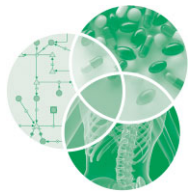
Advanced Review



Genome-scale metabolic networks

Marco Terzer[1] Nathaniel D. Maynard[2] Markus W. Covert[2] and Jörg Stelling[1]*

During the last decade, models have been developed to characterize cellular metabolism at the level of an entire metabolic network. The main concept that underlies whole-network metabolic modeling is the identification and mathematical definition of constraints. Here, we review large-scale metabolic network modeling, in particular, stoichiometric- and constraint-based approaches. Although many such models have been reconstructed, few networks have been extensively validated and tested experimentally, and we focus on these. We describe how metabolic networks can be represented using stoichiometric matrices and well-defined constraints on metabolic fluxes. We then discuss relatively successful approaches, including flux balance analysis (FBA), pathway analysis, and common extensions or modifications to these approaches. Finally, we describe techniques for integrating these approaches with models of other biological processes. *WIREs Syst Biol Med*

LARGE-SCALE NETWORK ANALYSIS

Metabolic networks are well-enough characterized that it is now possible to construct and analyze mathematical models of their behavior at a whole-genome level.[1,2] This is not due to rich, extensive datasets—in fact, the existing data are far from comprehensive.[3] Even in the best-understood organisms, the majority of kinetic parameters remain undetermined. Instead, whole-network modeling of metabolism has largely been enabled by the development of new computational methods that are able to make compelling and testable predictions even without many parameters. As a result, metabolic modeling has become a bellwether, leading the way toward a fundamental goal in biology: a computational model of an entire cell.

The main concept that underlies whole-network metabolic modeling is the identification and mathematical definition of constraints. These constraints then separate feasible and infeasible metabolic behaviors. Importantly, the constraints are often much easier to identify than kinetic parameters, making large-scale model building possible.

Most analysis techniques in this area employ up to three types of constraints.[1] *Physico-chemical constraints* are defined by conservation laws for mass and energy, dependency of reaction rates on metabolite concentrations, and negative free energy change for spontaneous reactions. *Environmental constraints* are imposed as a result of specific conditions, such as the availability of nutrients or electron acceptors. Finally, the effects of gene expression may result in *regulatory constraints* as the cell adapts to environmental changes.

Here, we review large-scale metabolic network modeling, focusing on *stoichiometric-* and *constraint-based* approaches. After discussing how these approaches work, we will highlight common and relatively successful approaches as well as existing models. Finally, we will comment on the possibility of integrating these models with models of other biological processes, such as transcriptional regulation and signal transduction, as another step toward true whole-cell modeling.

MODEL DEVELOPMENT

The Stoichiometric Matrix

Although there are a variety of metabolic modeling approaches, all of them share a fundamental requirement: a stoichiometric matrix based on a reconstructed metabolic network. Each column of the stoichiometric matrix corresponds to a chemical or

*Correspondence to: joerg.stelling@bsse.ethz.ch

[1]Department of Biosystems Science and Engineering, ETH Zurich, Switzerland

[2]Department of Bioengineering, Stanford University, Stanford, CA, USA

DOI: 10.1002/wsbm.037

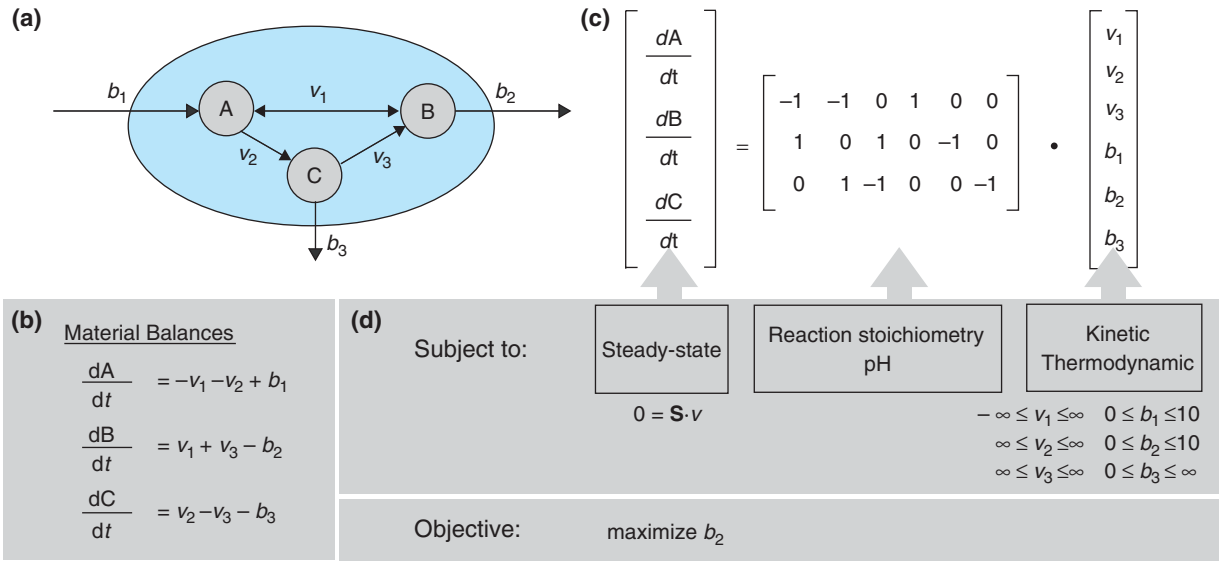


FIGURE 1 | (a) A small reaction network consisting of three metabolites (A, B, and C), three transport reactions, and three enzymatic reactions is constructed. v_i indicates the flux through reaction i and b_j represents the flux through transport protein j. (b) Material balance equations are shown for each metabolite. (c) A stoichiometric matrix is populated according to Eq. 1. (d) Assumptions, constraints, and an objective are listed for the system.

transport reaction, with non-zero values that identify the metabolites which participate in the reaction as well as the stoichiometric coefficients that correspond to each metabolite (Figure 1). The matrix also contains directionality: substrate and product metabolites in the matrix have negative and positive coefficients, respectively. By considering the matrix rows instead of the columns, the stoichiometric matrix can also be thought of as the list of reactions in which a given metabolite participates. This interpretation is useful when defining mass balances for each metabolite in the network.

These mass balances are expressed by a system of differential equations written for all the metabolite concentrations **c** as follows:

$$\frac{\mathrm{d}\mathbf{c}(t)}{\mathrm{d}t} = \mathbf{S} \cdot \mathbf{v}(t) \qquad (1)$$

where **S** is the stoichiometric matrix and $\mathbf{v}(t)$ the vector of reaction rates. Note that metabolism operates on a much faster time-scale than regulatory or cell division events. It is thus often reasonable to assume that metabolic dynamics have reached a *quasi-* or *pseudo-steady state*, where metabolite concentrations do not change. This leads to the *metabolite balancing equation*

$$\mathbf{S} \cdot \mathbf{v}(t) = 0 \qquad (2)$$

Equation (2) is a homogeneous system of linear equations. It requires that each metabolite is consumed in the same quantity as it is produced, and is the basis for further analysis of metabolic fluxes based on the stoichiometric matrix.

The process of building stoichiometric matrices has been amply described and reviewed elsewhere.[4] Briefly, this process involves gathering a variety of genomic, biochemical, and physiological data from the primary literature as well as databases, such as Uniprot,[5] BRENDA,[6] BioCyc,[7] KEGG,[8] and the Enzyme Commission database.[9] This information is used to synthesize a list of chemical and transport reactions together with their metabolite participants for a given cell. In addition, the chemical formula and charge of each metabolite should be inspected to verify that the chemical reaction is balanced. A reconstructed network model is only as good as its corresponding stoichiometric matrix, and the amount and quality of experimental evidence supporting the inclusion of a reaction in the matrix can vary significantly. Therefore, careful curation and continual updates to the matrix are critical.

Constraints on Reaction Rates

The stoichiometric matrix can be annotated by including further important information about either the reactions or the metabolites. The most common matrix annotations include the reversibility of each reaction and the cellular compartment in which

each reaction occurs. More generally, reaction rates are bounded as a consequence of kinetic constants, measured or estimated concentration ratios or to reflect the experimental setup. Upper and lower limits can apply to fluxes of individual reactions ($v_{min} \leq v \leq v_{max}$), and reaction directions can be defined by simply setting $v_{min} = 0$ or $v_{max} = 0$ for forward or backward irreversible reactions, respectively. Additional matrix annotation might include more detailed information about reaction kinetics. For example, it is possible to calculate metabolite uptake or secretion rates without difficulty in several cases; these can be used as a part of the metabolic model. In addition, ^{13}C chase experiments measure concentrations of internal metabolites, from which enzymatic fluxes can often be inferred.[10,11] At the present such 'fluxomic' approaches can only be applied to highly reduced metabolic networks, although substantial progress has been made in recent years.

Links to Genomic Information

The stoichiometric matrix can also be linked explicitly to the genome and gene expression data for use in certain applications. First, one can connect each chemical or transport reaction to the proteins and genes which enable it to occur. These relationships are often complicated, because proteins are frequently made up of multiple subunits, multiple enzymes sometimes catalyze the same chemical reaction, and certain enzymes catalyze more than one reaction. Delineating these gene–protein–reaction relationships enables the comparison of computational network analyses with experimentally determined gene knockout phenotypes. Gene expression data and transcription factor–gene relationships can also be critical to understanding more complex metabolic behaviors, as described below.

Existing Genome-Scale Models

To date, dozens of large-scale metabolic reconstructions have been published.[12] Most reconstructions are of human pathogens, model organisms, or organisms used in the biotechnology industry to produce valuable chemical products. For the purposes of this review, we will briefly highlight three of the most extensively tested reconstructions: *Escherichia coli*,[13] *Geobacter sulferreducens*,[14] and *Saccharomyces cerevisiae*.[15]

In terms of gene coverage, the most complete of these reconstructions is that of *E. coli*,[13] which has been the product of over a decade of sustained effort.[16] The most recent reconstruction includes 1260 of *E. coli's* 4405 genes, together with regulatory events, compartmentalization, P/O ratio, reaction thermodynamics, energetic requirements for maintenance, and known kinetic effects. The model has been used for a variety of applications, including the prediction of growth rates, substrate uptake, and by-product secretion for thousands of combinations of knockout strains and culture conditions.[17]

Geobacter species network reconstructions have also been developed due to their remarkable bioremediative potential.[18] These organisms have somewhat unique metabolic properties enabling them to oxidize organic compounds using a variety of toxic or radioactive metals as electron acceptors. The *G. sulferreducens* metabolic model of metabolism[14] has been used to engineer a potentially useful strain with high respiration capacity and low growth rate.[19]

The *S. cerevisiae* model is notable, in particular, for the recent efforts by the yeast community to develop a consensus network. Several reconstructions had previously been made and varied significantly in their content.[20–22] As a result, leading researchers in the yeast field along with metabolic modeling experts were brought together for a weekend 'jamboree' to agree on the specifics of yeast metabolism. The resulting consensus metabolic network reconstruction was represented in a standard format, Systems Biology Markup Language (SBML: http://www.sbml.org/)[23] and made publicly available. This consensus network is distinct from an actual *in silico* model, because it must be associated with analytical approaches to produce metabolic simulations. Many of the significant modeling approaches are detailed below.

STOICHIOMETRIC NETWORK ANALYSIS

The analysis of genome-scale metabolic models relies on three basic approaches: (1) characterizing the general network structure, (2) identifying particular flux distributions, or (3) analyzing all possible flux distributions in a network. The following sections are organized according to the corresponding analysis methods.

Structural Analysis: Characterizing the Nullspace

Nullspace analysis is the first simple tool to perform consistency validations with a metabolic network. Two or more metabolites are called *conserved moieties* if the overall concentration of all remains constant. In such cases, consumption of either metabolite involves production of the other. Some examples of conserved moieties include NAD^+ and NADH, or ATP, ADP, and AMP. Detecting conserved moieties requires only

the stoichiometric matrix **S**. Several metabolites are part of a *conserved group* if the corresponding rows in the stoichiometric matrix are *linearly dependent*. Different methods exist to evaluate dependent rows, for instance, *Gaussian elimination* or *singular value decomposition* (SVD).[24] All of these methods compute a basis for the *left nullspace* of **S** or, equivalently, a basis for the nullspace of the transpose of **S**. The set of basis vectors calculated by these methods—and the conserved moieties—are unfortunately non-unique. However, a *convex basis* can be constructed to derive a unique minimal definition for *conserved groups* (for an application example, see Famili and Palsson[25]).

Analysis of *balanced metabolites* also uses nullspace basis vectors. The *metabolite balancing equation* (2) defines a subspace of R^n, where n denotes the number of reactions, and is also the dimensionality of the space. The kernel matrix **K** is a basis for the nullspace, but it is not unique. However, any valid flux vector $\mathbf{v} = (v_1, v_2, \ldots, v_n)^{\mathrm{T}}$ that defines a flux value v_i for each reaction i is a linear combination of column vectors of **K**, that is,

$$\mathbf{v} = a_1 \cdot \mathbf{K}_1 + a_2 \cdot \mathbf{K}_2 + \ldots + a_n \cdot \mathbf{K}_n = \mathbf{K} \cdot \mathbf{a} \qquad (3)$$

Equation (3) enables the identification of key reaction properties, including those shown in the sidebar; see also Figure 3.

RESULTS OF NULLSPACE ANALYSIS

1. If the kernel matrix contains a zero-row, the corresponding reaction cannot carry a (non-zero) flux. We can remove this reaction for all analysis employing the steady-state assumption.
2. If two matrix rows differ only by a constant factor, the two reactions are coupled, that is, the flux through one reaction is always a multiple of the flux through the other reaction; consequently either both reactions are active or both are passive. Such reactions are presumably co-regulated.[26]
3. Given reversibility constraints, inconsistent reaction coupling can be detected. For example, two coupled forward-only reactions with a negative coupling factor cannot carry a non-zero flux without violating an irreversibility constraint, since one reaction would have to operate in backward mode.

Model Consistency

An important step in model building is determining the *consistency* of the network, meaning technical consistency without considering experimental data or biological interpretation. One method involves the steps shown in the sidebar; the method relies on the analysis of variability and coupling in metabolic fluxes as introduced in Burgard et al.[27] and Mahadevan and Schilling.[28] Another method, which

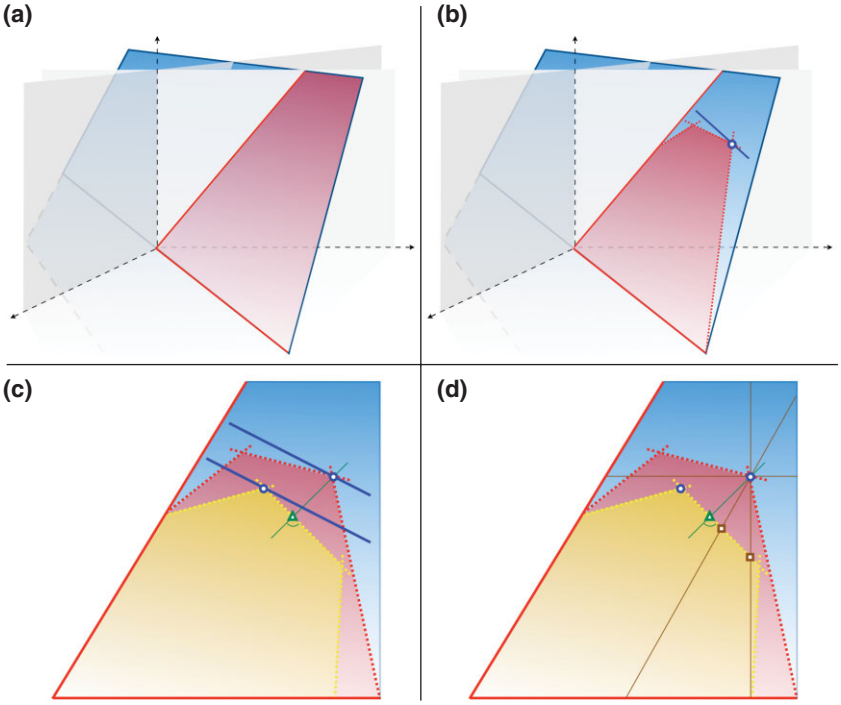


FIGURE 2 | (a) Nullspace (blue hyperplane) and the two-dimensional cone as intersection of the nullspace with the positive orthant. (b) Additional boundary constraints (dotted lines) shape a bounded convex region. The flux balance analysis (FBA) objective function (blue solid line) touches the region in the optimal point (blue circle). (c) The same cone, now in a two-dimensional view, with feasible regions for wild-type (red area) and mutant (yellow). The FBA objective function touches the regions at the optimal points (blue circles). If Minimization of metabolic adjustment (MoMA) is used instead, the distance to the best wild-type value is minimized, resulting in a different optimal value for the mutant (green triangle). (d) As a third alternative, *regulatory on/off minimization* (ROOM) minimizes the number of necessary changes. The brown lines indicate that one variable is kept constant, implying a minimal number of changes for this example. Here, two alternative optimal values are possible (brown squares).

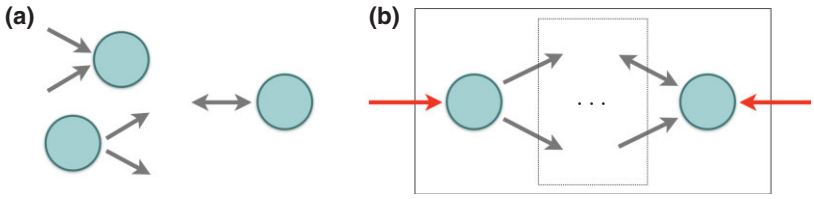


FIGURE 3 | Network inconsistencies due to dead-end metabolites (a) or reaction couplings (b). Nodes correspond to metabolites and arrows denote reactions.

focuses on stoichiometric inconsistencies, was recently described.[29]

MODEL CONSISTENCY

1. Minimize and maximize the flux value for each reaction.
   (a) If min *and* max value are zero, the reaction is a *zero flux* reaction, that is, it cannot have a flux value other than zero. It can be removed if no model corrections are made, without affecting the outcome of subsequent simulations.
   (b) If min *or* max value is zero and the reaction is reversible, we have an *unsatisfied reversibility*. Either the reversibility constraint is too lax or another component is missing, disabling the operation in one direction. Tightening this constraint might lead to better simulation performance.
   (c) If the minimal and maximal values are non-zero and have equal sign, the reaction is *essential*. Deletion of the reaction, for example, by gene knockout, is predicted to be lethal.
2. For reactions not of type (1c), set the bounds to zero. If biomass cannot be produced, the reaction is essential. Again, reaction removal is associated with lethality.

Identifying Particular Solutions: FBA

If reaction reversibilities and maximum throughput rates are incorporated as constraints on the steady-state model, the solution space reduces from the nullspace to a *convex polyhedral cone* called the *flux cone*. Because these constraints can be represented as linear equations and inequalities, *linear programming* (LP) methods can be used to identify points with optimal values of a given objective function (for an introduction to LP, see, for instance, Cormen et al.[30]). In the simplest case, a single reaction flux is optimized. Many different objective functions have been tested for their utility in predicting phenotypic behavior, relying on smaller-scale network models.[31,32] One common practice is to define an artificial *growth reaction* that takes the chemical dry-weight components of the cell, in their proper ratios, and 'produces' cellular biomass. This reaction can be used as the objective function under certain conditions, depending on the cell type and environmental conditions.[33] The term *flux balance analysis* stands for the application of LP methods to analyze fluxes under balanced metabolite conditions (Figure 2b).

As FBA was introduced, the objective function was probably the most discomforting component of the method. Objections to this approach ranged from the biological to the mathematical.[31,34] One major biological concern was whether the objective function made biological sense—do cells actually have objectives? Use of an objective function was also unsatisfying to those who doubted that FBA could predict cellular phenotypes in the absence of kinetic parameters. The FBA approach seemed to undermine traditional metabolic models which were composed of equations describing the properties of, and interactions between, small molecules and proteins.

To address these concerns, some of the earliest work in experimentally testing FBA focused on whether or not optimization of growth was a reasonable objective that could be observed in culture. Initial results were promising: model predictions of *E. coli* optimal growth and by-product secretion rates matched experimental measurements.[35] However, this did not hold true for certain substrates, like glycerol, where *E. coli* growth was suboptimal. This discrepancy between model and experiment was eventually reconciled when *E. coli* was grown on glycerol for several hundred generations and it was shown to evolve toward the computationally predicted optimal growth rate.[36] Interestingly, the phenotypic evolutionary endpoint was shown to be reproducible between cultures, whereas the underlying gene expression states varied.[37] The link between optimal growth and evolution was further strengthened when FBA was able to predict *a priori* the endpoint growth rates for several *E. coli* deletion strains.[38]

Note that often the optimal flux distributions are not unique. Rather, in general, a set of

alternate optima satisfy the LP problem, and the available solvers may therefore return different results. Sometimes, it is desirable to enumerate all alternate optima for a given objective, but this is computationally challenging.[39,40] Another way to deal with this ambiguity is *flux variability analysis*, which examines how individual fluxes can be changed without affecting optimality.[28,41] Some LP packages directly report variability by *sensitivity ranges*. The sensitivity range can also be computed by determining the optimal value for a given objective, fixing the optimal objective value (or a desired optimality range), and minimizing and maximizing fluxes for reactions of interest.

Finally, multiple iterations of FBA can be used to generate dynamic simulations. An initial optimization can be run for any given starting conditions, and using the resulting flux distribution, external and internal initial concentrations can be updated over a given time step. Critically, this time step must be large enough that the quasi-steady-state assumption in Eq. (2) still holds. However, in practice, a time step of $\sim$ 1 s should be large enough. The new conditions define the environment for the next iteration step, leading to a time-course for the environmental conditions as well as for optimal flux patterns. With this type of modeling, glucose uptake was predicted on minimal media under aerobic and anaerobic conditions.[42] Some groups have also integrated kinetic information with flux balance models.[43–46]

Deletion Strain Phenotypes

A common use of FBA is to compute the essentiality of all the genes in the network, by constraining the corresponding reactions fluxes to zero and comparing to observed deletion strain phenotypes.[47] However, some alternate methods have since been proposed. Using a reference flux vector of the wild-type (determined experimentally or estimated by an FBA simulation of the wild-type strain), these optimization strategies minimize the adjustment compared with wild-type fluxes. *Minimization of metabolic adjustment* (MoMA)[48] uses quadratic programming to minimize the sum-of-squares difference between mutant and wild-type reference flux distribution. Note that the MoMA solution is not optimal in terms of the wild-type objective. For instance, if the wild-type maximizes for biomass production, the mutant type might not exploit its full growth potential (see Figure 2c). Furthermore, because it depends on an initial non-unique FBA solution, the MoMA solution is also non-unique. *Regulatory on/off minimization* (ROOM)[49] minimizes the number of (significant) flux changes associated with regulation effort. ROOM defines a set of Boolean *on/off* variables for all reactions. An *on* state means that the corresponding reaction is up- or down-regulated. For *off* variables, an additional constraint ensures that the mutant flux lies within a predefined interval around the wild-type flux; deviations inside the interval are regarded *insignificant*. Due to the binary variables, *mixed integer linear programming* (MILP) is used to minimize the number of significant regulation changes (see Figure 2d). Overall, the main difference between MoMA and ROOM is in the motivation behind the approaches. MoMA has a mathematical origin in the formulation of a minimal response to perturbations. ROOM uses a more qualitative, biological approach to control of gene expression, assuming that a cell, in the long run, tries to minimize the number of significant flux changes.

Comprehensive Approaches

It is also possible to analyze flux cones without the need to define an objective function. Next, we introduce two approaches that treat the flux cone as a whole through *minimal sets of elements: pathway analysis* and the determination of *minimal cut sets*.

Pathway Analysis

As described above, all steady-state flux distributions can be constructed from nullspace basis vectors. *Elementary* (*flux*) *modes* have similar properties, but also some important differences: elementary modes (EMs) are *feasible* vectors, whereas nullspace basis vectors might violate reaction reversibility constraints. In addition, all feasible flux vectors—and only feasible ones—can be constructed from *non-negative* (*conic*) combinations of EMs. From nullspace vectors, feasible and infeasible vectors can be constructed, using any *linear* combination.

To calculate a unique, smallest possible set of EMs, an additional constraint is introduced: *minimality*. A feasible flux vector is minimal (or *elementary*) if no other flux vector has the same reactions with zero flux plus additional ones. As a consequence, EMs describe basic non-decomposable operation modes of the network.[50,51] *Extreme pathways* (EPs) constitute a very similar concept, associated with the 'shortest' possible paths in the network (note that 'shortest' is misleading). EMs are actually a superset of *EPs*, arising from a slightly different treatment of reversible transport reactions (see Klamt and Stelling[52] and Wagner and Urbanczik[53] for exact definitions). Moreover, in a strong mathematical sense, neither EMs nor EPs are minimal, and hence also not a minimal set of generators. EMs and EPs operate in a reconfigured

network with enhanced reaction dimensionality, and they are only minimal in the augmented flux space. We refer the reader to Klamt and Stelling[52] and Wagner and Urbanczik[53] and focus here on EMs. The concepts also apply to EPs unless stated otherwise.

Unfortunately, these *pathway sets* grow exponentially with increasing network size. It is not yet possible to compute EMs for genome-scale networks under general conditions.[51,54] Nevertheless, EMs have interesting properties because they are minimal, and hence contain information about the smallest functional units of complex networks. Pathway analysis uncovers *all* alternative pathways in contrast to FBA, where only a single (optimal) pathway is found. All modes of the network are superpositions of EMs, and an *alpha-spectrum*—the contribution of individual pathways to an observed (*in vivo*) flux pattern—has been analyzed for a simplified core metabolic network and for human red blood cell metabolism.[55] Reaction participation in EMs also indicates a reaction's importance for different substrate/product conversions.

Furthermore, the number of active reactions in a single EM is related to the necessary enzymes, and a cost function can be derived. In combination with benefit functions known from FBA, we get a *multiple objective,* which might indeed be better suited to represent the complex optimization strategy of the cell.[51,56] To analyze *flux variability*, only EMs above a certain optimality criterion can be considered, and the *coefficient of variation*—the relative standard deviation of reaction fluxes—indicates variability or flexibility of the fluxes[51] (see Figure 4 for an example). In addition, EMs can be used to reliably predict viability of gene deletion mutants. Therefore, only EMs are kept where the reaction of interest does not participate. An empty set for the perturbed network indicates that the network is structurally unable to operate under steady-state conditions.[56]

Minimal Cut Sets

Minimal cut sets give an opposite view of a metabolic network when compared with EMs. EMs describe minimal requirements to operate the network; *minimal cut sets* define smallest possible reaction sets that cause network failure with respect to a specific function such as biomass production.[57] In fact, both concepts completely describe the metabolic network (of course without involved regulation); formally, they are dual.[58] The simplest *minimal cut sets* are essential reactions, since they have minimal size, and their removal disables any target reaction. *Synthetic lethals* are minimal cut sets of size two: lethality is caused by the simultaneous knockout of two non-essential reactions. We can proceed with this concept for larger

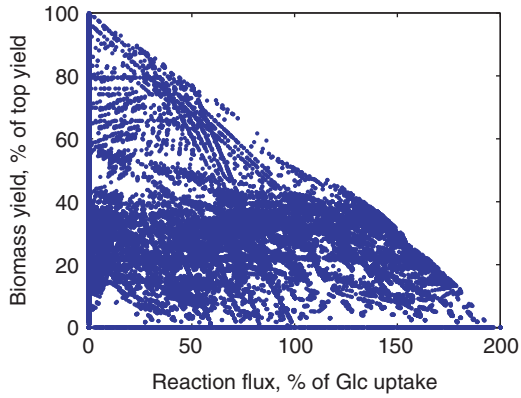


FIGURE 4 | Ethanol excretion rate related to biomass yield for 178,575 elementary modes (EMs) of an *Escherichia coli* central metabolism network with 97 metabolites and 120 reactions; only growth on glucose (Glc) was considered. Ethanol production is possible only for suboptimal growth.

sets, requiring minimality or *irreducibility*, similar to the non-decomposability of EMs. Not surprisingly, it is computationally expensive to enumerate all minimal cut sets for large networks.

Network dysfunction is associated with minimal cut sets. Such dysfunction may be caused internally through spontaneous mutations, or have external reasons and be intentional in the case of gene deletions or RNA interference. Using minimal cut sets, internal structural fragility and robustness can be analyzed. In metabolic engineering, minimal cut sets identify potential drug targets, driving new hypotheses, or narrowing down test candidates for expensive experiments. Finally, minimal cut sets have practical applications in the design and optimization of biotechnological processes. Methods such as OptKnock[59] can be used to study minimal intervention strategies for overproduction of target biochemicals in microbial strains.

INCORPORATING REGULATION

Constraint-based analysis has for the most part only focused on metabolism, with a notable exception in signal transduction.[60] For instance, it is unclear whether the assumptions required for FBA can be justifiably applied to other biological processes. Instead, some studies have integrated FBA-based metabolic models with possibly more appropriate models for transcriptional regulatory and signaling networks. Also, these approaches start to address the dynamic behavior of metabolic networks.

However, gene expression has a dramatic additional effect on metabolic networks. For example,

only about 50% of the *E. coli* genome is expressed under typical culture conditions, and therefore half of the stoichiometric matrix might be significantly constrained at any given time. This has major ramifications for convex-based analyses that ignore gene expression. The calculated pathways will include several false positives which never occur in the living cell. Furthermore, the optimal solutions determined by FBA will most likely be incorrect in all but relatively simple growth conditions.

To explicitly account for the effects of gene expression on metabolic behavior, an integrated procedure was developed whereby transcriptional regulatory events were described using Boolean logic and used to constrain the solution space further.[61] The status of regulated transcription is found by evaluating intra- or extra-cellular conditions. Transcription may be switched on or off by the presence or absence of particular metabolites, proteins, or signaling molecules. Sometimes, concentrations above a certain threshold are required to trigger regulatory events. If an activating or inhibiting expression changes—caused by updated concentration values—the rate constraints for the regulated reaction change, resulting in up-/down-regulations. In the simplest case, inhibition sets the constraints to zero, and activation causes a reset back to the original boundary values. This method was found to significantly reduce the number of EPs calculated in a simple system.[62]

Combination of the method with FBA in an approach called regulatory FBA or rFBA (Figure 5) produced dramatically more accurate model predictions in organisms such as *E. coli*[63] and yeast.[20] More specifically, Covert et al.[64] have used rFBA to derive time courses for *E. coli* with a genome-scale model and to correctly predict viability for 106 of 116 mutant strain/growth medium conditions. In Covert et al.,[63] an extended model with 1010 genes was used in an iterative process to generate and test hypotheses, and missing components and interactions in regulatory or metabolic networks could be identified.

The initial work focused on reconstructing transcriptional regulatory networks based directly on findings from the primary literature. Boolean logic best represented the typically qualitative conclusions of experimental studies (e.g., 'transcription factor X was observed to repress expression of target gene Y'). However, subsequent studies have used microarray data to constrain FBA[65] as well as machine learning techniques to generate more sophisticated regulatory network models.[66] rFBA models stand to gain significantly from these developments.

CURRENT CHALLENGES

Automated Network Reconstruction

The earliest metabolic network reconstructions were manually generated. Since that time, several tools that support network reconstruction have been developed, such as Pathway Tools,[67] KEGG Pathways,[68] PUMA2,[69] and SimPheny.[70] However, unknown reactions and the necessary validation of database entries still result in time-intensive manual network curation.[71] Therefore, although significant effort is dedicated to developing computational approaches for fully automatic network reconstruction[72] and reconciliation,[29] curation-based efforts such as the jamboree currently produce the 'gold standard' reconstructions.

One automated reconstruction method aims to identify necessary reactions from an organism-wide database such that these reactions could allow growth of mutants that are experimentally viable, but predicted to be inviable by an existing stoichiometric model.[73] This approach considers

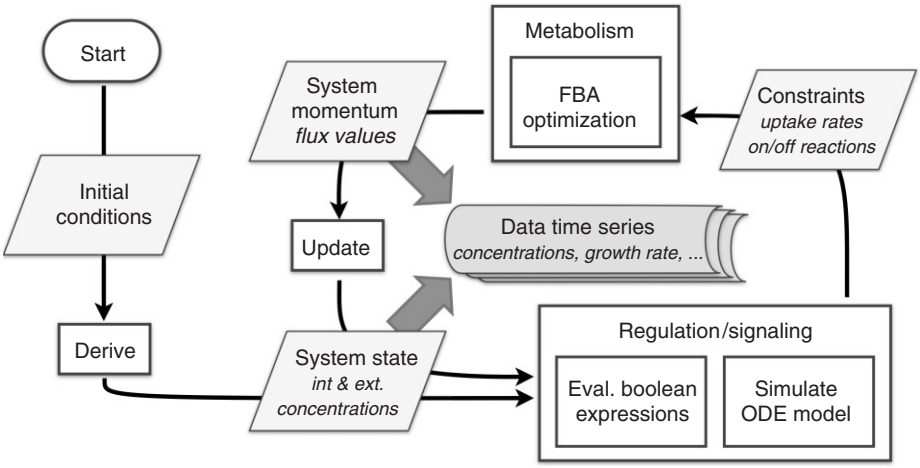


FIGURE 5 | Schematic representation of regulated flux balance analysis (rFBA) using Boolean expressions to simulate regulatory elements. The concept is generalized to integrated FBA (iFBA), also incorporating ordinary differential equations (ODEs) to simulate regulation. The algorithm is an iterative procedure, generating time series output at each iteration.

only one experimental condition at a time, it yields (potentially large) sets of candidate reactions, and it is computationally expensive because for each condition and candidate reaction FBA analysis has to be performed. Other methods rely on information fusion from pathway databases to reconstruct models *de novo* but so far they have not yielded functional models,[74] or only prototypes that lack validation with experimental data.[75] In addition, powerful methods exist to identify metabolic genes for a given enzymatic function[76,77]—but this function has to be already contained in the model. Optimization-based methods help identifying gaps in metabolic network reconstructions, and they consolidate the models by introducing new reactions, or by modifying existing reactions. Existing algorithms, however, do not consider global changes in network structures or potential effects on the quality of model predictions.[78] Available methods, thus, have limitations in automatically generating predictive network models, and novel concepts, such as Clauset et al.,[79] are needed.

Cellular Optimality and Design

The choice of a biologically meaningful objective function is critical for FBA. Identifying the objective function—or cellular design principles—can be regarded as the *inverse problem* of FBA. Where FBA finds an optimal flux vector given some objective function, a more challenging problem is to infer the objective for an experimentally determined reference flux vector. Early work on this problem used a bi-level optimization approach to test existing hypothetical objectives.[80] The *OptFind* method solves two optimizations in one step, using the duality theorem of LP to flatten the two optimality layers. In an application to *E. coli* under aerobic and anaerobic conditions, a high coefficient of importance was found for the biomass function.[80] Method refinements allowed to derive objectives *de novo*.[81] Similar techniques—in combination with binary variables, requiring the use of *MILP*— can find optimal knockout strategies leading to the overproduction of a desired product.[59] By automatically querying reaction databases, an optimal strain is composed, and the method yields optimal substrates for different microbes.[82] Similar approaches have been established using stochastic optimization[83] and sensitivity analysis of metabolic flux distributions.[84]

Another way to analyze biological networks begins with the hypothesis that cells have optimized their operation over evolutionary time-scales, as assumed in FBA. However, FBA does not provide insight into specific control mechanisms. Further elaborations of principles of optimal control theory (already present in MoMA) could elucidate large-scale metabolic control circuits. For instance, the cybernetic approach[85] treats metabolic control as a set of dynamic, optimal resource allocation problems that are solved in parallel with the mass balances. Predictions on gene expression and enzyme activity result from choices between competing alternatives, each with a relative cost and benefit for the organism. In addition, postulates for specific pathway architectures have resulted from this approach.[86]

Toward Large-Scale Network Integration and Dynamics

To build on these successes toward creation of a whole-cell model, approaches to model integration must be developed, in particular, with large-scale kinetic models.[87] The integration of metabolism with transcriptional regulation was already described above. More recently, these integration methods were expanded to include ordinary differential equations (ODEs).[43] The method (called 'integrated FBA' or iFBA; see Figure 6) was used to build an FBA model of *E. coli* central metabolism, together with a Boolean logic-based regulatory network and a set of ODEs that described catabolite repression.[88] The resulting model had significant advantages over either the rFBA or ODE-based models alone, particularly in predicting the consequences of gene perturbation. Another approach has been developed which incorporates coarse-grained time-scale information about signaling dynamics with metabolic models.[89] Finally, recent advances toward large-scale dynamic models include those that rely on simplified reaction kinetics[90,91] or on ensembles of models.[92]

In addition to explicitly modeling cellular regulation, one can exploit that stoichiometric constraints restrict the systems dynamics, for instance, by conserved moieties.[25] Early work addressed this topic for chemical reaction networks. For instance, in the 1970s, Feinberg, Horn, and Jackson started deriving theorems to determine the possible dynamic regimes, such as multistability and oscillations, based on network structure alone. The specific challenges posed by biological systems lead to application studies as well as further theory development.[93] For instance, the theory can be used in stability analysis and for model discrimination by safely rejecting hypotheses on reaction mechanisms; this analysis relies on a modular approach where subnetworks that correspond to EMs are investigated individually.[94] Other algorithms for the identification of dependent species in large

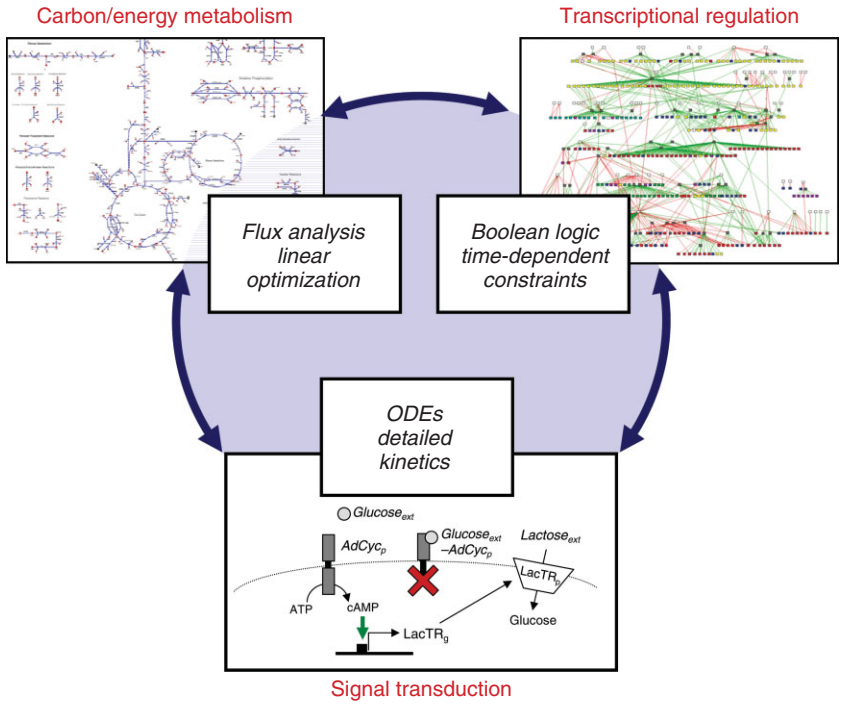


FIGURE 6 | Models have been constructed which build on the constraint-based framework to integrate metabolic, transcriptional regulatory, and signal transduction networks (each represented using different mathematics) into a single model.

biochemical systems—to be employed, for instance, in model reduction—have recently become available.[95]

CONCLUSION

In sum, large-scale metabolic network modeling has matured as a field, with a library of computational techniques, published networks for a large and increasing number of organisms, and an extensive body of supporting experimental evidence. It is clear that such modeling can be extremely useful, notwithstanding its limitations. These studies have also established the essentiality of other biological models—metabolism alone cannot explain most observed phenotypic behaviors. It seems unrealistic to expect that data which would allow detailed genome-scale modeling of other biological networks will arrive in the next few years. However, we re-emphasize that scientists felt the same way about metabolism just over a decade ago.

REFERENCES

1. Price ND, Reed JL, Palsson BO. Genome-scale models of microbial cells: evaluating the consequences of constraints. *Nat Rev Microbiol* 2004, 2:886–897.
2. Rocha I, Forster J, Nielsen J. Design and application of genome-scale reconstructed metabolic models. *Methods Mol Biol* 2008, 416:409–431.
3. Jaqaman K, Danuser G. Linking data to models: data regression. *Nat Rev Mol Cell Biol* 2006, 7:813–819.
4. Covert MW, et al. Metabolic modeling of microbial strains *in silico*. *Trends Biochem Sci* 2001, 26:179–186.
5. Apweiler R, et al. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2004, 32:D115–D119.
6. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res* 2002, 30:47–49.
7. Karp PD, et al. Expansion of the BioCyc collection of pathway/genome databases to 160 genomes. *Nucleic Acids Res* 2005, 33:6083–6089.
8. Ogata H, et al. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 1999, 27:29–34.
9. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000, 28:304–305.
10. Wiechert W. 13C metabolic flux analysis. *Metab Eng* 2001, 3:195–206.
11. Wittmann C. Metabolic flux analysis using mass spectrometry. *Adv Biochem Eng Biotechnol* 2002, 74:39–64.
12. Reed JL, et al. Towards multidimensional genome annotation. *Nat Rev Genet* 2006, 7:130–141.

13. Feist AM, et al. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007, 3:121.
14. Mahadevan R, et al. Characterization of metabolism in the Fe(III)-reducing organism Geobacter sulfurreducens by constraint-based modeling. *Appl Environ Microbiol* 2006, 72:1558–1568.
15. Herrgard MJ, et al. A consensus yeast metabolic network reconstruction obtained from a community approach to systems biology. *Nat Biotechnol* 2008, 26:1155–1160.
16. Reed JL, Palsson BO. Thirteen years of building constraint-based *in silico* models of *Escherichia coli*. *J Bacteriol* 2003, 185:2692–2699.
17. Feist AM, Palsson BO. The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*. *Nat Biotechnol* 2008, 26:659–667.
18. Lovley DR. Cleaning up with genomics: applying molecular biology to bioremediation. *Nat Rev Microbiol* 2003, 1:35–44.
19. Izallalen M, et al. Geobacter sulfurreducens strain engineered for increased rates of respiration. *Metab Eng* 2008, 10:267–275.
20. Herrgard MJ, et al. Integrated analysis of regulatory and metabolic networks reveals novel regulatory mechanisms in *Saccharomyces cerevisiae*. *Genome Res* 2006, 16:627–635.
21. Kuepfer L, Sauer U, Blank LM. Metabolic functions of duplicate genes in *Saccharomyces cerevisiae*. *Genome Res* 2005, 15:1421–1430.
22. Caspi R, et al. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 2006, 34:D511–D516.
23. Hucka M, et al. The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics* 2003, 19:524–531.
24. Sauro HM, Ingalls B. Conservation analysis in biochemical networks: computational issues for software writers. *Biophys Chem* 2004, 109:1–15.
25. Famili I, Palsson BO. The convex basis of the left null space of the stoichiometric matrix leads to the definition of metabolically meaningful pools. *Biophys J* 2003, 85:16–26.
26. Notebaart RA, et al. Co-regulation of metabolic genes is better explained by flux coupling than by network distance. *PLoS Comput Biol* 2008, 4:e26.
27. Burgard AP, et al. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res* 2004, 14:301–312.
28. Mahadevan R, Schilling CH. The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. *Metab Eng* 2003, 5:264–276.
29. Gevorgyan A, Poolman MG, Fell DA. Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics* 2008, 24:2245–2251.
30. Cormen TH, et al., *Introduction to Algorithms*. Burr Ridge, IL: McGraw-Hill; 2001.
31. Schuetz R, Kuepfer L, Sauer U. Systematic evaluation of objective functions for predicting intracellular fluxes in *Escherichia coli*. *Mol Syst Biol* 2007, 3:119.
32. Savinell JM, Palsson BO. Network analysis of intermediary metabolism using linear optimization. I. Development of mathematical formalism. *J Theor Biol* 1992, 154:421–454.
33. Fischer E, Sauer U. Large-scale *in vivo* flux analysis shows rigidity and suboptimal performance of *Bacillus subtilis* metabolism. *Nat Genet* 2005, 37:636–640.
34. Pramanik J, Keasling JD. Effect of *Escherichia coli* biomass composition on central metabolic fluxes predicted by a stoichiometric model. *Biotechnol Bioeng* 1998, 60:230–238.
35. Edwards JS, Ibarra RU, Palsson BO. *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 2001, 19:125–130.
36. Ibarra RU, Edwards JS, Palsson BO. *Escherichia coli* K-12 undergoes adaptive evolution to achieve *in silico* predicted optimal growth. *Nature* 2002, 420:186–189.
37. Fong SS, Joyce AR, Palsson BO. Parallel adaptive evolution cultures of *Escherichia coli* lead to convergent growth phenotypes with different gene expression states. *Genome Res* 2005, 15:1365–1372.
38. Fong SS, Palsson BO. Metabolic gene-deletion strains of *Escherichia coli* evolve to computationally predicted growth phenotypes. *Nat Genet* 2004, 36:1056–1058.
39. Lee S, et al. Recursive MILP model for finding all the alternate optima in LP models for metabolic networks. *Comput Chem Eng* 2000, 24:711–716.
40. Phalakornkule C, et al. A MILP-based flux alternative generation and NMR experimental design strategy for metabolic engineering. *Metab Eng* 2001, 3:124–137.
41. Bilu Y, et al. Conservation of expression and sequence of metabolic genes is reflected by activity across metabolic states. *PLoS Comput Biol* 2006, 2:e106.
42. 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, 60:3724–3731.
43. Covert MW, et al. Integrating metabolic, transcriptional regulatory and signal transduction models in *Escherichia coli*. *Bioinformatics* 2008, 24:2044–2050.
44. Mahadevan R, Edwards JS, Doyle FJ 3rd. Dynamic flux balance analysis of diauxic growth in *Escherichia coli*. *Biophys J* 2002, 83:1331–1340.
45. Smallbone K, et al. Something from nothing: bridging the gap between constraint-based and kinetic modelling. *FEBS J* 2007, 274:5576–5585.

46. Yugi K, et al. Hybrid dynamic/static method for large-scale simulation of metabolism. *Theor Biol Med Model* 2005, 2:42.
47. Edwards JS, Palsson BO. The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. *Proc Natl Acad Sci USA* 2000, 97:5528–5533.
48. Segre D, Vitkup D, Church GM. Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci USA* 2002, 99:15112–15117.
49. Shlomi T, Berkman O, Ruppin E. Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci USA* 2005, 102:7695–7700.
50. Gagneur J, Klamt S. Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics* 2004, 5:175.
51. Terzer M, Stelling J. Large-scale computation of elementary flux modes with bit pattern trees. *Bioinformatics* 2008, 24:2229–2235.
52. Klamt S, Stelling J. Two approaches for metabolic pathway analysis? *Trends Biotechnol* 2003, 21:64–69.
53. Wagner C, Urbanczik R. The geometry of the flux cone of a metabolic network. *Biophys J* 2005, 89:3837–3845.
54. Klamt S, Stelling J. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep* 2002, 29:233–236.
55. Wiback SJ, Mahadevan R, Palsson BO. Reconstructing metabolic flux vectors from extreme pathways: defining the alpha-spectrum. *J Theor Biol* 2003, 224:313–324.
56. Stelling J, et al. Metabolic network structure determines key aspects of functionality and regulation. *Nature* 2002, 420:190–193.
57. Klamt S, Gilles ED. Minimal cut sets in biochemical reaction networks. *Bioinformatics* 2004, 20:226–234.
58. Klamt S. Generalized concept of minimal cut sets in biochemical networks. *Biosystems* 2006, 83:233–247.
59. Burgard AP, Pharkya P, Maranas CD. Optknock: a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng* 2003, 84:647–657.
60. Papin JA, Palsson BO. The JAK-STAT signaling network in the human B-cell: an extreme signaling pathway analysis. *Biophys J* 2004, 87:37–46.
61. Covert MW, Schilling CH, Palsson B. Regulation of gene expression in flux balance models of metabolism. *J Theor Biol* 2001, 213:73–88.
62. Covert MW, Palsson BO. Constraints-based models: regulation of gene expression reduces the steady-state solution space. *J Theor Biol* 2003, 221:309–325.
63. Covert MW, et al. Integrating high-throughput and computational data elucidates bacterial networks. *Nature* 2004, 429:92–96.
64. Covert MW, Palsson BO. Transcriptional regulation in constraints-based metabolic models of *Escherichia coli*. *J Biol Chem* 2002, 277:28058–28064.
65. Akesson M, Forster J, Nielsen J. Integration of gene expression data into genome-scale metabolic models. *Metab Eng* 2004, 6:285–293.
66. Bonneau R, et al. A predictive model for transcriptional control of physiology in a free living cell. *Cell* 2007, 131:1354–1365.
67. Karp PD, Paley S, Romero P. The pathway tools software. *Bioinformatics* 2002, 18:S225–S232.
68. Kanehisa M, et al. The KEGG databases at GenomeNet. *Nucleic Acids Res* 2002, 30:42–46.
69. Maltsev N, et al. PUMA2—grid-based high-throughput analysis of genomes and metabolic pathways. *Nucleic Acids Res* 2006, 34:D369–D372.
70. Price ND, et al. Genome-scale microbial *in silico* models: the constraints-based approach. *Trends Biotechnol* 2003, 21:162–169.
71. Ott MA, Vriend G. Correcting ligands, metabolites, and pathways. *BMC Bioinformatics* 2006, 7:517.
72. Nikoloski Z, et al. Metabolic networks are NP-hard to reconstruct. *J Theor Biol* 2008, 254:807–816.
73. Reed JL, et al. Systems approach to refining genome annotation. *Proc Natl Acad Sci USA* 2006, 103:17480–17484.
74. DeJongh M, et al. Toward the automated generation of genome-scale metabolic networks in the SEED. *BMC Bioinformatics* 2007, 8:139.
75. Arakawa K, et al. GEM System: automatic prototyping of cell-wide metabolic pathway models from genomes. *BMC Bioinformatics* 2006, 7:168.
76. Kharchenko P, et al. Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics* 2006, 7:177.
77. Kharchenko P, Vitkup D, Church GM. Filling gaps in a metabolic network using expression information. *Bioinformatics* 2004, 20:i178–i185.
78. Satish Kumar V, Dasika MS, Maranas CD. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics* 2007, 8:212.
79. Clauset A, Moore C, Newman ME. Hierarchical structure and the prediction of missing links in networks. *Nature* 2008, 453:98–101.
80. Burgard AP, Maranas CD. Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol Bioeng* 2003, 82:670–677.
81. Gianchandani EP, et al. Predicting biological system objectives *de novo* from internal state measurements. *BMC Bioinformatics* 2008, 9:43.
82. Pharkya P, Burgard AP, Maranas CD. OptStrain: a computational framework for redesign of

microbial production systems. *Genome Res* 2004, 14:2367–2376.

83. Patil KR, et al. Evolutionary programming as a platform for *in silico* metabolic engineering. *BMC Bioinformatics* 2005, 6:308.
84. Park JH, et al. Metabolic engineering of *Escherichia coli* for the production of L-valine based on transcriptome analysis and *in silico* gene knockout simulation. *Proc Natl Acad Sci USA* 2007, 104:7797–7802.
85. Kompala DS, Ramkrishna D, Tsao GT. Cybernetic modeling of microbial growth on multiple substrates. *Biotechnol Bioeng* 1984, 26:1272–1281.
86. Young JD, et al. Integrating cybernetic modeling with pathway analysis provides a dynamic, systems-level description of metabolic control. *Biotechnol Bioeng* 2008, 100:542–559.
87. Jamshidi N, Palsson BO. Formulating genome-scale kinetic models in the post-genome era. *Mol Syst Biol* 2008, 4:171.
88. Kremling A, Bettenbrock K, Gilles ED. Analysis of global control of *Escherichia coli* carbohydrate uptake. *BMC Syst Biol* 2007, 1:42.
89. Min Lee J, et al. Dynamic analysis of integrated signaling, metabolic, and regulatory networks. *PLoS Comput Biol* 2008, 4:e1000086.
90. Liebermeister W, Klipp E. Bringing metabolic networks to life: integration of kinetic, metabolic, and proteomic data. *Theor Biol Med Model* 2006, 3:42.
91. Liebermeister W, Klipp E. Bringing metabolic networks to life: convenience rate law and thermodynamic constraints. *Theor Biol Med Model* 2006, 3:41.
92. Tran LM, Rizk ML, Liao JC. Ensemble modeling of metabolic networks. *Biophys J* 2008, 95:5606–5617.
93. Craciun G, Tang Y, Feinberg M. Understanding bistability in complex enzyme-driven reaction networks. *Proc Natl Acad Sci USA* 2006, 103:8697–8702.
94. Conradi C, et al. Subnetwork analysis reveals dynamic features of complex (bio)chemical networks. *Proc Natl Acad Sci USA* 2007, 104:19175–19180.
95. Vallabhajosyula RR, Chickarmane V, Sauro HM. Conservation analysis of large biochemical networks. *Bioinformatics* 2006, 22:346–353.