass as an objective function, cell growth, depicted as a nonzero value for $Z$, will only occur if all the components in the biomass function can be provided for by the network in the correct relative amounts.

### 6.2.5 A Simple FBA Example

In order to demonstrate the concepts previously introduced, this section presents a specific example using a simple system. Figure 6-4a shows a hypothetical four metabolite (A,B,C,D), eight reaction ($v_1$,$v_2$,$v_3$,$v_4$,$v_5$,$v_6$,$b_1$,$b_2$) network. By convention, each internal reaction is associated with a flux $v_i$ whereas reactions that span the system boundary are denoted with flux $b_i$. Furthermore, external metabolites A and D are denoted with subscript "o" to distinguish them from their corresponding internal metabolite. However, external metabolites need not be explicitly considered in the stoichiometric network representation.



$b_1$:     ? A
$v_1$: A ? B
$v_2$: A ? C
$v_3$: B ? C
$v_4$: C ? B
$v_5$: B ? D
$v_6$: C ? D
$b_2$: D ?

(*b*)

$$\mathbf{S} = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{cccccccc} v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & b_1 & b_2 \\ \end{array}$$

$$\begin{array}{c|cccccccc} & v_1 & v_2 & v_3 & v_4 & v_5 & v_6 & b_1 & b_2 \\ \hline A & -1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 \\ B & 1 & 0 & -1 & 1 & -1 & 0 & 0 & 0 \\ C & 0 & 1 & 1 & -1 & 0 & -1 & 0 & 0 \\ D & 0 & 0 & 0 & 0 & 1 & 1 & 0 & -1 \end{array}$$

(*c*)

**Figure 6-4** An example system. (a) A four-metabolite, eight-reaction system is first decomposed into individual reactions in (b), and then represented mathematically in the *S* matrix depicted in (c). By convention, internal reactions are denoted by $v_i$ and reactions that span the system boundary are denoted by $b_i$ External metabolites $A_0$ and $D_0$ need not be explicitly represented explicitly within this framework as they are outside the system under consideration.

Figure 6-4b outlines the reaction list associated with the system. Notice that the conversion of metabolite B to C is reversible. Rather than treating this as a single reaction, for simplicity this reaction is decoupled into two separate reactions with individual corresponding fluxes.

The *S* matrix for this system is detailed in Figure 6-4c. Again, notice how this representation follows directly from the reaction list. Metabolite substrates and products are represented with negative and positive coefficients, respectively. Recall that LP problems take on the following form:

$$\begin{array}{ll} \text{Maximize} & Z = c^{\mathbf{T}} v \\ \text{Subject to} & S \cdot v = 0 \\ \alpha \leq v_i \leq \beta & \text{for all reactions } i \end{array}$$

For example, if the metabolite D output is to be maximized, corresponding to maximizing the flux through $b_2$ the objective function is defined as follows:

$$Z = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1) \cdot (v_1 \quad v_2 \quad v_3 \quad v_4 \quad v_5 \quad v_6 \quad b_1 \quad b_2)^{\mathrm{T}}$$

Furthermore, in addition to the mass and charge, balance constraints imposed by the *S* matrix, lower ($\alpha$), and upper ($\beta$) bound vectors must be specified for the reaction vector *v*. Since all reactions in this network are irreversible, which constrains all fluxes to be positive, the lower bound vector $\alpha$ is set to zero.

$$\alpha = (0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0)^{\mathrm{T}}$$

Upper bound values specified in vector β can be chosen to incorporate experimentally determined maximal enzyme capacities, also known as $V_{\max}$ values, or some arbitrarily chosen values to explore network properties. An acceptable example vector is

$$\beta = (2 \quad 10 \quad 4 \quad 6 \quad 10 \quad 8 \quad 100 \quad 100)^{\mathrm{T}}$$

The latter two upper bound values for the respective input and output fluxes are set to an arbitrarily large number in this case to reflect an effectively unlimited capacity. Given the constraints on the internal fluxes, however, the actual values of these fluxes in the calculated optimal flux distribution will never approach these values.

Utilizing the information compiled above, the Matlab function **linprog()** can be used to solve for a steady-state flux distribution that maximizes for the output of metabolite D under wild-type conditions, as detailed in Box 6-1. It should be noted that the default Matlab optimization solver is only suitable for problems of this and slightly

## BOX 6-1 FBA USING MATLAB

Here, we use Matlab to solve an FBA problem for three cases using the system shown in Fig. 6-4. The **linprog()** function accepts six arguments and returns two values in the following form:

$$[v, Z] = \text{linprog}(c, \text{Aeq}, \text{beq}, S, b, \alpha, \beta)$$

This solves the following LP problem:

$$\begin{aligned}&\text{Minimize} \quad Z = c^T \times v\\&\text{Subject to} \quad \text{Aeq} \times v \leq \text{beq}\\&S \cdot v = b\\&a \leq v \leq b\end{aligned}$$

Since the system does not have inequality constraints other than flux vector bounds, Aeq is set equal to the identity matrix and beq to β, so that

$$\text{Aeq} \cdot v \leq \text{beq}$$

is equivalent to

$$v \leq b$$

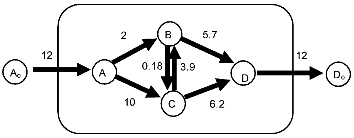
The code to solve the wild type problem (Case 6-1) of interest in Matlab's framework follows, using the **linprog()** function and α and β as defined in the text:

```
>> S = [-1 (1 0 0 0 0 1 0; 1 0 -1 1 -1 0 0 0; 0 1 1 -1 0 -1 0 0; 0 0 0 0 1 1 0 -1];
>> b = [0 0 0 0]';
>> alpha = [0 0 0 0 0 0 0 0 ]';
>> beta = [2 10 4 6 10 8 100 100]';
>> c = [0 0 0 0 0 0 0 1];
>> Aeq = eye (8);
>> [v, z] = linprog (-c, Aeq, beta, S, b, alpha, beta)Optimization
terminated successfully.

v = 2.0000 10.0000 0.1822 3.9137 5.7315 6.2685 12.0000 12.0000
Z = -12.0000
```

Note that since **linprog()** defaults to solving a minimization problem we use the negative of the optimization weight vector *c*. Use the Matlab Help for more details on **linprog()**.

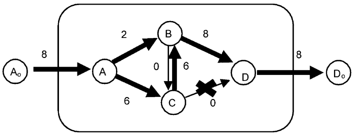
*Case 6-1*: Wild Type



Case 6-2, shown below, solves the same problem but this time after knocking out reaction $v_5$ by modifying the $\beta$ vector stored in the beta variable:

```
>> beta = [2 10 4 6 10 0 100 100]';
```

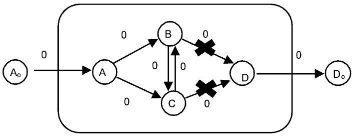
*Case 6-2*: Growth Impaired $v_6$ Knockout



Finally, in Case 6-3 depicted below, by again modifying the beta variable a “lethal” deletion strain can be simulated by knocking out both $v_5$ and $v_6$:

```
>> beta = [2 10 4 6 0 0 100 100]';
```

*Case 6-3*: Lethal $v_5$ and $v_6$ Double Knockout



larger magnitude. Typical biological problems that involve many more variables and constraints require more sophisticated optimization software such as the packages available through LINDO Systems, Inc. and GAMS. A thorough discussion of the algorithmic details that underlie solving FBA and other LP problems is beyond the scope of this text. For further details, see Refs [54,59,60] .

Having used the above information to simulate the wild-type case, the upper bound $\beta$ vector can be modified to simulate a gene deletion. For example, if we want to examine the effects of deleting the enzyme responsible for the conversion of metabolite C to D, flux $v_6$ is restricted to zero simulation.

$$\beta = (2 \quad 10 \quad 4 \quad 6 \quad 10 \quad \mathbf{0} \quad 100 \quad 100)^T$$

Similarly, a $v_5$, $v_6$ double mutant is simulated using the following vector:

$$\beta = (2 \quad 10 \quad 4 \quad 6 \quad \mathbf{0} \quad \mathbf{0} \quad 100 \quad 100)^T$$

## 6.3 COMPUTATIONAL CHALLENGES

This chapter presents the basic steps required to reconstruct and analyze genome-scale metabolic networks. These model systems quickly grow in size and scale, introducing computational challenges that need to be addressed. As noted previously, with large-scale models it becomes necessary to use a robust computational platform designed specifically for sophisticated optimization problems, such as those developed by LINDO Systems, Inc. and available through GAMS.

Furthermore, data management becomes difficult as models scale up in size. For example, the most current published *E. coli* model contains 904 genes and 931 unique biochemical reactions [61]. Analyzing a genome-scale model within the framework proposed in Sections 6.2.4 and 6.2.5 is possible, but would be slow, cumbersome, and error prone. In recent years, an integrative data management and analysis software platform called SimPheny™ (Genomatica, San Diego, CA) has been developed specifically to address the data management and computational challenges inherent in building large-scale cellular models. This versatile platform provides network visualization, database, and various analytical tools that greatly facilitate the construction and study of genome-scale cellular models.

Currently, more than a dozen genome-scale metabolic models have been published and are available (Table 6-1) for further research and analysis. Most of these models represent bacteria and range from the important model organism *E. coli* [61–63] to pathogenic microbes such as *H. pylori* [64,65], and *S. aureus* [66]. Furthermore, recently developed models of *G. sulfurreducens* [67] and *S. coelicolor* [68] are potentially important for their facilitation of studies that probe these organisms' respective potential bioenergetic and therapeutics-producing properties.

Representative constraint-based models have also appeared from the other two major branches of the tree of life. The recently developed metabolic reconstruction of *M. barkeri* [69], an interesting methanogen with bioenergetic potential, represents the first constraint-based model of an archaea that has been used to aid in the analysis of experimental data from this relatively obscure group of organisms. Furthermore, several eukaryotic models also have been developed. The metabolic models of the baker's or brewer's yeast *S. cerevisiae* [70-72] are second only to the *E. coli* models in terms of relative maturity and have been used in a variety of studies designed to assess network properties (for recent examples, see Refs [73,74]). Metabolic models of higher order systems are also becoming available such as a model of mouse (*Mus muculus* [75]), human cardiac mitochondria [56], and red blood cell [76].

As more of these genome-scale models are developed, the issue of making their contents available to the broader research community is of primary concern. Given their inherent complexity, there is a need for a standardized format in which their contents can be represented in order to circumvent potential problems associated with the current typical means of distribution of models via nonstandard flat file or spreadsheet format. In an effort to mitigate this deficiency, for example, the Systems Biology Markup Language (SBML) [77], has been developed to provide a uniform framework in which models can be represented, and the recently initiated MIRIAM ("minimum information requested in the annotation of biochemical models") project [78] and affiliated databases have appeared to provide greater transparency as to the contents, and potential deficiencies of models. The adoption of these or similar standards will be important to the advancement of the field and in promoting its general utility in biological research.

### 6.3.1 Predicting Gene Essentiality

One application of constraint-based modeling in conjunction with FBA that has been particularly successful in computationally assessing metabolic networks is in studies of gene essentiality. Recent studies have used genome-scale constraint-based models to assess gene essentiality for several organisms under various growth conditions. Each study simulated gene deletions by constraining the flux through the associated reaction(s) to zero as described in Section 6.2.5 and Box 6.1. In this section, we will review the results from studies performed using models of *E. coli* [53,63], *H. Influenzae* [79], *H. pylori* [64,65], and *S. cerevisiae* [70-72] as a platform on which we can highlight some of the benefits and limitations of genome-scale metabolic models.

***6.3.1.1 Escherichia coli*** The bacterium *E. coli* is historically the most studied and perhaps the best characterized model organism to date, and is of important industrial, genetic, and pathologic importance. Thus, *E. coli* is among the most suitable organisms for metabolic reconstruction and constraints based analysis. Accordingly, constraint-based models of *E. coli* have been under development since 1990 (for a historical review of the *E. coli* constraint-based model development [62]). Prior to the complete determination of its genome sequence in 1997 [80], *E. coli* models were limited by data availability and thus included only 300 reactions. The success of genome sequencing efforts, coupled with other high-throughput technological advances led to a dramatic increase in size and scope of available models, yielding the first genome-scale models. The most current model of *E. coli* K-12 MG1655 metabolism includes 904 genes, 931 unique biochemical reactions, and 625 metabolites [61] and will soon crest the 1000 gene, 1000 reaction mark (A. Feist, B. Palsson, personal communication).

Following the completion of the first genome-scale model, gene essentiality was examined by investigating the effects of single-gene deletions on the metabolic capability of *E. coli* [63]. Gene deletions were simulated by restricting flux through the corresponding enzymatic reaction to zero. Each individual gene involved in the central metabolic pathways (glycolysis, pentose phosphate pathway, TCA, and respiration processes) was subjected to deletion under an environment of aerobic growth on minimal glucose medium.

Eleven (*rpiAB, pgk, acnAB, gltA, icdA, tktAB, gapAC*) of the simulated deletion strains failed to grow, and an additional 12 (*atp, fba, pfkAB, tpiA, eno, gpmAB, nuo, ackAB, pta*) deletion strains were impaired in their growth characteristics relative to wild type. When grown on glucose, these genes are involved in the three-carbon stage of glycolysis, three reactions of the TCA cycle, and several points within the pentose phosphate pathway.

This study also simulated gene deletion effects on *E. coli* when grown on other minimal medium formulations. Of 79 cases tested, 68 (86 percent) of the *in silico* predictions matched experimental observations. Most of the mischaracterized growth predictions were failure to predict no growth. A later study showed that the incorporation of transcriptional regulatory information can improve performance of the *E. coli* model [53]. Furthermore, the most recent model of *E. coli* is enhanced by containing elementally and charge balanced reactions, as well as GPR associations [61]. As models are further enhanced and a more detailed knowledge and representation of the biomass formulation is acquired, the predictive performance of these models will continue to improve.

***6.3.1.2 Haemophilus influenzae*** *H. influenzae* is a Gram-negative pathogen adapted to living in the upper-respiratory mucosa, causing ear infections as well as acute- and chronic-respiratory infections primarily in children. Prior to development of the first vaccine in 1985, *H. influenzae* type b was the leading cause of bacterial meningitis in children less than 5 years of age. Based upon the annotated genome sequence and known biochemical information, the metabolic network of this microbe was reconstructed [79,81]. Four hundred of the approximately 1743 open reading frames (ORFs) predicted to exist in *H. influenzae* were included in the model. By including 49 additional reactions based on general metabolic information on related prokaryotes, the final network consists of 461 reactions acting on 367 internal and 84 external metabolites.

In addition to studying various systems properties of the network, gene essentiality was assessed in *H. influenzae* by examining the effects of simulated gene deletions on growth characteristics. Gene deletions were simulated by constraining the flux through the corresponding enzyme catalyzed reaction to zero. One study examined the single, double, and triple deletion of a set of 36 enzymes involved in central intermediary metabolism [81]. For each simulation, the ability of *H. influenzae* to exhibit *in silico* growth was assessed on a defined media, with fructose and glutamate being the two key substrates. Under these conditions, 12 genes (*eno, fba, fbp, pts, gapA, gpmA, pgi, pgk, ppc, rpiA, tktA, tpiA*) were essential for growth and an additional 10 enzyme deletion strains (*cudABCD, atp, ndh, ackA, pta, gnd, pgl, zwf, talB, rpe*) exhibited impaired growth. This result suggests that *H. influenzae*'s metabolic network is less robust against single central metabolic gene deletions than *E. coli* [63]. Examination of all possible double mutants for evidence of so-called synthetic lethal interactions [82] within this set of enzymes revealed only 7 of 361 lethal gene pairs where each single deletion mutant was viable. Similarly, only 7 of 5270 lethal gene triple knockouts were observed where the double deletion of any two of the gene products does not result in a null phenotype.

A related study also examined simulated gene deletion effects on an expanded 42 enzyme set under two different growth conditions [79]. When simulating growth on a minimal media with fructose as the primary carbon source, the 11 single-gene deletions (*fba, fbp, tpiA, gapA, pgk, pgmA, eno, rpiA, tktA, prsA, ppc*) failed to grow. In order to study these same single-gene mutants under more relevant *in vivo* conditions, simulations were performed in similar fashion, this time using media supplemented with a number of carbon sources likely to be found in the host, mucosal environment. These include fructose, glucose, glycerol, galactose, fucose, ribose, and sialic acid. Under these conditions six mutant strains (*gapA, pgk, pgmA, eno, ppc*) again failed to grow. While these predictions require experimental verification using methods described elsewhere in this volume, these studies show the utility of computational studies in directing the researcher to the most interesting targets.

**6.3.1.3 Helicobacter pylori** *H. pylori* is a human bacterial pathogen that colonizes the gastric mucosa. Infection results in acute inflammation and damage to epithelial cells, ultimately progressing to a number of disease states, including gastritis, peptic ulceration, and gastric cancer. In comparison with the examples of *E. coli*, *H. influenzae, and S. cerevisiae*, provided elsewhere in this chapter, a little experimental data is available to complement the known genome sequence of *H. pylori*. Recent work shows that even in the absence of extensive experimental data, detailed metabolic models of great utility can be developed, using *H. pylori* as an example [64]. Relying primarily on annotated genome sequence, a model comprised of 388 enzymatic reactions, corresponding to 291 of 1590 known ORFs and 403 metabolites, was developed for *H. pylori*.

Similar to the other modeling efforts described in this chapter, systemic properties of this model were examined, as was gene essentiality by simulating gene deletions. The effect of the loss of enzymatic function corresponding to a gene deletion was assessed under four different simulated growth conditions. The growth conditions include a previously determined minimal medium required to support *in silico* growth, minimal medium supplemented with glucose, other carbon sources, and amino acids.

Each of the 34 reactions in the central intermediary metabolic network was individually eliminated by constraining the flux through the reaction to zero. Of these simulated gene deletions, only four (*aceB, ppa, prsA, and tpi*) failed to exhibit simulated growth under all four conditions. These particular knockouts affect malate synthase activity, the pyrophosphate to inorganic phosphate conversion, synthesis of nucleotides and deoxynucleotides, and impaired glycolytic ability.

The predictive performance of the model was also assessed via comparison with available experimental gene deletion data for *H. pylori* [64]. The model accurately predicted the growth ability for 10–17 gene deletions. Of the seven incorrect predictions, six were predicted to be nonessential when experimental evidence showing their essentiality. This discrepancy could signal deficiencies in the model. For example, given that the model is not complete, all of the relevant information may not be available to accurately predict the given case. In contrast, it could also be because of the experimental conditions not corresponding exactly with simulated conditions. In any case, these discrepancies identify keen areas of interest to probe with further experiments in an effort to ultimately improve and enhance the model.

An updated *H. pylori* was also recently developed and published [65]. This expanded model includes 341 genes and 476 intracellular metabolic reactions. A single-gene deletion analysis was carried out in which the growth capability of all 341 knockouts was assessed by constraining the flux through the associated reaction(s) to zero for FBA simulations of growth on both minimal and rich medium. More than 70 percent of the 72 predictions for which experimental data was available were in congruence. Furthermore, this result represents an improvement over the previous version of the *H. pylori* reconstruction [64]. A simulated double-deletion, or synthetic lethal, screen was also carried out using this network by constraining the flux for reactions associated with all pairwise combinations of genes in the model to zero. Of the more than 22,000 combinations that were tested, 47 pairs involving 64 unique genes were found to be lethal. While no corresponding experimental data exists for validation, this effort is still quite useful in that it can direct researchers to the potentially more interesting portions of the network for experimental investigation in the absence of a labor-intensive high-throughput screen.

***6.3.1.4 Saccharomyces cerevisiae*** In recent years, using the vast quantities of data available for the baker's yeast, *S. cerevisiae*, researchers have developed a genome-scale reconstructed metabolic model using the constraint based approach [70,83]. A total of 708 metabolism related ORFs were accounted for in the reconstructed network, corresponding to 1035 metabolic reactions. An additional 140 reactions were included based on biochemical evidence without direct knowledge of a responsible enzyme, ultimately yielding a reconstructed network containing 1175 metabolic reactions and 584 metabolites. This model has been shown in most cases to predict growth characteristics consistent with observed phenotypic functions [83].

Gene essentiality was assessed by using this large-scale model of *S. cerevisiae* to computationally evaluate the effect of 599 single-gene deletions on viability [84]. In this study, growth of yeast was simulated under aerobic conditions and on complete medium containing glucose, the 20 essential amino acids, and nucleic acids. Ammonia, phosphate, and sulfate were also supplied. Gene deletions were simulated by constraining the flux through the corresponding reactions to zero and optimizing for growth, as in previous studies in *E. coli* [85], and performance was gauged through comparison with experimental data.

The model performed remarkably well, accurately predicting the effect of 90 percent of these mutant alleles. In concurrence with experimental observation [86], a small fraction of these deletion strains exhibit impaired growth and fewer are still lethal. It should be noted, however, that the model had the most difficulty in correctly predicting null mutant phenotypes. This can be attributed to incomplete biochemical information, inadequate biomass equation definition, and gene regulatory effects. Addressing each of these issues in future work will likely improve the model's predictive capability. Future studies might also include the examination of lethal double mutants, also known as synthetic lethality [83,87], as these may provide better insight than single deletion mutants into gene essentiality and network robustness in *S. cerevisiae*.

### 6.3.2 Model Performance Assessment

Validating model predictions is a critical component in constraint-based model analysis. Growth phenotype data, available for a number of knockout strains and organisms, can be acquired from biochemical literature [88] and online databases, including ASAP [89] for *E. coli*, as well as CYGD and SGD for *S. cerevisiae*. As noted in the previous section, experimental growth phenotype data is available to assess directly the predictive power of the model for three of the four organisms listed previously, and shows that correct predictions were made in approximately 60, 86, and 83 percent of cases for *H. pylori* [64], *E. coli* [53], and *S. cerevisiae* [71], respectively. These comparisons serve two important functions: Validation of the general predictive potential of the model and identification of areas that require refinement. In this sense, constraint-based models are particularly useful in experimental design by directing research to the most or least poorly understood biological components. The next section details how to interpret incorrect model predictions and their likely causes.

### 6.3.3 Troubleshooting Incorrect Predictions

In the studies discussed in Section 6.3.1, the model predictions when compared to experimental findings failed most often by falsely predicting growth when the gene deletion leads to a lethal phenotype *in vivo*. This trend indicates that the most common cause of false predictions is because of the lack of information included in the network. For example, certain important pathways not related to metabolism in which the deleted gene participates may not be represented. In addition, the objective function may not be defined properly by failing to include the production of a compound required for growth. This case was shown to account for many false predictions when using a yeast metabolic model to account for strain lethality [72] when a few relatively minor changes to the biomass function dramatically improved the model's predictive capability. Alternatively, the gene deletion may lead to the production of a toxic by-product that ultimately kills the cell, a result for which this approach cannot account. Furthermore, certain isozymes are known to be dominant whereas metabolic models typically assign equal ability to each isozyme. The model would predict viable growth for the dominant isozyme deletion whereas *in vivo*, the minor isozyme(s) would not sufficiently rescue the strain from the lethal phenotype perhaps due to lower gene expression or enzymatic activity.

An additional major error source stems from the lack of regulatory information incorporated into the previously described models. Including transcription factor, metabolic gene interactions, using a Boolean logic approach, enhance the accuracy of constraint-based model predictions [53]. Regulatory information is available in the primary literature, in addition to online resources such as EcoCyc and RegulonDB [90]. Furthermore, these interactions can be derived from ChIP-chip analysis of transcription factors and corresponding gene expression microarray data [91]. A more detailed treatment of this latter topic is presented in Section 6.4.3.

Incorrect predictions are less often due to false predictions of lethality. These uncommon cases often suggest the presence of previously unidentified enzyme activities, which if added to the model, would lead to accurate predictions. They may also reflect improper biomass function definition, but in a different sense from the situation described above. For example, rather than failing to include compounds required for growth, it is also possible that certain compounds are included in the biomass function erroneously, and may actually not be essential to support biological growth. In any case, inaccurate [12,92] predictions are most often attributed to a paucity of information available for inclusion in the model and not simply a failure of the technique, thus validating the general strategy of constraint-based modeling.

### 6.3.4 Additional Analytical Tools

A rapidly growing collection of analytical methods have been developed for use in conjunction with constraint-based models [12], some of which we briefly introduce in this section. Although many of the examples in this chapter focus on the use of constraint-based models to assess gene essentiality, these models can also be used to predict behavior of viable gene deletions. For example, FBA uses LP to identify the optimal metabolic state of the mutant strain. In contrast, Minimization of Metabolic Adjustment (MOMA) uses quadratic programming (QP) to identify optimal solutions that minimize the flux distribution distance between a wild type and simulated gene deletion strain [93,94]. Experimental data seems to confirm the MOMA assumption that knockout strains utilize the metabolic network similar to wild type [93]. It remains to be determined if this is true in all situations or if the network optimizes for growth over time following gene deletion.

A more recent method known as regulatory on/off minimization (ROOM) [95] is another constraint-based analysis technique that uses a mixed integer linear programming strategy to predict the metabolic state of an organism following a gene deletion by minimizing the number of flux changes that occur with respect to wild type. In other words, this algorithm aims to identify flux distributions that are qualitatively the most similar to wild-type in terms of the number and types of reactions that are utilized. While MOMA seems to better predict the initial metabolic adjustment that occurs following the genetic perturbation, ROOM, like FBA, better predicts the later, stabilized growth phenotype.

Constraint-based modeling also has applications in the metabolic engineering field. Identifying optimal metabolic behavior of mutant strains using a bilevel optimization framework has been employed by OptKnock [96]. This metabolic engineering strategy uses genome-scale metabolic models and a dual-level, nested optimization structure to predict which gene deletion(s) will lead to a desired biochemical production while retaining viable growth characteristics. This technique establishes a framework for microbial strain design and improvement and has the potential for significant impact. These and other analytical techniques and applications that rely on constraint-based modeling will be discussed in detail in Chapter 11.

Additional methods have been developed to specifically assess the systemic or topological properties of these networks [12]. Extreme pathway analysis [97] represents one such technique that utilizes convex analysis of the *S* matrix to define a cone that circumscribes all allowable steady-state solutions within the space defined by the *S* matrix and its associated constraints (see Fig. 6-1 for a conceptual representation of the this space, also known as the "solution space"). Accordingly, all possible routes through the network can be described by nonnegative combinations of the generated extreme pathways. This technique and analysis of the extreme pathways themselves have been fruitful in a variety of studies (for examples, see Refs [98–100]) and can be readily calculated for reasonably sized networks using available software [79,101].

## 6.4 FUTURE DIRECTIONS FOR CONSTRAINT-BASED MODELING

Thus far, constraint-based models have had their primary success in assessing the metabolic capabilities of cells, but fail to account for many other important aspects of cellular biology. In the past several years, however, several efforts have been initiated to apply the constraint-based modeling and analysis techniques to other cellular processes. Below we briefly describe relatively recent work that is setting the stage for including RNA and protein synthesis [102] as well as other processes governed by cell signaling [103] and transcriptional regulatory networks (TRNs) into genome-scale, constraint-based models of the cell.

### 6.4.1 Modeling of RNA and Protein Synthesis

RNA and protein synthesis represent two of the primary energy drains on the cell [58] and are of obvious vital importance in that these processes give rise to many of the active components responsible for cellular activities. Existing constraint-based genome-scale metabolic models do not explicitly account for these processes, rather they are included as abstract, lumped sum quantities of monomeric amino acid, and nucleotide triphosphate demand required to support cellular growth [104]. The specific values for these quantities are determined from measurements of biomass constituents [58] and are independent of the genome sequence. In order to meet this deficiency in the field, a scalable, constraint-based framework was developed to capture the metabolic requirements for gene expression and protein synthesis directly from the genome sequence [102].

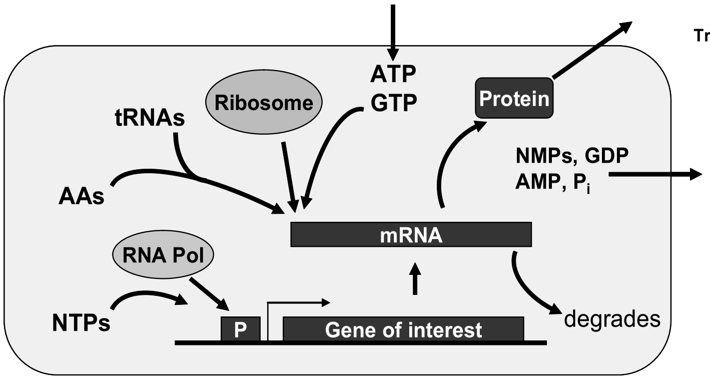
The general strategy stems from the observation that RNA and protein synthesis can be broken down into constitutive biochemical reactions that underlie the processing of these polymers. As illustrated in Figure 6-5, the expression of a given gene and the synthesis of the protein that it encodes can be modeled by six essential biochemical reactions. These reactions include transcription initiation, transcription elongation, mRNA degradation, translation initiation, translation elongation, and tRNA charging. Biochemical equations representing each of these processes can be compiled (Fig. 6-5b) and used to formulate an associated *S* matrix (Fig. 6-5c).

Many of the previously introduced analytical tools can then be used to computationally assess the properties of the *S* matrix. For example, by choosing protein production as the objective, FBA can be used to determine how much the protein that the RNA and protein synthesis machinery within the cell can produce for a given set of environmental conditions and resources [102]. One can also incorporate promoter strength, transcription elongation, and translational initiation constraints on the system if such information is known or can be approximated. Extreme pathway analysis can also be used to assess the capabilities of these systems and their characteristic states [102]. Thus far, however, this framework and analysis methods have only been applied to small biological systems, namely the malate dehydrogenase (*mdh*) gene and the *lac* operon [102]. Accordingly, the limitations associated with studying large-scale systems in this manner remain to be assessed, although an ongoing study of the *E. coli* RNA and protein synthesis network (I. Thiele and B. Palsson, personal communication) is certain to be illuminating.

### 6.4.2 Modeling of Cell Signaling Networks

The signal transduction pathways that comprise cell signaling networks are responsible for many critical processes. Signaling events operate both on relatively quick timescales, such as those that cause posttranslational protein changes, and long timescales, such as cell cycle control, cell proliferation and migration, as well as apoptosis. Cell signaling networks are often highly connected and complex involving many molecular players. In an effort to quantitatively characterize their properties, researchers are beginning to reconstruct these networks and apply mathematical methods to analyze them.

One approach to computationally analyzing cell signaling networks relies on many of the same constraint-based modeling principles discussed earlier in this chapter for metabolic networks [103,105]. The key insight is to treat signaling pathways as a series of biochemical transformations starting with an input (the signal) and resulting in an output (posttranslational protein modification, apoptosis, etc.). Accordingly, just as in modeling metabolic networks the first steps of this process focus on network reconstruction. One must first identify the components in the signaling network of interest and the interactions that occur between them. In contrast to modeling of metabolic networks where enzymes and metabolites are the primary players, signaling networks typically include receptors and their corresponding receptor ligands,



(*a*)

| | | | |
|---|---|---|---|
| **Transcription initiation:** | Gene + RNA Pol | ⟶ | Gene* |
| **Transcription:** | Gene* + *n* NTP | ⟶ | mRNA + Gene + RNA Pol + 2*n* $P_i$ |
| **mRNA decay:** | mRNA | ⟶ | *n* NMP |
| **Translation initiation:** | mRNA + Ribosome | ⟶ | Ribosome* |
| **Translation:** | Ribosome* + *a* AAtRNA +2*a* GTP | ⟶ | *a* tRNA +2*a* GDP + 2*a* $P_i$ +<br>Ribosome + mRNA + Protein |
| **tRNA charging:** | AA + tRNA + ATP | ⟶ | AMP + 2$P_i$ + AAtRNA |
| **Exchange reactions:** | $AA_{ext}$ | ⟶ | AA |
| | $NTP_{ext}$ | ⟶ | NTP |
| | $ATP_{ext}$ | ⟶ | ATP |
| | $GTP_{ext}$ | ⟶ | GTP |
| | Protein | ⟶ | $Protein_{ext}$ |
| | NMP | ⟶ | $NMP_{ext}$ |
| | AMP | ⟶ | $AMP_{ext}$ |
| | GDP | ⟶ | $GDP_{ext}$ |
| | Pi | ⟶ | $P_{i\ ext}$ |

(*b*)

**Figure 6-5** Constraint-based modeling of RNA and protein synthesis.(a) A hypothetical system that represents the RNA and protein synthesis network associated with the transcription and translation of a single gene is depicted. The processes of transcription initiation, transcription, mRNA decay, translation initiation, translation, and tRNA charging are depicted. Also shown are some of the exchange fluxes required to balance the system. (b) A biochemical reaction list for the included processes and appropriate exchange reactions can be compiled. Note that the precise stoichiometry can and should be included in each reaction definition. In this system the gene and associated protein length can vary. Accordingly, variables for the number of bases (*n*) and number of amino acids (*a*) are included in the reaction stoichiometry. (c) The *S* matrix can then be formulated based on the reaction list. System components are represented in respective rows and each column denotes individual system reactions. AA, amino acid; AAtRNA, charged tRNA; Gene*, gene undergoing transcription; NMP, nucleotide monophosphate; Pi, inorganic phosphate; Ribosome*, actively translating ribosome; RNA Pol, RNA polymerase.

| | Transcription Init. | Transcription | mRNA decay | Trans-lation Init. | Trans-lation | tRNA Charging | AA influx | NTP influx | ATP influx | GTP influx | Protein efflux | NMP efflux | AMP efflux | GDP efflux | $P_i$ efflux |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Gene** | —1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Gene*** | 1 | —1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **RNA Pol** | —1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **NTP** | 0 | —n | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **mRNA** | 0 | 1 | —1 | —1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **$P_i$** | 0 | 2n | 0 | 0 | 2a | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| **NMP** | 0 | 0 | n | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Ribosome** | 0 | 0 | 0 | —1 | 1 | 0 | 0 | —1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Ribosome*** | 0 | 0 | 0 | 1 | —1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **AAtRNA** | 0 | 0 | 0 | 0 | —a | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Protein** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| **GTP** | 0 | 0 | 0 | 0 | —2a | 0 | 0 | 0 | 0 | —1 | 0 | 1 | 0 | 0 | 0 |
| **GDP** | 0 | 0 | 0 | 0 | 2a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| **AA** | 0 | 0 | 0 | 0 | 0 | —1 | —1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **tRNA** | 0 | 0 | 0 | 0 | a | —1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **ATP** | 0 | 0 | 0 | 0 | 0 | —1 | 0 | 0 | —1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **AMP** | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

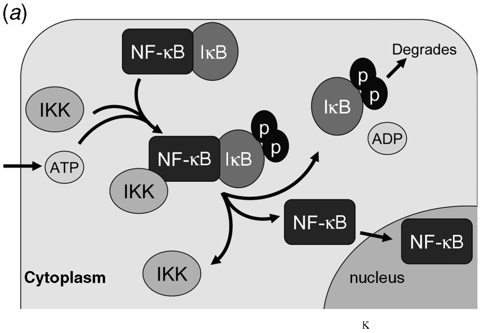
(*c*)

**Figure 6-5** (*Continued*)

metabolites such as ATP and ADP, as well as intracellular signal transducing proteins. Also, these networks often include transcription factors, transcription factor binding sites, and the resulting target genes.

The data from which components and their interactions are derived have been traditionally difficult to obtain due to the often laborious effort involved in mapping signaling pathways using standard molecular biology techniques. However, recently developed high-throughput, genome-scale techniques are mitigating this issue. For example, whole genome sequencing and annotation identifies the possible network components, ChIP-chip assays identify protein-DNA interactions, and yeast two-hybrid assays identify protein–protein interactions. As previously noted, Table 6-2 summarizes many useful online resources that contain publicly accessible data. Several strategies for mapping signaling pathways and networks have been developed in recent years by integrating these and other high-throughput data [4]. These methods have been employed to map DNA damage response as well as developmental pathways [4] among others.

Having identified the components and interactions that occur between them, a list of biochemical reactions that describes the cell signaling network can be listed. A stoichiometric matrix is then derived from this list (Fig. 6-6) in very much the same



(*b*)

| | | | |
|---|---|---|---|
| **IKK complex formation:** | IKK + IκB-NF-κB + 2ATP | ⟶ | IKK-IκB-NF-κB-ATP |
| **IKK complex dissociation** | IKK-IκB-NF-κB-ATP | ⟶ | IKK + IκBpp-NF-κB + 2ADP |
| **NF-κB release:** | IκBpp-NF-κB | ⟶ | IκBpp + NF-κB (cytoplasm) |
| **NF-κB nuclear translocation:** | NF-κB (cytoplasm) | ⟶ | NF-κB (nucleus) |
| **IκBpp decay:** | IκBpp | ⟶ | |
| **Exchange** fluxes: | ATP (external) | ⟶ | ATP |
| | ADP | ⟶ | ADP (external) |

(*c*)

| | IKK complex form. | IKK complex dissoc. | NF-κB release | NF-κB nuclear transloc. | IκBpp decay | ATP influx | ADP efflux |
|---|---|---|---|---|---|---|---|
| **IKK** | –1 | 1 | 0 | 0 | 0 | 0 | |
| **IKK-IκB-NF-κB-ATP** | 1 | –1 | 0 | 0 | 0 | 0 | 0 |
| **IκB-NF-κB** | –1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **IκBpp-NF-κB** | 0 | 1 | –1 | 0 | 0 | 0 | 0 |
| **IκBpp** | 0 | 0 | 1 | 0 | –1 | 0 | 0 |
| **NF-κB (cytoplasm)** | 0 | 0 | 1 | –1 | 0 | 0 | 0 |
| **NF-κB (nucleus)** | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **ATP** | –2 | 0 | 0 | 0 | 0 | –1 | 0 |
| **ADP** | 0 | 2 | 0 | 0 | 0 | 0 | 1 |

**Figure 6-6** Constraint-based modeling of cell signaling networks. (a) A schematic that includes a portion of the nuclear factor (NF)-*κ*B signaling-related network is depicted. (b) A reaction list that corresponds to the schematic in (a) is detailed. Reactions are included for the interaction of I*κ*B kinase (IKK) with the inhibitor of NF-*κ*B (I*κ*B)-NF-*κ*B complex. The subsequent phosphorylation of I*κ*B and release of NF-*κ*B are also shown in addition to the degradation of phosphorylated I*κ*B (I*κ*Bpp) and NF-*κ*B translocation to the nucleus, and exchange fluxes required for the system. (c) The associated *S* matrix is compiled based on the reaction list. System components are depicted in each respective row, and reactions are represented in each column.

manner as previously described for metabolic as well as RNA and protein synthesis networks. It is important to note that each state of a component must be explicitly accounted for in the network. For example, a protein must be differentially represented in separate phosphorylated and unphosphorylated forms [105,106].

This stoichiometric framework explicitly defines the underlying network reactions in a chemically consistent form. Accordingly, network properties can be readily and quantitatively assessed using previously introduced analytical tools. Extreme pathway analysis, in particular, is an immensely useful tool for characterizing cell signaling networks. Using existing software [101,107], one can enumerate the extreme pathways using the stoichiometric matrix from the reconstructed cell signaling network.

All routes through the cell signaling network can be described by nonnegative linear combinations of the extreme pathways. Accordingly, network cross talk, signaling redundancy, correlated reaction sets, as well as reaction participation and likely relative importance are all properties that can be derived from this analysis. Network cross talk refers to an analysis of how disjoint, overlapping, or identical inputs can lead to disjoint, overlapping, or identical outputs within a signaling network and are derived from pairwise comparisons of individual extreme pathways. Strictly speaking, signaling redundancy is the multiplicity of routes through a network by which identical inputs lead to identical outputs, but it can be further delineated into considerations of input and redundancy alone. Correlated reaction sets are a collection of reactions that is always either present or absent in all of the extreme pathways. In other words, these sets of reactions represent functional modules that act together in a given network, although the reactions themselves may not necessarily be adjacent on the reaction map. Finally, reaction participation is the percentage of pathways in which a given signaling reaction is used. This relatively simple calculation can indicate important biological insights. For example, reactions with high participation values are likely to be critical for network functionality while low participation values indicate more specialized portions of the network.

Thus far, the stoichiometric approach to modeling signaling networks has only been applied to a prototypic network [105] and the human B-cell JAK (Janus activated kinase)–STAT (signal transducer and activator of transcription) signaling network [106]. While the prototypic network study served simply as proof of concept, the work on the JAK–STAT network showed that the constraint-based approach can be used to analyze real biological systems and yield quantitative insights into its properties. Accordingly, as more signaling networks are delineated and reconstructed, this approach will likely be of great utility.

### 6.4.3 Modeling of Transcriptional Regulatory Networks

With the huge success of whole genome sequencing efforts and the appearance of hundreds of genome sequences, there is an increased interest in understanding how the genes within a given genome are regulated through complex TRNs). Consequently, efforts are underway to define and catalog the set of regulatory rules for model organisms. Due to the large number of regulated genes and associated regulatory proteins as well as their extensive interconnectivity, there is a significant need for a structured framework to integrate regulatory rules and interrogate TRN functions in a systematic fashion.

Previous work has integrated models of regulatory and metabolic networks to analyze and predict the effect of transcriptional regulation on cellular metabolism at the genome scale [50,52,53,108]. These studies developed and utilized a framework in which regulatory rules are represented as Boolean logic rules that control the expression of enzyme-encoding genes that ultimately facilitate metabolic reactions within a constraint-based metabolic model of the type described previously within this chapter. The regulatory rules are defined such that metabolic enzyme genes are determined to be present or absent based on the presence or absence of extracellular and intracellular metabolites. If an enzyme-encoding gene is determined to be absent then the flux through that enzyme is set to zero in the metabolic model, which in effect adds a temporary constraint on the system. In effect, this is equivalent to carrying out FBA on the network following a gene deletion.

Using an iterative computational scheme in which time, $t$, is divided into small steps (usually on the order of minutes), a dynamic profile of growth can be simulated. At $t = 0$ the metabolic model is used to predict the optimal flux distribution for the network using FBA, as described in Section 6.2.4. The resulting flux distribution is used as initial conditions from which the Boolean transcriptional regulatory rules are evaluated. The rule evaluations specify the transcriptional status of enzymes for the next time step. As noted above, if the transcriptional state of an enzyme-encoding gene results in the absence of the corresponding enzyme the reaction flux(es) mediated by it are set to zero for the FBA carried out on the system for the next time step. This process of iterative Boolean rule evaluation and FBA calculation continues for the user-defined time span [50,52].

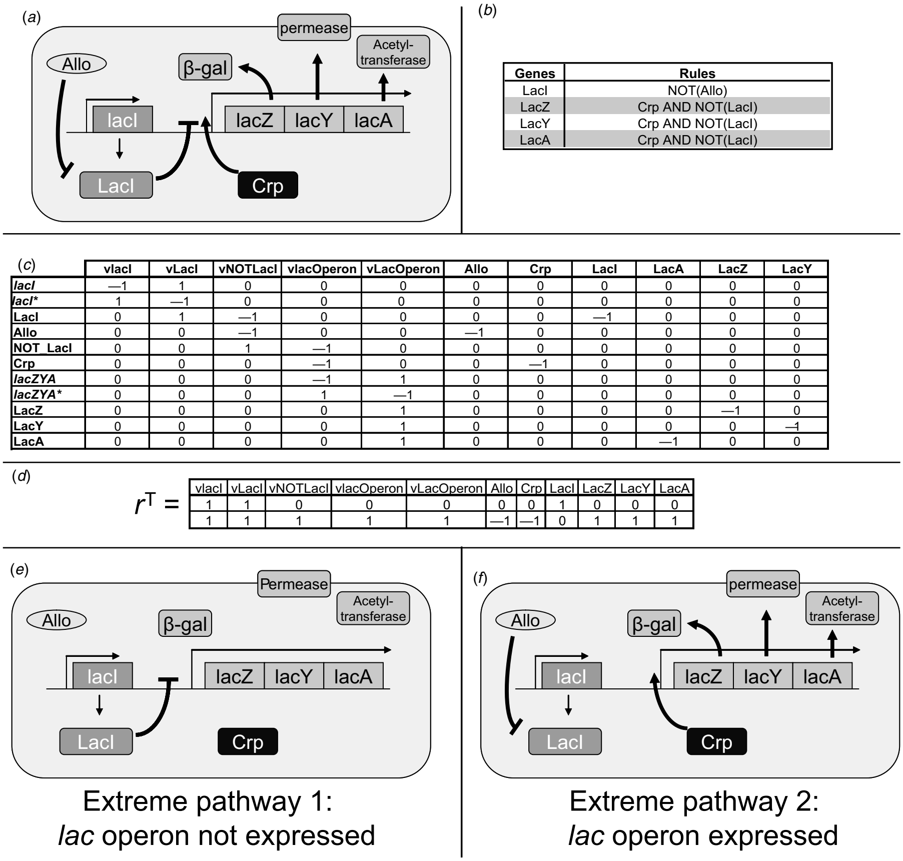
This type of integrated analysis of metabolic and regulatory networks has been performed for both small prototypic systems [52] as well as for a genome-scale model of *E. coli* [50] and more recently in yeast [108]. In the study of *E. coli*, this analysis was performed in conjunction with dual perturbation growth experiments coupled with genome-wide expression analysis. This systematic approach to reconstructing and interrogating the integrated network of *E. coli* led to the identification of many novel regulatory rules, and an expanded characterization of the genome-scale TRN, based on a model-driven analysis of multiple high-throughput data sets. Furthermore, a recent study has also used this model in a large-scale simulation project to study all potential network states and found them to be organized primarily based upon terminal electron acceptor availability [57]. However, one shortcoming of this framework is that it does not facilitate a detailed analysis of transcriptional regulatory network properties.

In an effort to address this limitation, a structured and self-contained representation of TRNs that can be quantitatively interrogated has been developed relying on the principles of the constraint-based approach [109]. This strategy, which effectively connects environmental cues to transcriptional responses, is conceptually similar to the previously described constraint-based approach to modeling cell signaling networks. The first step in the process involves defining the components of the system and interactions between them based on legacy data from traditional molecular biology studies or from recently generated high-throughput data. In particular, ChIP-chip data provides direct information regarding transcription factor-target gene relationships and genome-wide expression profiling data can yield insight as to the type of regulation. For example, when examined using conditions in which a given transcription factor is known to be active, the upregulation of a gene identified to be a target of the given transcription factor indicates that the transcription factor serves as an activator, whereas downregulation of the target gene suggests that the transcription factor acts as a repressor for that particular target gene. The environmental cues or stimuli to which the transcription factors respond as well as any required cofactors such as cyclic AMP (cAMP) must also be identified and included in this representation.

Having gathered this type of information that describes the regulatory system of interest, the next step is to write quasi stoichiometric, biochemical equations that describe the regulatory logic for each interaction in the network (Fig. 6-7b). The quasi stoichiometric nature of these equations is not required of course, but rather is used due to the general lack of specific chemical detail for most regulatory interactions. As the specific stoichiometry of regulatory interactions becomes available [110], however, higher levels of detail can be readily incorporated into this framework. As is the case in reconstructing cell signaling networks, it is important to reiterate that each state of a component must be explicitly accounted for in the network. For example, for regulatory networks, this case is encountered when transcription factors interact with cofactors to form activating or inhibitory complexes.

One peculiarity of this methodology is that it requires the inclusion of the converse of regulatory rules in addition to the regulatory rules themselves. The *converse* of the regulatory rules—the regulatory reactions that lead to the inhibition of gene transcription in our sample system—is necessary to reflect the lack of protein production for a given set of environmental cues. Many regulatory rules are inhibitory, such that the expression of a protein depends on the absence of a given metabolite or protein product. Additional reactions that include the converse of the regulatory rules and the absence of metabolites and protein products where appropriate must be included in the system. Also, note that regulatory rules of the Boolean type "OR" require two separate reactions to indicate that there are two independent ways in which the target gene can be transcribed.

A matrix can then be compiled from this list of biochemical reactions (Fig. 6-7c) in much the same way as was done for the other network types described previously in this chapter. Each row of the matrix describes a component of the system and each column represents regulatory events, or reactions. As a reminder, notice that each metabolite is represented in both present and absent forms, as is each transcription factor. Furthermore, the quasi stoichiometric formalism needs to be supplemented by *exchange* reactions that balance the entry of external cues or stimuli into the system as well as the production of proteins and their exit from the system. These exchange reactions describe the role of external cues and stimuli as inputs to the regulatory system and the role of the proteins as outputs of the transcriptional regulatory system. Therefore, columns representing the exchange of external stimuli as well as protein products are incorporated.



(*b*)

| Genes | Rules |
|---|---|
| LacI | NOT(Allo) |
| LacZ | Crp AND NOT(LacI) |
| LacY | Crp AND NOT(LacI) |
| LacA | Crp AND NOT(LacI) |

(*c*)

| | vlacI | vLacI | vNOTLacI | vlacOperon | vLacOperon | Allo | Crp | LacI | LacA | LacZ | LacY |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***lacI*** | —1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***lacI**** | 1 | —1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **LacI** | 0 | 1 | —1 | 0 | 0 | 0 | 0 | —1 | 0 | 0 | 0 |
| **Allo** | 0 | 0 | —1 | 0 | 0 | —1 | 0 | 0 | 0 | 0 | 0 |
| **NOT_LacI** | 0 | 0 | 1 | —1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **Crp** | 0 | 0 | 0 | —1 | 0 | 0 | —1 | 0 | 0 | 0 | 0 |
| ***lacZYA*** | 0 | 0 | 0 | —1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| ***lacZYA**** | 0 | 0 | 0 | 1 | —1 | 0 | 0 | 0 | 0 | 0 | 0 |
| **LacZ** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | —1 | 0 |
| **LacY** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | —1 |
| **LacA** | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | —1 | 0 | 0 |

(*d*)

$r^{\mathrm{T}} =$

| vlacI | vLacI | vNOTLacI | vlacOperon | vLacOperon | Allo | Crp | LacI | LacZ | LacY | LacA |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 | 1 | —1 | —1 | 0 | 1 | 1 | 1 |

Extreme pathway 1:
*lac* operon not expressed

Extreme pathway 2:
*lac* operon expressed

**Figure 6-7** Constraint-based modeling and analysis of transcriptional regulatory networks. (a) The *lac* operon regulatory system is depicted and defined to include the *lac* operon genes (*lacZ*, *lacY*, *lacA*), the inhibitor gene *lacI*, the activator *Crp*, and the inducer allolactose (*Allo*). (b) A reaction list that summarizes the Boolean rules that capture the regulatory logic of the system is shown. (c) The *R* matrix that corresponds to the regulatory rule list from (b) is depicted with each row corresponding to system components and each column specifying regulatory reactions in a quasi stoichiometric formalism. Accordingly, a "$-1$" represents a "consumed" component, whereas a "$+1$" represents a "produced" component. (d) The two extreme pathways for this system are listed in **r** with the corresponding reaction labels listed as well for reference. A nonzero value indicates that the corresponding reaction is active. The negative coefficients in the second extreme pathway reflect that *Allo* and *Crp* can be thought of as conceptually flowing into the system. (e) Pathway 1 is graphically illustrated and reflects the conditions for the LacI-mediated inhibition of the *lac* operon. (f) The graphical depiction of Pathway 2 shows the activation of the *lac* operon (i.e., inhibition of LacI by allolactose, thus allowing for derepression and Crp-activated expression of *lacZYA*). $r^{\mathrm{T}}$, the transpose of the extreme pathway vectors reported in *r* (depicted in this way simply out of space considerations).

With a regulatory matrix in hand, many of the analytical tools previously discussed can be applied to assess its properties. For example, extreme pathway analysis generates a set of vectors that encompasses all possible expression states of the network. Recall that all possible regulatory pathways, and thus expression states can be described as a nonnegative linear combination of these extreme pathways. Consequently, extreme pathway analysis represents an *in silico* technique for evaluating global characteristics of gene expression. Furthermore, the pervasiveness of signal inputs, percentage of environments in which a given gene is expressed, numbers of genes coordinately expressed, and correlated gene sets represent the type of data that can be readily generated based on extreme pathway analysis for a transcriptional regulatory network.

To illustrate some of these regulatory matrix ideas, we briefly consider the *lac* operon in *E. coli*. For the purpose of this investigation, the system is defined to include the *lac* operon (*lacZYA*) and the proteins each gene encodes, the inhibitor of the operon (*lacI*), an activator of the operon (Crp), and the intracellular inducer molecule allolactose, which inhibits the LacI inhibitor thus activating *lacZYA* transcription (Fig. 6-7a) by way of derepression.

Having defined the system (Fig. 6-7a) and Boolean rules that specify the regulatory logic of this small transcriptional regulatory network (Fig. 6-7b), the system can be formulated and the associated $R$ matrix constructed (Fig. 6-7c). For the purposes of this analysis, each gene/operon is depicted within the matrix twice: *lacI* and *lacI*$^*$, as well as *lacZYA* and *lacZYA*$^*$. The former entity represents the open form, whereas the latter, asterisk-marked entity represents the actively transcribed form of the gene. This level of detail is not required in formulating $R$ as the actively transcribed form of the gene is only a transient entity between transcription and translation. Rather, this is meant to show concretely that such mechanistic detail about ORFs and other network relationships can be readily incorporated into the current formalism as the data becomes available.

Extreme pathway analysis on this system yields two vectors, denoted by $r$ (Fig. 6-7d). Each entry in the vectors represents the activity of a reaction in the expression state, or pathway. For reaction names prefaced with a ‘‘*v*,’’ a 1 indicates that the reaction is active, and a 0 indicates that it is inactive. In the remaining reactions that specify flow across the system boundary, a 1 indicates flow out of the system (for example, a protein is produced), a -1 indicates flow into the system, and a 0 indicates that the associated component is neither produced nor consumed. Note that the entries are not quantitative but denote an active connection, and further, that a series of connections leads to a ‘‘causal path.’’ The first vector represents the LacI-mediated inhibition of the *lac* operon. The second vector defines the inhibition of LacI by allolactose, thus resulting in derepression and Crp-activated expression of *lacZYA*. These two vectors thus represent the two expression states of the *lac* operon system, as depicted graphically in Figure 6-7e and f.

Thus far, this approach has only been applied to the small *lac* operon system described above and a larger 25 gene prototypic network [111]. While this proof of concept study validates the utility of this approach for small systems, potential complications associated with scaling this approach up to genome-scale systems remain to be determined. Nonetheless, transcriptional regulatory network matrix reconstructions for model organisms will likely be important not only in studies of regulatory network properties but also in guiding experimental programs based upon results from these analyses.

### 6.4.4 The Next Big Challenge

The constraint-based approach has proven immensely successful for modeling metabolic models and, as described in this section, is showing promise for RNA and protein synthesis, cell signaling, and transcriptional regulatory networks. However, as the field currently stands, each respective framework produces models that exist as independent entities. Arguably, the ultimate goal of systems biology is to integrate data from disparate sources and generate comprehensive models that reflect biological reality for entire cells. Therefore, these modeling strategies present an opportunity to take a significant step forward in realizing this aim through integrative modeling efforts.

To elaborate, the interconnectivity between these distinct networks is clear. For example, a simplistic, but illustrative conceptual picture (Fig. 6-8) can be envisioned in

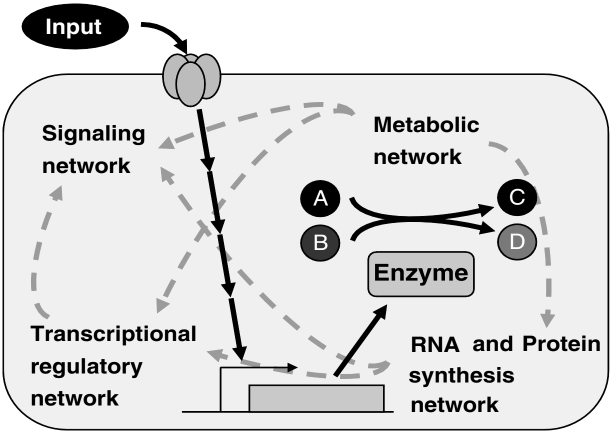

Input
Signaling network
Metabolic network
A
B
C
D
Enzyme
Transcriptional regulatory network
RNA and Protein synthesis network

**Figure 6-8** The next big challenge: model integration. This chapter has illustrated the utility of constraint-based modeling and analysis in computationally representing many cellular processes. To date, however, these models have been developed and analyzed in isolation despite the fact that these systems are all interrelated, as shown in this conceptual figure. For example, cellular signals or inputs are recognized by the cell signaling network, which in turn stimulate regulatory processes. These regulatory processes mediate RNA and protein synthesis ultimately leading to the production of enzymes that perform metabolic processes that result in cell growth or maintenance. The dashed arrows highlight the interconnectivity of these networks in the form of shared molecular components or feedback mechanisms. In principle, the constraint-based formalism can be used as a platform to capture these systems into a single picture. Accordingly, one of the next major challenges facing the field is to integrate these models of disparate cellular processes, thus pushing toward one of the field of systems biology's foundational goals: To computationally represent and analyze models of entire cells and biological systems.

which system inputs are recognized by cell signaling networks that in turn stimulate regulatory processes. These regulatory processes mediate RNA and protein synthesis ultimately leading to the production of enzymes that perform metabolic processes and lead to cell growth or maintenance. Additional connectivity between the systems also exists in the form of feedback processes and shared currency metabolites such as ATP and GTP, for example. Thus, in principle, the stoichiometric and pseudo-stoichiometric representations of the networks described in this chapter could be integrated into a unified model of the cell. While there are certainly computational challenges that will need to be overcome in order to facilitate the development and analysis of such a model, this notion seems feasible and is sure to be tackled in the near future. Representing additional cellular processes, such as differentiation, and accounting for multicellularity await novel research efforts and represent open problems to be addressed in the more distant future.

## 6.5 CONCLUSIONS

Despite the challenges outlined in the previous section associated with pushing the field forward, constraint-based modeling and its associated analyses are (and will remain as) powerful tools that facilitate system-level modeling [11,53,103] and analysis of biological networks [57,99,111,112]. Furthermore, these model-based studies can be used to help researchers prioritize experimental projects and save considerable time at the bench. Beyond its utility as a tool for basic biological research and in metabolic engineering applications [97,113], this computational approach also has potential medical relevance. For example, in pathogenic microbial models, each gene that is predicted to be essential by constraint-based modeling and analysis represents a potential drug target that could be used to develop effective therapeutics in the future. As more genome-scale models are developed and existing models enhanced, additional applications—a broad range of fields—will likely become apparent. Consequently, the flexibility of constraint-based models will continue to be exploited to drive the exploration of countless exciting biological questions in the future.

## REFERENCES

1. Wheeler DL, et al. Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res* 2004;32(Database issue): D35–D40.
2. Wyrick JJ, Young RA. Deciphering gene expression regulatory networks. *Curr Opin Genet Dev* 2002;12(2): 130–136.
3. Sanford K, et al. Genomics to fluxomics and physiomics—pathway engineering. *Curr Opin Microbiol* 2002;5(3): 318–322.
4. Joyce AR, Palsson BO. The model organism as a system: integrating omics' data sets. *Nat Rev Mol Cell Biol* 2006;7(3):198–210.
5. Arkin AP. Synthetic cell biology. *Curr Opin Biotechnol* 2001;12(6): 638–644.
6. Tomita M,et al. E-CELL: software environment for whole-cell simulation. *Bioinformatics* 1999;15(1): 72–84.

7. Hoffmann A, et al. The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science* 2002;298(5596): 1241–1245.
8. Elowitz MB, et al. Stochastic gene expression in a single cell. *Science* 2002;297(5584): 1183–1186.
9. Arkin A, Ross J, McAdams HH. Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* 1998;149(4): 1633–1648.
10. Sarkar A, Franza BR. A logical analysis of the process of T cell activation: different consequences depending on the state of CD28 engagement. *J Theor Biol* 2004;226(4): 455–466.
11. Reed JL, et al. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;7(2): 130–141.
12. 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.
13. Edwards JS, Covert M, Palsson B. Metabolic modelling of microbes: the flux-balance approach. *Environ Microbiol* 2002;4(3):133–140.
14. Covert MW, Famili I, Palsson BO. Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? *Biotechnol Bioeng* 2003;84 (7): 763–772.
15. Price ND, et al. Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol* 2003;21(4): 162–169.
16. Varma A, Palsson BO. Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl Environ Microbiol* 1994;60(10): 3724–3731.
17. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol* 2003;14(5): 491–496.
18. Neidhardt FC, Curtiss R. *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. Washington, DC: ASM Press, 2 v. xx, 1996, p. 2822.
19. Scheffler IE. *Mitochondria*. New York: Wiley-Liss, 1999,xiv p. 367.
20. Liolios K, et al. The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res* 2006;34(Database issue): D332–D334.
21. Consortium CSAA. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005;437(7055): 69–87.
22. Gibbs RA,et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* 2004;428(6982): 493–521.
23. Istrail S, et al. Whole-genome shotgun assembly and comparison of human genome assemblies. *Proc Natl Acad Sci USA* 2004;101(7): 1916–1921.
24. Kirkness EF, et al. The dog genome: survey sequencing and comparative analysis. *Science* 2003;301(5641): 1898–1903.
25. Stein L. Genome annotation: from sequence to biology. *Nat Rev Genet* 2001;2(7): 493–503.
26. Brent MR. Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res* 2005;15(12): 1777–1786.
27. Mao X, et al. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. *Bioinformatics* 2005;21(19): 3787–3793.

28. Chen Z. Assessing sequence comparison methods with the average precision criterion. *Bioinformatics* 2003;19(18): 2456–2460.

29. Karp PD, Paley S, Romero P. The Pathway Tools software. *Bioinformatics* 2002;18(Suppl 1): S225–S232.

30. Cash P. Proteomics of bacterial pathogens. *Adv Biochem Eng Biotechnol* 2003;8393–115.

31. Taylor SW, Fahy E, Ghosh SS. Global organellar proteomics. *Trends Biotechnol* 2003;21 (2): 82–88.

32. Kanehisa M, et al. The KEGG resource for deciphering the genome. *Nucleic Acids Res* 2004;32(Database issue): D277–D280.

33. Karp PD, et al. The EcoCyc Database. *Nucleic Acids Res* 2002;30(1): 56–58.

34. Mewes HW, et al. MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* 2004;32(Database issue): D41–D44.

35. Christie KR, et al. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* 2004;32(Database issue): D311–D314.

36. Caspi R, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 2006;34(Database issue): D511–D516.

37. Karp PD, et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005;33(19): 6083–6089.

38. Harris MA, et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res* 2004;32(Database issue): D258–D261.

39. The Gene Ontology (GO) project in 2006 Nucleic Acids Res 2006;34(Database issue): D322–D326.

40. Serres MH, Goswami S, Riley M. GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res* 2004;32(Database issue): D300–D302.

41. Coulton G. Are histochemistry and cytochemistry 'Omics'? *J Mol Histol* 2004;35(6): 603–613.

42. Arita M, Robert M, Tomita M. All systems go: launching cell simulation fueled by integrated experimental biology data. *Curr Opin Biotechnol* 2005;16(3): 344–349.

43. Huh WK,et al. Global analysis of protein localization in budding yeast. *Nature* 2003;425 (6959): 686–691.

44. Guda C Subramaniam S TARGET: a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics* 2005.

45. Fields S. High-throughput two-hybrid analysis. The promise and the peril. *Febs J* 2005;272(21): 5391–5399.

46. Deeds EJ, Ashenberg O, Shakhnovich EI. A simple physical model for scaling in protein–protein interaction networks. *Proc Natl Acad Sci USA* 2006;103 (2): 311–316.

47. Sprinzak E, Sattath S, Margalit H. How reliable are experimental protein–protein interaction data? *J Mol Biol* 2003;327(5): 919–923.

48. Palsson B. Two-dimensional annotation of genomes. *Nat Biotechnol* 2004;22(10): 1218–1219.

49. Beard DA, Liang SD, Qian H. Energy balance for analysis of complex metabolic networks. *Biophys J* 2002;83(1): 79–86.

50. Covert MW, et al. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004;429(6987): 92–96.
51. Covert MW, Palsson BO. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J Theor Biol* 2003;221(3): 309–325.
52. Covert MW, Schilling CH, Palsson B. Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* 2001;213(1): 73–88.
53. Covert MW, Palsson BO. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J Biol Chem* 2002;277(31): 28058–28064.
54. Chvatal V. *Linear Programming*. New York: W.H. Freeman and Company, 1983.
55. 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.
56. 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.
57. Barrett CL, et al. The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states. *Proc Natl Acad Sci USA* 2005;102(52): 19103–19108.
58. Neidhardt FC, Ingraham JL, Schaechter M. *Physiology of the Bacterial Cell*. Sunderland, MA: Sinauer Associates, Inc., 1990.
59. Hillier FS, Lieberman GJ. *Introduction to Mathematical Programming*. New York: McGraw-Hill, xv, 1990,p. 649.
60. Williams HP, NetLibrary Inc. *Model Building in Mathematical Programming*, 4th ed. New York: Wiley, 1999,p. 368.
61. Reed JL, et al. An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol* 2003;4(9): R54.
62. Reed JL, Palsson BO. Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J Bacteriol* 2003;185(9): 2692–2699.
63. 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.
64. Schilling CH, et al. Genome-scale metabolic model of *Helicobacter pylori* 26695. *J Bacteriol* 2002;184(16): 4582–4593.
65. Thiele I, et al. An expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): an *in silico* genome-scale characterization of single and double deletion mutants. *J Bacteriol* 2005;187(16): 5818–5830.
66. 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.
67. Mahadevan R, 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.
68. Borodina I, Krabben P, Nielsen J. Genome-scale analysis of *Streptomyces coelicolor* A3 (2) metabolism. *Genome Res* 2005;15(6): 820–829.
69. Feist AM, et al. Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. 2006;2(1): msb4100046-E1–msb4100046-E14.

70. Forster J, et al. Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* 2003;13(2): 244–253.

71. Duarte ND, 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.

72. Kuepfer L, Sauer U, Blank LM. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res* 2005;15(10): 1421–1430.

73. Almaas E, Oltvai ZN, Barabasi AL. The activity reaction core and plasticity of metabolic networks. *PLoS Comput Biol* 2005;1(7): e68

74. Segre D, et al. Modular epistasis in yeast metabolism. *Nat Genet* 2005;37(1): 77–83.

75. 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.

76. Wiback SJ, Palsson BO. Extreme pathway analysis of human red blood cell metabolism. *Biophys J* 2002;83(2): 808–818.

77. Hucka M, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003;19 (4): 524–531.

78. Novere NL, et al. Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat Biotechnol* 2005;23(12): 1509–1515.

79. 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.

80. Blattner FR, et al. The complete genome sequence of *Escherichia coli* K-12. *Science* 1997;277(5331): 1453–1474.

81. Edwards JS, Palsson BO. Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J Biol Chem* 1999;274(25): 17410–17416.

82. Tong AH,et al. Global mapping of the yeast genetic interaction network. *Science* 2004;303(5659): 808–813.

83. Famili I, et al. *Saccharomyces cerevisiae* phenotypes can be predicted by using constraint-based analysis of a genome-scale reconstructed metabolic network. *Proc Natl Acad Sci USA* 2003;100(23): 13134–13139.

84. Forster J, et al. Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae*. *Omics* 2003;7(2): 193–202.

85. Edwards JS, Palsson BO. Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics* 2000;1(1): 1.

86. Giaever G, et al. Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 2002;418(6896): 387–391.

87. Hartwell L, Genetics. Robust interactions. *Science* 2004;303(5659): 774–775.

88. Baba T, et al. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. 2006;2 (1): pmsb4100050-E1–msb4100050-E11.

89. Glasner JD, et al. ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res* 2003;31(1): 147–151.

90. Salgado H, et al. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res* 2004;32(Database issue): D303–D306.

91. Covert MW, et al. Integrating high-throughput data and computational models leads to *E. coli* network elucidation. *Nature* 2004;429(6987): 92–96.
92. Palsson BO. *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, 2006.
93. Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA* 2002;99(23): 15112–15117.
94. Segre D, et al. From annotated genomes to metabolic flux models and kinetic parameter fitting. *Omics* 2003;7(3): 301–316.
95. Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* 2005;102(21): 7695–7700.
96. 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.
97. Schilling CH, Letscher D, Palsson BO. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol* 2000;203(3): 229–248.
98. Price ND, et al. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys J* 2003;84(2): 794–804.
99. Papin JA, Price ND, Palsson BO. Extreme pathway lengths and reaction participation in genome-scale metabolic networks. *Genome Res* 2002;12(12): 1889–1900.
100. Papin JA, et al. The genome-scale metabolic extreme pathway structure in *Haemophilus influenzae* shows significant network redundancy. *J Theor Biol* 2002;215(1): 67–82.
101. Bell SL, Palsson BO. Expa: a program for calculating extreme pathways in biochemical reaction networks. *Bioinformatics* 21(8): 2005;1739–1740.
102. Allen TE, Palsson BO. Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. *J Theor Biol* 2003;220(1): 1–18.
103. Papin JA, et al. Reconstruction of cellular signalling networks and analysis of their properties. *Nat Rev Mol Cell Biol* 2005;6(2): 99–111.
104. Varma A, Palsson BO. Metabolic capabilities of *Escherichia coli*: II. Optimal growth patterns. *Journal of Theoretical Biology* 1993;165(4): 503–522.
105. Papin JA, Palsson BO. Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J Theor Biol* 2004;227(2): 283–297.
106. Papin JA, Palsson BO. The JAK–STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys J* 2004;87(1): 37–46.
107. 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.
108. Herrgard MJ, et al. Integrated Analysis of Regulatory and Metabolic Networks Reveals Novel Regulatory Mechanisms in *Saccharomyces cerevisiae*. *Genome Res* 2006;16(5): 627–635.
109. Gianchandani EP, et al. Matrix formalism to describe functional states of transcriptional regulatory systems. *PLoS Comput Biol* 2006;2(8): e101.
110. von Hippel PH. Biochemistry. Completing the view of transcriptional regulation. *Science* 2004;305(5682): 350–352.

111. Price ND, et al. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. *Biophys J* 2003;84(2 Pt 1): 794–804.

112. Price ND, Papin JA, Palsson BO. Determination of redundancy and systems properties of the metabolic network of *Helicobacter pylori* using genome-scale extreme pathway analysis. *Genome Res* 2002;12(5): 760–769.

113. Fong SS, et al. *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol Bioeng* 2005;91(5): 643–648.

114. Oh YK, et al. Genome-scale reconstruction of metabolic network in *Bacillus subtilis* based on high-throughput phenotyping and gene essentiality data. *J Biol Chem* 2007;282 (39): 28791–28799.

115. Oliveira AP, Nielsen J, Forster J. Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol* 2005; 539.

116. Hong SH et al. The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat Biotechnol*, 2004;22(10): 1275–1281.

117. Taylor SW, et al. Characterization of the human heart mitochondrial proteome. *Nat Biotechnol* 2003;21(3): 281–286.