30

Predicting Gene Essentiality Using Genome-Scale *in Silico* Models

Andrew R. Joyce and Bernhard Ø. Palsson

Summary

Genome-scale metabolic models of organisms can be reconstructed using annotated genome sequence information, well-curated databases, and primary research literature. The metabolic reaction stoichiometry and other physicochemical factors are incorporated into the model, thus imposing constraints that represent restrictions on phenotypic behavior. Based on this premise, the theoretical capabilities of the metabolic network can be assessed by using a mathematical technique known as flux balance analysis (FBA). This modeling framework, also known as the constraint-based reconstruction and analysis approach, differs from other modeling strategies because it does not attempt to predict exact network behavior. Instead, this approach uses known constraints to separate the states that a system can achieve from those that it cannot. In recent years, this strategy has been employed to probe the metabolic capabilities of a number of organisms, to generate and test experimental hypotheses, and to predict accurately metabolic phenotypes and evolutionary outcomes. This chapter introduces the constraint-based modeling approach and focuses on its application to computationally predicting gene essentiality.

Key Words: computational modeling; constraint-based reconstruction and analysis; flux balance analysis (FBA); gene essentiality prediction; metabolic phenotype; systems biology.

1. Introduction

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 hundreds 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 further 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 biological

From: *Methods in Molecular Biology, vol. 416: Microbial Gene Essentiality*
Edited by: A. L. Osterman and S. Y. Gerdes © Humana Press Inc., Totowa, NJ

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

A prudent approach to gaining 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 implicitly limited by requiring knowledge of unknown parameters that are difficult or impossible to experimentally determine or approximate. 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 many of the aforementioned limitations. This approach, known as constraint-based reconstruction and analysis ***(11–15)***, has been employed to generate genome-scale models for organisms from all three major branches of the tree of life. Although bacterial models dominate this growing collection, a model from archaea has recently appeared, and several eukaryotic models are also available (*see* **Note 1** and **Table 1** for an overview of existing constraint-based metabolic models).

Among other uses (*see* **Note 2** and Ref. ***12***), these models have facilitated the computational investigation of gene essentiality. Flux balance analysis (FBA) ***(16, 17)*** is a powerful mathematical approach that uses optimization by linear programming 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 across the metabolic model that leads to this result. Accordingly, this methodology allows the investigator to assess wild-type growth capabilities of the modeled organism. Furthermore, metabolic gene knockout strains can be simulated simply by removing associated reaction(s) from the model. By comparing predicted growth rates before and after introducing the simulated gene deletion, the gene's essentiality can be assessed (i.e., growth will be zero if the removed gene is essential for biomass production). Given that this type of analysis relies on computer simulation, computational results must be confirmed by generating and studying the effects of gene knockouts at the lab bench. However, by first investigating these situations at the computer workstation, or *in silico,* 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 directly introduce the method and associated concepts. Furthermore, we discuss both the utility and potential shortcomings of these models in studying gene essentiality by reviewing results from several published studies. Finally, we briefly discuss additional interesting applications and some potential future directions for constraint-based modeling and analysis.

Table 1
Currently Available Constraint-Based Models

Organism	Total genes	Model genes	Model metabolites	Model reactions	Reference
Bacteria					
Bacillus subtilis	4,225	614	637	754	***(91)***
Escherichia coli	4,405	904	625	931	***(68)***
		720	438	627	***(55)***
Geobacter sulfurreducens	3,530	588	541	523	***(71)***
Haemophilus influenzae	1,775	296	343	488	***(56)***
		400	451	461	***(92)***
Helicobacter pylori	1,632	341	485	476	***(58)***
		291	340	388	***(57)***
Lactococcus lactis	2,310	358	422	621	***(93)***
Mannheimia succiniproducens	2,463	335	352	373	***(94)***
Staphylococcus aureus	2,702	619	571	641	***(70)***
Streptomyces coelicolor	8,042	700	500	700	***(72)***
Archaea					
Methanosarcina barkeri	5,072	692	558	619	***(59)***
Eukarya					
Mus musculus	28,287	1,156	872	1,220	***(76)***
Saccharomyces cerevisiae	6,183	750	646	1,149	***(62)***
		672	636	1,038	***(61)***
		708	584	1,175	***(73)***
Human cardiac mitochondria	615*	298	230	189	***(50)***
Human red blood cell	NA	NA	39	32	***(77)***

This table summarizes model statistics for the models developed and published to date. *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 ***(95)***. NA, not applicable.

2. Materials

1. Scientific literature and textbooks; for example, the PubMed database (www.pubmed.gov) and biochemical and organism-specific texts.
2. Online Genomic Databases and Resources (**Table 2**).
3. Software; for example, Microsoft Excel (office.microsoft.com), MATLAB (www.mathworks.com), Mathematica (www.wolfram.com), LINDO (www.lindo.com), GAMS (www.gams.com), and SimPheny (www.genomatica.com).

3. Methods

This section outlines the general procedure (**Fig. 1**) followed in constructing and using a constraint-based model in conjunction with FBA to computationally investigate gene essentiality. This model building and analysis procedure can be divided approximately into four successive steps:

Table 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
	The SEED	Database resource for genome annotations using the subsystem approach	http://www.theseed.org
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 for Community Analysis of Genomes (ASAP)	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* K-12 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.

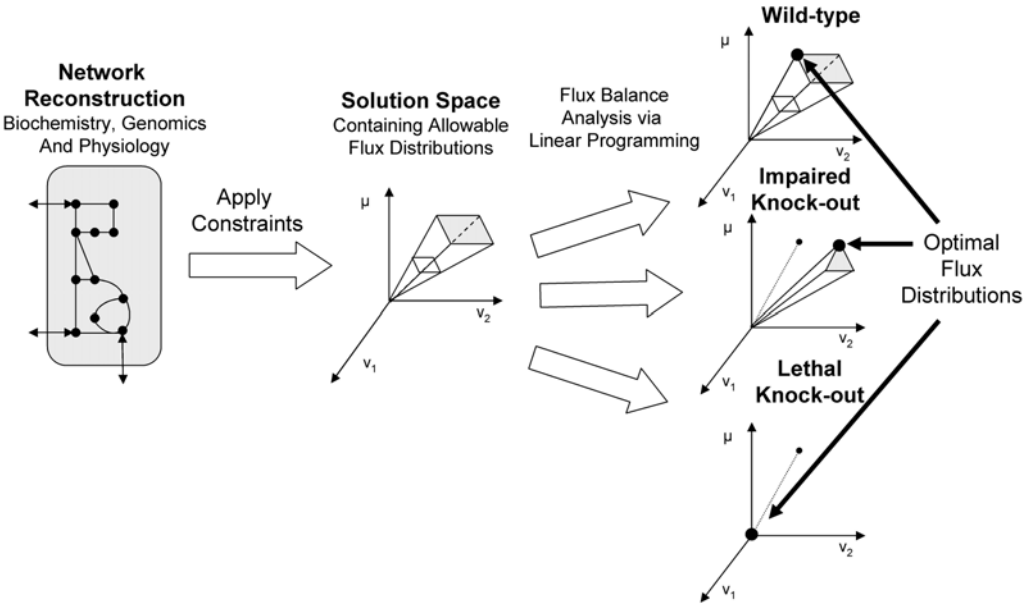


Fig. 1. Constraint-based modeling. Application of constraints to a reconstructed metabolic network leads to a defined solution space that specifies a cell's allowable metabolic phenotypes. Flux balance analysis (FBA) uses linear programming to find solutions in the space that maximize or minimize a given objective. In the graphical representation on the right, the optimal flux distributions that maximize μ, which represents growth/biomass production for the purposes of this chapter, are highlighted. 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. Impaired knockout strains are those that have a lower maximum value for the objective function than wild type, and lethal knockout strains are those that have a zero value for the objective function, indicating no growth capability when the strain harbors that particular gene deletion. As a reference, the wild-type flux distribution vector is also depicted by the dashed line on the impaired and lethal knockout plots.

1. Network reconstruction.
2. Stoichiometric (*S*) matrix compilation.
3. Identification and assignment of appropriate constraints to molecular components.
4. Assessment of gene essentiality via flux balance analysis (FBA).

In this section, each of the above components will be discussed in turn. In addition, a simple example will be provided in **Section 3.5** to illustrate directly the concepts described herein.

3.1. 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 con-

struction. We detail each of these components, known individually as data collection, metabolic reaction list generation, and gene-protein-reaction (GPR) relationship determination in this section.

3.1.1. Model Component Data Collection

Perhaps the most critical component of the constraint-based modeling approach involves the collection of data that is relevant to the system of interest. Not long ago, this was among the most challenging steps as researchers had access to very limited amounts of biochemical data. However, the success of recent genome sequencing ***(18)*** and annotation ***(19, 20)*** projects and advances in high-throughput technologies as well as the development of detailed and extensive online database resources has improved matters dramatically.

After identifying the system or organism of interest, relevant data sources must be identified to begin compiling the 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.

3.1.1.1. Biochemical Literature

Direct biochemical information found in the primary literature usually contains the best-quality data for use in reconstructing biochemical networks. Important details, such as precise reaction stoichiometry, in addition to its reversibility, are often directly available. Given that scrutinizing each study individually is an excessively time-consuming and tedious task, biochemical textbooks and review articles should be utilized when available and the primary literature used to resolve conflicts. Furthermore, many volumes devoted to individual organisms and organelles, such as *Escherichia coli* ***(21)*** and the mitochondria ***(22)***, are increasingly becoming available and are typically excellent resources.

3.1.1.2. 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 for hundreds of organisms has been determined ***(18)***. Furthermore, extensive bioinformatics-based annotation efforts ***(20)*** have made great strides toward identifying all coding regions contained within the sequence. For those biochemical reactions known to occur in the organism, but whose corresponding genes are unknown, sequence alignment tools such as BLAST and FASTA ***(23)*** can be utilized to assign putative functions based on similarity to orthologous genes and proteins of known function. The subsystem approach ***(19)*** is another strategy available to researchers looking for functional gene assignments. Rather than focusing on the annotation of individual genomes, the subsystem approach calls for the annotation of cellular pathways and processes across all sequenced organisms. The associated online resource known as SEED is becoming an increasingly useful tool in constraint-based model-building efforts. It should be emphasized, however, that putative assignments are hypothetical and subject to revision upon direct biochemical characterization. As one final note on genome annotation,

interesting efforts are also under way to automatically reconstruct networks based on annotated sequence information alone ***(24)***. However, these automated approaches are limited in that they can only be 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 ***(25, 26)***. 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 and listed in **Table 2**.

3.1.1.3. 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 pathway maps. Kyoto Encyclopedia of Genes and Genomes (KEGG) ***(27)*** is perhaps the most extensive and well-known among these resource types. Pathway maps for numerous metabolic processes are available through KEGG as is 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. For example, EcoCyc ***(28)*** incorporates gene and regulatory information as well as enzyme reaction pathways particular to *E. coli*. The Comprehensive Yeast Genome Database (CYGD) ***(29)*** and Saccharomyces Genome Database (SGD) ***(30)*** are other examples of *Saccharomyces cerevisiae*–specific comprehensive resources. Finally, the BioCyc resource ***(31, 32)*** contains automated annotation-derived pathway/genome databases for 250 individual organisms.

Additional important resources provide functional information for individual genes and gene products. These ontology-based tools strive to describe how gene products behave in a cellular context as they typically contain information regarding the function and localization of gene products within the cell. Perhaps the most well-known resource is Gene Ontology Consortium (GO) ***(33, 34)***, which contains ontological information for a variety of organisms. In recent years, organism-specific ontologies, such as GenProtEC ***(35)*** for *E. coli*, have also appeared. In sum, these online resources are valuable in that they typically incorporate information regarding individual genes and proteins as well as information regarding their regulation, cellular localization, and participation in enzymatic reactions into a single integrative resource.

3.1.2. 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. Each reaction in a metabolic network always 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 2A** is neither elementally nor charge balanced. However, inclusion of hydrogen in **Figure 2B** balances the reaction in both regards.

Biological boundaries also must be considered when defining reaction lists. Metabolic networks are composed of both intracellular and extracellular reactions. For example, in bacteria the reactions of glycolysis and the tricarboxylic acid cycle (TCA) take 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 into the cell. 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 that will assist in this process 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 ***(36–38)***. Furthermore, computational tools have also been developed to predict subcellular location of proteins in eukaryotes ***(39)***.

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 2**) are two useful resources that catalogue enzyme reversibility.

3.1.3. Determining GPR Relationships

Upon completing the reaction list, the protein or protein complexes that facilitate each metabolite substrate to product conversion must be determined. Each subunit of a protein complex must be assigned to the same reaction. Additionally, some reactions can be catalyzed by different enzymes. These so-called 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 3.1.1** and **Table 2** assist this process. In particular, KEGG and GO

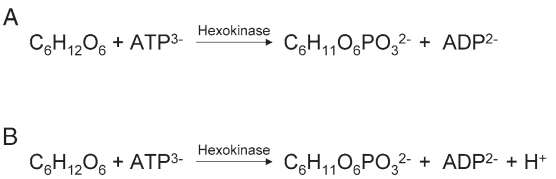


Fig. 2. Charge and elementally balanced reactions. **(A)** This depiction of the hexokinase-mediated conversion of glucose to glucose-6-phosphate is neither elementally nor charge balanced. **(B)** Inclusion of hydrogen both elementally and charge balances the reaction.

provide considerable enzyme-reaction information for a variety of organisms. Furthermore, protein-protein interaction data sets, derived from yeast two-hybrid experiments ***(40)***, for example, may be useful resources for defining enzymatic complexes in less-defined situations. One must take care in using these data, however, given their generally high false-positive rate and questionable reproducibility ***(41, 42)***.

3.2. Defining the Stoichiometric Matrix

The compiled reaction list can be represented mathematically in the form of a stoichiometric (*S*) matrix. The *S* matrix is formed from the stoichiometric coefficients of the reactions that participate in a 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 with a reaction, and every row corresponds with 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, 43)***.

Figure 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 individual proteins or protein complexes that catalyze the reactions. Element S_{ij} represents the coefficient of metabolite *i* in reaction *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 element is assigned a zero value.

3.3. Identifying and Applying Appropriate Constraints

Having developed a mathematical representation of a metabolic network, 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 defined in the previous section is a constraint in that it defines the mass and charge balance requirements for all possible metabolic reactions that are available to the cell. These stoichiometric constraints establish a geometric solution space (*see* **Fig. 1** for a graphical

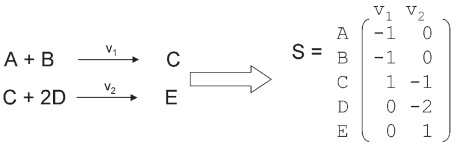


Fig. 3. Generating the stoichiometric (*S*) matrix. The reaction list on the left is mathematically represented by the *S* matrix on the right. As a convention, each row represents a metabolite, and each column represents a reaction in the network. Additionally, input or reactant metabolites have negative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

representation of the solution space concept) that 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_{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 ***(44)***, internal metabolic flux determinations ***(13)***, and transcriptional regulation ***(45–48)***.

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 (FBA) then can be used to examine the simulated knockout properties relative to wild type, as outlined in the next section.

3.4. Assessing Gene Essentiality via Flux Balance Analysis

Flux balance analysis (FBA) is a powerful computational method that relies on optimization by linear programming 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, FBA can 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.

3.4.1. Linear Programming

The solution space defined by constraint-based models can be explored via linear optimization by utilizing linear programming (LP). The LP problem corresponding with the optimal flux distribution determination through a metabolic network can be formulated as follows:

$$\begin{aligned} &\text{Maximize} && Z = \mathbf{c}^{\mathrm{T}} v \\ &\text{Subject to} && S \cdot v = 0 \\ & && \alpha_i \le v_i \le \beta_i \quad \text{for all reactions } i. \end{aligned}$$

In the above representation, Z represents the objective function, and $\mathbf{c}$ is a vector of weights on the fluxes $\mathbf{v}$. 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 by α_i and β_i, respectively.

The solution to this problem yields not only a maximum value for the objective function Z, but also results in an optimal flux distribution ($\mathbf{v}$) that allows the highest

Box 1: FBA using Matlab

Here we use Matlab to solve an FBA problem for 3 cases using the system in Figure 4. The **linprog()** function accepts six arguments and returns two values in the following form:

[v, Z] = linprog(c, Aeq, beq, S, b, α, β).

This solves the following LP problem:

Minimize $Z = c \cdot v$
Subject to $Aeq \cdot v \leq beq$
$S \cdot v = b$
$\alpha \leq v \leq \beta$

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

$Aeq \cdot v \leq beq$

is equivalent to

$v \leq \beta$.

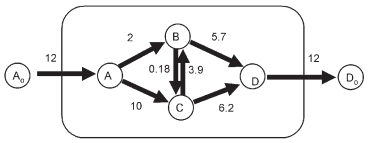
The code to solve the wild type problem (Case 1) of interest in Matlab's framework follows, using α and β as defined in the text :

```
>> 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 Matlab defaults to solving a minimization problem we use the negative of the optimization vector.

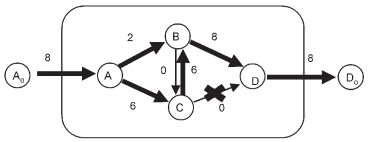
Case 1: Wild Type



Case 2 solves the same problem, but this time after knocking out reaction v5 by modifying the β vector:

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

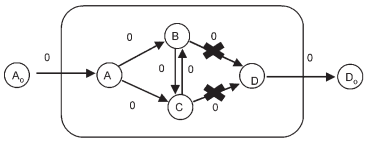
Case 2: v6 Knockout



Finally, Case 3 simulates a "lethal" deletion strain by knocking out both v5 and v6:

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

Case 3: v5 & v6 Double Knockout



flux through *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 with the gene(s) of interest, a simulated gene deletion strain may be created. The examination of simulation results from before and after introducing the simulated gene deletion leads directly to gene essentiality predictions.

Problems of this type can be readily formulated and solved by commercial software packages, such as MATLAB, Mathematica, LINDO, as well as tools available through the General Algebraic Modeling System (GAMS). **Section 3.5** and **Box 1** present simple, hypothetical examples that can be 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 exist for a given network and simulation conditions. This topic has been explored using mixed-integer linear programming (MILP) techniques with genome-scale metabolic models ***(49, 50)*** but is beyond the scope of this chapter and will not be further discussed.

3.4.2. Constraints

As previously stated, the *S* matrix constrains the system by defining all possible metabolic reactions. In mathematical terms, the stoichiometric (*S*) matrix is a linear transformation of the reaction flux vector,

$$\mathbf{v} = (v_1, v_2, \ldots, v_n)$$

to a vector of time derivatives of metabolic concentrations

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

such that

$$\frac{d\mathbf{x}}{dt} = S \cdot \mathbf{v}.$$

Therefore, a particular flux distribution **v** represents the flux levels through each reaction in the network. Because the time constants that describe metabolic transients are fast (of the order tens of seconds or less), whereas the time constants for cell growth are comparatively slow (of the order hours to days), the behavior of cellular components can be considered as existing in a quasi-steady state ***(51)***. This assumption leads to the reduction of the previous equation to:

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

By focusing only on the steady-state condition, assumptions or rough approximations 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 ***(52)***.

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 0. Similarly, if the enzyme capacity, or V_{max}, is experimentally defined, setting β to the known experimental value limits the allowable reaction flux through the enzyme within the model. In contrast, a gene knockout is simulated by setting $\beta_i = 0$ for gene i (**Section 3.5** and **Box 1**). If constraints on flux values through reaction v_i cannot be identified, then α_i and β_i are set to $-\infty$ 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 (*see* **Section 3.5** and **Box 1** for examples). Finally, if a flux is "known," for example, from detailed experimentation, α_i and β_i can be set to the same non-zero value to explicitly define the flux value associated with reaction v_i.

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 must also 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 represent 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, only those inputs that are 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 (*see* Ref. ***53*** for an example of a large-scale analysis of *E. coli* growth simulations performed using minimal media). 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.

3.4.3. The Objective Function

Given that multiple possible flux distributions exist for any given network, optimization can be used to identify a particular flux distribution 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 wild-type or simulated mutant 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 in the given environment, and further, is the reaction network able to support growth despite a simulated gene deletion?" 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* ***(54)***. In sum, with the choice of biomass as an objective function, cell growth, depicted as a non-zero 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. Accordingly, if the *in silico* knockout fails to exhibit simulated growth (i.e., $Z = 0$) (*see* **Fig. 1** for a graphical representation of this case), the associated gene is predicted to be essential.

3.5. A Simple FBA Example

In order to demonstrate the concepts previously introduced, this section presents a specific example using a simple system. **Figure 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 the corresponding internal metabolite. External metabolites need not be explicitly considered in the stoichiometric network representation, however.

Figure 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,

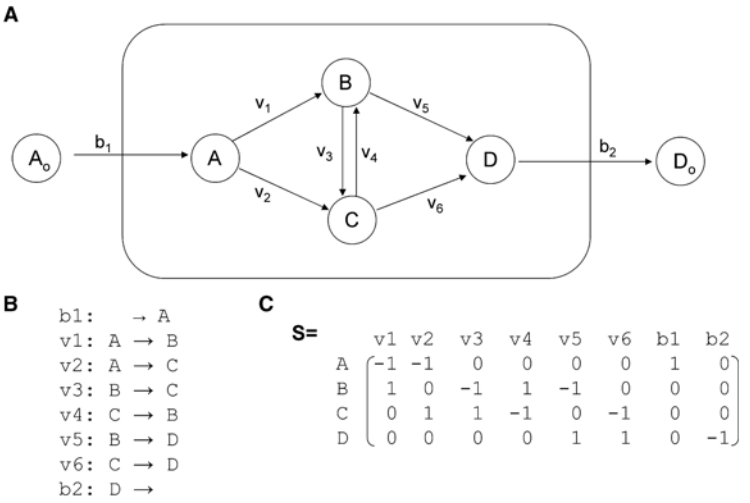


Fig. 4. An example system. **(A)** A four-metabolite, eight-reaction system is first decomposed into individual reactions in **(B)** and then represented mathematically in the *S* matrix depicted in **(C)**. By convention, internal reactions are denoted by v_i, and reactions that span the system boundary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

however, for simplicity the reaction is decoupled into two separate reactions with individual corresponding fluxes.

The *S* matrix for this system is detailed in **Figure 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:

$$\begin{aligned} &\text{Maximize} && Z = \mathbf{c}^{\mathrm{T}}\mathbf{v} \\ &\text{Subject to} && S \cdot \mathbf{v} = 0 \\ & && \alpha \le v_i \le \beta \quad \text{for all reactions } i. \end{aligned}$$

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

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

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

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

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

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

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. Accordingly, given the relatively low upper bounds on the internal fluxes, the actual values of these fluxes in the calculated optimal flux distribution will never approach these levels.

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 1**. It should be noted that the default MATLAB optimization solver is only suitable for problems of this and slightly larger magnitude. Typical biological problems that involve many more variables and constraints require more sophisticated optimization software such as the packages available through LINDO and GAMS (**Note 1**).

Having used the above information to simulate the wild-type case, the upper bound β vector is 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 0:

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

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

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

Previous studies utilized this general strategy to simulate gene knockouts in computational investigations of gene essentiality using genome-scale bacterial models (*see*, for example, *E. coli* ***[48, 55]***, *H. influenzae* ***[56]***, *H. pylori* ***[57, 58]***) as well as in the archaeal model of *M. barkeri* ***(59)*** and in the eukaryotic model of *S. cerevisiae* ***(60–62)*** (**Notes 3** and **4**).

4. Conclusion

Constraint-based modeling and its associated analyses are powerful tools that can be used to computationally predict gene essentiality with a high degree of success. This strategy aids researchers by identifying the most interesting knockouts that warrant future study, thus prioritizing experimental projects and saving considerable time. Beyond addressing the biological question associated with determining gene essentiality, this computational approach also has medical relevance. In pathogenic microbial models, each identified essential gene suggests a potential drug target that could be used to develop effective therapeutics in the future. Furthermore, progress is being made in applying this modeling framework to other aspects of the cell, such as in RNA and protein synthesis ***(63)***, cell signaling ***(64–66)***, and transcriptional regulatory networks ***(67)***. Because each of these network types are interrelated in terms of shared components and metabolites, these efforts are setting the stage for pushing the field a significant step forward toward generating integrated models of the entire cell (**Fig. 5**). As more genome-scale models are developed (**Note 1**), existing models enhanced (**Notes 4** and **5**), and different types of models integrated, additional applications for the constraint-based modeling approach will become apparent (**Note 2**). Consequently, the flexibility of the constraint-based modeling framework will continue to be exploited

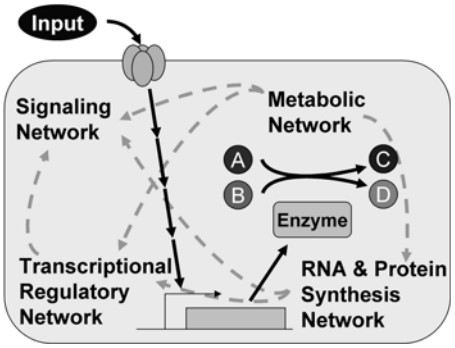


Fig. 5. The next big challenge: model integration. This chapter has illustrated the utility of constraint-based modeling and analysis in computationally assessing gene essentiality for metabolism. The constraint-based approach has been applied to other systems as well. To date, however, these models 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 stimulates 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 system biology's foundational goals: to computationally represent and analyze models of entire cells and biological systems.

to aid in the prediction of gene essentiality and drive the exploration of countless other exciting biological questions.

Notes

1. 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 previously noted, with large-scale models it may be necessary to use a robust computational platform designed specifically for 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 *E. coli* model contains 904 genes and 931 unique biochemical reactions ***(68)***. Building a genome-scale model within the framework proposed in **Section 3** is possible using ubiquitous spreadsheet software such as Excel (Microsoft, Redmond, WA), but this effort would likely be slow, unwieldy, 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 support, 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 1**) for further research and analysis. Most of these models represent bacteria and range from the important model organism *E. coli* ***(55, 68, 69)*** to pathogenic microbes such as *H. pylori* ***(57, 58)*** and *S. aureus* ***(70)***. Furthermore, recently developed models of *G. sulfurreducens* ***(71)*** and *S. coelicolor* ***(72)*** may become 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* ***(59)***, 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* **(61, 62, 73)** 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. ***74*** and ***75***). Metabolic models of higher-order systems are also becoming available, such as a model of mouse (*Mus musculus* **[76]**), as well as human cardiac mitochondria ***(50)*** and the human red blood cell ***(77)***.

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, the Systems Biology Markup Language (SBML) ***(78)***, for example, 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 ***(79)*** 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.

2. A rapidly growing collection of analytical methods have been developed for use in conjunction with constraint-based models (reviewed in Ref. ***12***), some of which we briefly introduce in this section. Although the focus of this chapter is 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 ***(86, 87)***. Experimental data seem to confirm the MOMA assumption that knockout strains utilize the metabolic network similar to wild type ***(86)***. It remains to be determined if this is true in all situations or if the network optimizes for growth over time after gene deletion.

   A more recently developed method known as regulatory on/off minimization (ROOM) ***(88)*** is another constraint-based analysis technique that uses a mixed-integer linear programming (MILP) strategy to predict the metabolic state of an organism after 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. Whereas MOMA seems to better predict the initial metabolic adjustment that occurs after 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 ***(89)***. 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 ***(90)*** and has the potential for significant impact.

3. Many studies have used genome-scale constraint-based models to assess gene essentiality, in particular using models of *E. coli* ***(48, 55)***, *H. influenzae* ***(56)***, *H. pylori* ***(57)***, *M. barkeri* ***(59)***, and *S. cerevisiae* ***(60, 62)*** under various growth conditions. Each study simulated gene deletions by constraining the flux through the associated reaction(s) to zero, as described in **Section 3.4.2** and **Box 1**. Relatively few central metabolic genes are predicted to be lethal, as shown in **Table 3**. This observation likely reflects the inherent redundancy and high degree of interconnectivity that is characteristic of central metabolism. In addition, *H. influenzae* seems to be less robust than *E. coli* against single-gene deletions as a higher percentage of central metabolic genes are predicted to be essential. Furthermore, given that these networks appear generally robust against single-gene deletions, perhaps future studies should focus on lethal double mutants, known as synthetic lethal mutants, which are commonly studied in

Table 3
Computationally Predicted Gene Essentiality

Organism	No growth	Impaired growth
E. coli ***(49, 55)***	*rpiAB, pgk, acnAB, gltA, icdA, tktAB, gapAC*	*atp, fba, pfkAB, tpiA, eno, gpmAB, nuo, ackAB, pta*
H. influenzae ***(56)***	*eno, fba, fbp, pts, gapA, gpmA, pgi, pgk, ppc, prsA, rpiA, tktA, tpiA*	*cudABCD, atp, ndh, ackA, pta, gnd, pgl, zwf, talB, rpe*
H. pylori ***(57)***	*aceB, ppa, prsA, tpi, tktA, eno*, pgi*, pgk*, gap*, pgm*, ppaA*, rpe*, rpi*, fba**	
M. barkeri ***(59)***	*ackA*, pta*, cdhABCDE*, cooS*, fmdABCDEF*, fwdBDEG*, ftr*, mch*, mtd*, mer*, mtrABCDEFGH*, mtaABC*, mcrABG*, hdrABCDE*, fpoABCDFHIJKLMNO*, frhABDG*, echABCDEF*, ahaABCDEFHIK**	
S. cerevisiae ***(60, 62)***	*ERG13, ACS2, ERG10, IPP1, CDS1, PSA1, TRR1, GUK1, PMI40, SAH1, SEC53, ERG26, OLE1, ERG25, ERG1, ERG11, ERG7, ERG9, ERG20, FAS1, ERG27, ERG12, ERG8, ACC1, MVD1, IDI1, FAS2, PIS1, DPM1*	*ATP16, RKI1, ILV3, ILV5, PGI1, TPI1, FBA1, PGK1*

This table summarizes some results from studies that used constraint-based metabolic models to predict gene essentiality. The "No growth" column lists the gene-deletion strains that had a simulated lethal phenotype (i.e., $Z = 0$). The "Impaired growth" column lists gene-deletion strains whose simulated phenotype was less than the wild-type strain, but not lethal (i.e., $Z_{\text{wild-type}} > Z_{\text{deletion-strain}}$).

*These genes are essential under some, but not all, tested environmental conditions.

S. cerevisiae ***(80, 81)***. Results from such studies are beginning to appear ***(58, 61)*** and may provide additional insight into gene and reaction essentiality as well as metabolic network robustness.

4. 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 ***(82)*** and online databases, including ASAP ***(83)*** for *E. coli* as well as CYGD and SGD for *S. cerevisiae*. Experimental growth phenotype data are available to assess directly the predictive power of the model for four of the five organisms listed previously and shows that correct predictions were made in ~60%, 86%, 83%, and 92% of cases for *H. pylori* ***(57)***, *E. coli* ***(48)***, *S. cerevisiae* ***(62)***, and *M. barkeri* ***(59)***, 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. Note 5 details how to interpret incorrect model predictions and their likely causes.
5. In the studies discussed in **Note 3** and **Note 4**, the model predictions, when compared with 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 due to 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 latter case was shown to account for many false predictions when using a yeast metabolic model to account for strain lethality ***(61)*** as 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 current genome-scale metabolic models typically assign equal ability to each isozyme. If this in fact is the case, 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 deletion of its dominant counterpart.

   An additional major error source stems from the lack of regulatory information incorporated into the previously described models. A Boolean logic approach has been used to include transcription factor–metabolic gene interactions and enhance the accuracy of constraint-based model predictions ***(48)*** and in genome-scale models of *E. coli* ***(45)*** and yeast ***(84)***. Regulatory information is available in the primary literature in addition to online resources such as EcoCyc and RegulonDB ***(85)***. Furthermore, these interactions can be derived from ChIP-chip analysis of transcription factors and corresponding gene expression microarray data ***(45)***.

   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 predictions often can be attributed to a paucity of information and not simply a technique failure, thus validating the general strategy of constraint-based modeling.

References

1. Wheeler, D. L., Church, D. M., Edgar, R., Federhen, S., Helmberg, W., Madden, T. L., et al. (2004) Database resources of the National Center for Biotechnology Information: update. *Nucleic Acids Res.* **32** (Database issue), D35–40.
2. Wyrick, J. J., and Young, R. A. (2002) Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev.* **12**, 130–136.
3. Sanford, K., Soucaille, P., Whited, G., and Chotani, G. (2002) Genomics to fluxomics and physiomics—pathway engineering. *Curr. Opin. Microbiol.* **5**, 318–322.
4. Joyce, A. R., and Palsson, B. O. (2006) The model organism as a system: integrating "omics" data sets. *Nat. Rev. Mol. Cell. Biol.* **7**, 198–210.
5. Arkin, A. P. (2001) Synthetic cell biology. *Curr. Opin. Biotechnol.* **12**, 638–644.
6. Tomita, M., Hashimoto, K., Takahashi, K., Shimizu, T. S., Matsuzaki, Y., Miyoshi, F., et al. (1999) E-CELL: software environment for whole-cell simulation. *Bioinformatics* **15**, 72–84.
7. Hoffmann, A., Levchenko, A., Scott, M. L., and Baltimore, D. (2002) The IkappaB-NF-kappaB signaling module: temporal control and selective gene activation. *Science* **298**, 1241–1245.
8. Elowitz, M. B., Levine, A. J., Siggia, E. D., and Swain, P. S. (2002) Stochastic gene expression in a single cell. *Science* **297**, 1183–1186.
9. Arkin, A., Ross, J., and McAdams, H. H. (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* **149**, 1633–1648.
10. Sarkar, A., and Franza, B. R. (2004) A logical analysis of the process of T cell activation: different consequences depending on the state of CD28 engagement. *J. Theor. Biol.* **226**, 455–466.
11. Reed, J. L., Famili, I., Thiele, I., and Palsson, B. O. (2006) Towards multidimensional genome annotation. *Nat. Rev. Genet.* **7**, 130–141.
12. Price, N. D., Reed, J. L., and Palsson, B. O. (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat. Rev. Microbiol.* **2**, 886–897.
13. Edwards, J. S., Covert, M., and Palsson, B. (2002) Metabolic modelling of microbes: the flux-balance approach. *Environ. Microbiol.* **4**, 133–140.
14. Covert, M. W., Famili, I., and Palsson, B. O. (2003) Identifying constraints that govern cell behavior: a key to converting conceptual to computational models in biology? *Biotechnol. Bioeng.* **84**, 763–772.
15. Price, N. D., Papin, J. A., Schilling, C. H., and Palsson, B. O. (2003) Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol.* **21**, 162–169.
16. Varma, A., and Palsson, B. O. (1994) Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. *Appl. Environ. Microbiol.* **60**, 3724–3731.
17. Kauffman, K. J., Prakash, P., and Edwards, J. S. (2003) Advances in flux balance analysis. *Curr. Opin. Biotechnol.* **14**, 491–496.
18. Liolios, K., Tavernarakis, N., Hugenholtz, P., and Kyrpides, N. C. (2006) The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide. *Nucleic Acids Res.* **34**, D332–334.
19. Overbeek, R., Begley, T., Butler, R. M., Choudhuri, J. V., Chuang, H. Y., Cohoon, M., et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702.

20. Brent, M. R. (2005) Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res.* **15**, 1777–1786.
21. Neidhardt, F. C., and Curtiss, R. (1996) *Escherichia coli and Salmonella: cellular and molecular biology*, 2nd ed. Washington, DC: ASM Press.
22. Scheffler, I. E. (1999) *Mitochondria*. New York: Wiley-Liss.
23. Chen, Z. (2003) Assessing sequence comparison methods with the average precision criterion. *Bioinformatics* **19**, 2456–2460.
24. Karp, P. D., Paley, S., and Romero, P. (2002) The Pathway Tools software. *Bioinformatics* **18** (Suppl 1), S225–232.
25. Cash, P. (2003) Proteomics of bacterial pathogens. *Adv. Biochem. Eng. Biotechnol.* **83**, 93–115.
26. Taylor, S. W., Fahy, E., and Ghosh, S. S. (2003) Global organellar proteomics. *Trends Biotechnol.* **21**, 82–88.
27. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res.* **32** (Database issue), D277–280.
28. Karp, P. D., Riley, M., Saier, M., Paulsen, I. T., Collado-Vides, J., Paley, S. M., et al. (2002) The EcoCyc Database. *Nucleic Acids Res.* **30**, 56–58.
29. Mewes, H. W., Amid, C., Arnold, R., Frishman, D., Guldener, U., Mannhaupt, G., et al. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res* **32** (Database issue), D41–44.
30. Christie, K. R., Weng, S., Balakrishnan, R., Costanzo, M. C., Dolinski, K., Dwight, S. S., et al. (2004) Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from *Saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Res* **32** (Database issue), D311–314.
31. Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., et al. (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* **34**, D511–516.
32. Karp, P. D., Ouzounis, C. A., Moore-Kochlacs, C., Goldovsky, L., Kaipa, P., Ahren, D., et al. (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res.* **33**, 6083–6089.
33. Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., et al. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **32** (Database issue), D258–261.
34. Gene Ontology Consortium (2006) The Gene Ontology (GO) project in 2006. *Nucleic Acids Res.* **34**, D322–326.
35. Serres, M. H., Goswami, S., and Riley, M. (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res.* **32** (Database issue), D300–302.
36. Coulton, G. (2004) Are histochemistry and cytochemistry "Omics"? *J. Mol. Histol.* **35**, 603–613.
37. Arita, M., Robert, M., and Tomita, M. (2005) All systems go: launching cell simulation fueled by integrated experimental biology data. *Curr. Opin. Biotechnol.* **16**, 344–349.
38. Huh, W. K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003) Global analysis of protein localization in budding yeast. *Nature* **425**, 686–691.
39. Guda, C., and Subramaniam, S. (2005) pTARGET [corrected] a new method for predicting protein subcellular localization in eukaryotes. *Bioinformatics* **21**, 3963–3969.

40. Fields, S. (2005) High-throughput two-hybrid analysis. The promise and the peril. *FEBS J.* **272**, 5391–5399.
41. Deeds, E. J., Ashenberg, O., and Shakhnovich, E. I. (2006) A simple physical model for scaling in protein-protein interaction networks. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 311–316.
42. Sprinzak, E., Sattath, S., and Margalit, H. (2003) How reliable are experimental protein-protein interaction data? *J. Mol. Biol.* **327**, 919–923.
43. Palsson, B. (2004) Two-dimensional annotation of genomes. *Nat. Biotechnol.* **22**, 1218–1219.
44. Beard, D. A., Liang, S. D., and Qian, H. (2002) Energy balance for analysis of complex metabolic networks. *Biophys. J.* **83**, 79–86.
45. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. O. (2004) Integrating high-throughput and computational data elucidates bacterial networks. *Nature* **429**, 92–96.
46. Covert, M. W., and Palsson, B. O. (2003) Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J. Theor. Biol.* **221**, 309–325.
47. Covert, M. W., Schilling, C. H., and Palsson, B. (2001) Regulation of gene expression in flux balance models of metabolism. *J. Theor. Biol.* **213**, 73–88.
48. Covert, M. W., and Palsson, B. O. (2002) Transcriptional regulation in constraints-based metabolic models of *Escherichia coli. J. Biol. Chem.* **277**, 28058–28064.
49. Reed, J. L., and Palsson, B. O. (2004) Genome-scale in silico models of *E. coli* have multiple equivalent phenotypic states: assessment of correlated reaction subsets that comprise network states. *Genome Res.* **14**, 1797–1805.
50. Vo, T. D., Greenberg, H. J., and Palsson, B. O. (2004) Reconstruction and functional characterization of the human mitochondrial metabolic network based on proteomic and biochemical data. *J. Biol. Chem.* **279**, 39532–39540.
51. Palsson, B. O. (2006) *Systems Biology: Properties of Reconstructed Networks*. New York: Cambridge University Press.
52. Schilling, C. H., Letscher, D., and Palsson, B. O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J. Theor. Biol.* **203**, 229–248.
53. Barrett, C. L., Herring, C. D., Reed, J. L., and Palsson, B. O. (2005) The global transcriptional regulatory network for metabolism in *Escherichia coli* exhibits few dominant functional states. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 19103–19108.
54. Neidhardt, F. C., Ingraham, J. L., and Schaechter, M. (1990) *Physiology of the Bacterial Cell*. Sunderland, MA: Sinauer Associates, Inc.
55. Edwards, J. S., and Palsson, B. O. (2000) The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5528–5533.
56. Schilling, C. H., and Palsson, B. O. (2000) Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. *J. Theor. Biol.* **203**, 249–283.
57. Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. (2002) Genome-scale metabolic model of *Helicobacter pylori* 26695. *J. Bacteriol.* **184**, 4582–4593.
58. Thiele, I., Vo, T. D., Price, N. D., and Palsson, B. (2005) An Expanded Metabolic Reconstruction of *Helicobacter pylori* (*i*IT341 GSM/GPR): An *in silico* genome-scale characterization of single and double deletion mutants. *J. Bacteriol.* **187**, 5818–5830.

59. Feist, A. M., Scholten, J. C. M., Palsson, B. O., Brockman, F. J., and Ideker, T. (2006) Modeling methanogenesis with a genome-scale metabolic reconstruction of *Methanosarcina barkeri*. *Mol. Syst. Biol.* **2**, msb4100046-E4100041-msb4100046-E4100014.
60. Forster, J., Famili, I., Palsson, B. O., and Nielsen, J. (2003) Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae*. *Omics* **7**, 193–202.
61. Kuepfer, L., Sauer, U., and Blank, L. M. (2005) Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res.* **15**, 1421–1430.
62. Duarte, N. C., Herrgard, M. J., and Palsson, B. O. (2004) Reconstruction and validation of *Saccharomyces cerevisiae* iND750, a fully compartmentalized genome-scale metabolic model. *Genome Res.* **14**, 1298–1309.
63. Allen, T. E., and Palsson, B. O. (2003) Sequence-based analysis of metabolic demands for protein synthesis in prokaryotes. *J. Theor. Biol.* **220**, 1–18.
64. Papin, J. A., and Palsson, B. O. (2004) Topological analysis of mass-balanced signaling networks: a framework to obtain network properties including crosstalk. *J. Theor. Biol.* **227**, 283–297.
65. Papin, J. A., Hunter, T., Palsson, B. O., and Subramaniam, S. (2005) Reconstruction of cellular signalling networks and analysis of their properties. *Nat. Rev. Mol. Cell. Biol.* **6**, 99–111.
66. Papin, J. A., and Palsson, B. O. (2004) The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys. J.* **87**, 37–46.
67. Gianchandani, E. P., Papin, J. A., Price, N. D., Joyce, A. R., and Palsson, B. O. (2006) Matrix formalism to describe functional States of transcriptional regulatory systems. *PLoS Comput. Biol.* **2**, e101.
68. Reed, J. L., Vo, T. D., Schilling, C. H., and Palsson, B. O. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Biol.* **4**, R54.
69. Reed, J. L., and Palsson, B. O. (2003) Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J. Bacteriol.* **185**, 2692–2699.
70. Becker, S. A., and Palsson, B. O. (2005) Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiol.* **5**, 8.
71. Mahadevan, R., Bond, D. R., Butler, J. E., Esteve-Nunez, A., Coppi, M. V., Palsson, B. O., Schilling, C. H., and Lovley, D. R. (2006) Characterization of metabolism in the Fe(III)-reducing organism *Geobacter sulfurreducens* by constraint-based modeling. *Appl. Environ. Microbiol.* **72**, 1558–1568.
72. Borodina, I., Krabben, P., and Nielsen, J. (2005) Genome-scale analysis of *Streptomyces coelicolor* A3(2) metabolism. *Genome Res.* **15**, 820–829.
73. Forster, J., Famili, I., Fu, P., Palsson, B. O., and Nielsen, J. (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res.* **13**, 244–253.
74. Almaas, E., Oltvai, Z. N., and Barabasi, A. L. (2005) The Activity Reaction Core and Plasticity of Metabolic Networks. *PLoS Comput. Biol.* **1**, e68.
75. Segre, D., DeLuna, A., Church, G. M., and Kishnoy, R. (2005) Modular epistasis in yeast metabolism. *Nat. Genet.* **37**, 77–83.
76. Sheikh, K., Forster, J., and Nielsen, L. K. (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of *Mus musculus*. *Biotechnol. Prog.* **21**, 112–121.
77. Wiback, S. J., and Palsson, B. O. (2002) Extreme pathway analysis of human red blood cell metabolism. *Biophys. J.* **83**, 808–818.

78. Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., et al. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* **19**, 524–531.
79. Novere, N. L., Finney, A., Hucka, M., Bhalla, U. S., Campagne, F., Collado-Vides, J., et al. (2005) Minimum information requested in the annotation of biochemical models (MIRIAM). *Nat. Biotechnol.* **23**, 1509–1515.
80. Hartwell, L. (2004) Genetics. Robust interactions. *Science* **303**, 774–775.
81. Tong, A. H., Lesage, G., Bader, G. D., Ding, H., Xu, H., Xin, X., et al. (2004) Global mapping of the yeast genetic interaction network. *Science* **303**, 808–813.
82. Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., et al. (2006) Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* **2**, msb4100050-E4100051-msb4100050-E4100011.
83. Glasner, J. D., Liss, P., Plunkett, G. 3rd, Darling, A., Prasad, T., Rusch, M., et al. (2003) ASAP, a systematic annotation package for community analysis of genomes. *Nucleic Acids Res.* **31**, 147–151.
84. Herrgard, M. J., Lee, B. S., Portnoy, V., and Palsson, B. O. (2006) Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae. Genome Res.* **16**, 627–635.
85. Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., et al. (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.* **32** (Database issue), D303–306.
86. Segre, D., Vitkup, D., and Church, G. M. (2002) Analysis of optimality in natural and perturbed metabolic networks. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15112–15117.
87. Segre, D., Zucker, J., Katz, J., Lin, X., D'Haeseleer, P., Rindone, W. P., et al. (2003) From annotated genomes to metabolic flux models and kinetic parameter fitting. *Omics* **7**, 301–316.
88. Shlomi, T., Berkman, O., and Ruppin, E. (2005) Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7695–7700.
89. Burgard, A. P., Pharkya, P., and Maranas, C. D. (2003) Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol. Bioeng.* **84**, 647–657.
90. Fong, S. S., Burgard, A. P., Herring, C. D., Knight, E. M., Blattner, F. R., Maranas, C. D., and Palsson, B. O. (2005) *In silico* design and adaptive evolution of *Escherichia coli* for production of lactic acid. *Biotechnol. Bioeng.* **91**, 643–648.
91. Oh, Y.K., Palsson, B.O., Park, S.M., Schilling, C.M., and Mahadevon, R. (2007) Genome-scale reconstruction of metabolic network in Bacillus subtilis based on high-throughput phenotyping and gene essentiality data. *J. Biol. Chem.*, in press.
92. Edwards, J. S., and Palsson, B. O. (1999) Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. *J. Biol. Chem.* **274**, 17410–17416.
93. Oliveira, A. P., Nielsen, J., and Forster, J. (2005) Modeling *Lactococcus lactis* using a genome-scale flux model. *BMC Microbiol.* **5**, 39.
94. Hong, S. H., Kim, J. S., Lee, S. Y., In, Y. H., Choi, S. S., Rih, J. K., et al. (2004) The genome sequence of the capnophilic rumen bacterium *Mannheimia succiniciproducens*. *Nat. Biotechnol.* **22**, 1275–1281.
95. Taylor, S. W., Fahy, E., Zhang, B., Glenn, G. M., Warnock, D. E., Wiley, S., et al. (2003) Characterization of the human heart mitochondrial proteome. *Nat. Biotechnol.* **21**, 281–286.