Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2012, Article ID 323472, 10 pages
doi:10.1155/2012/323472

*Research Article*

# Gap Detection for Genome-Scale Constraint-Based Models

**J. Paul Brooks,[1,2] William P. Burns,[1] Stephen S. Fong,[1,3]
Chris M. Gowen,[1,3] and Seth B. Roberts[1,3]**

[1] *Center for the Study of Biological Complexity, Virginia Commonwealth University, P.O. Box 843083, Richmond, VA 23284, USA*
[2] *Department of Statistical Sciences and Operations Research, Virginia Commonwealth University, P.O. Box 843083, Richmond, VA 23284, USA*
[3] *Department of Chemical and Life Science Engineering, Virginia Commonwealth University, P.O. Box 843083, Richmond, VA 23284, USA*

Correspondence should be addressed to J. Paul Brooks, jpbrooks@vcu.edu

Received 23 April 2012; Accepted 16 July 2012

Academic Editor: T. Akutsu

Constraint-based metabolic models are currently the most comprehensive system-wide models of cellular metabolism. Several challenges arise when building an *in silico* constraint-based model of an organism that need to be addressed before flux balance analysis (FBA) can be applied for simulations. An algorithm called FBA-Gap is presented here that aids the construction of a working model based on plausible modifications to a given list of reactions that are known to occur in the organism. When applied to a working model, the algorithm gives a hypothesis concerning a minimal medium for sustaining the cell in culture. The utility of the algorithm is demonstrated in creating a new model organism and is applied to four existing working models for generating hypotheses about culture media. In modifying a partial metabolic reconstruction so that biomass may be produced using FBA, the proposed method is more efficient than a previously proposed method in that fewer new reactions are added to complete the model. The proposed method is also more accurate than other approaches in that only biologically plausible reactions and exchange reactions are used.

## 1. Introduction

Flux balance analysis (FBA) is the use of a linear program (LP) to model the flow of metabolites through the network of reactions in a cell [1]. FBA simulations give insight into the relative rates at which reactions occur when the cell is optimized for a specific objective. A fundamental assumption of FBA is that organisms can function optimally (often as a result of adaptive evolution) in that they make optimal use of scarce resources to serve the needs of the organism. This characterization of cell behavior naturally leads to a math programming modeling paradigm. FBA has been used to predict growth rates, gene essentiality, and other features of multiple organisms [2–5].

Several related challenges are encountered in the building of metabolic reconstructions. To apply FBA to a constraint-based model, both a reaction network (representing organism-specific biochemical capabilities) and an objective (representing a desired or measurable physiological goal) need to be specified. Currently, complete reaction networks for organisms are not known. There may be reactions in a cell that must be active for the production of biomass that have not been cataloged in biological databases or documented in the literature. Another challenge is modeler error; the modeler can mistakenly omit a reaction or transport process that is necessary for the production of biomass. Aside from establishing a model that can produce biomass, a common difficulty in using FBA models is that of finding a culture medium that can allow the *in silico* cell to send flux through the biomass reaction.

Several methods for restoring functionality in broken FBA models, those incapable of a desired level of flux through the biomass reaction, have been previously proposed. GapFind [6] is a procedure that determines which metabolites in a network cannot be produced, and GapFill [6] determines a minimal set of reactions to add from a universal database so that a specified set of metabolites may be produced. These optimization-based procedures have already been integrated into the Model SEED metabolic reconstruction pipeline with some success [7]. Reed et al. [8] utilize a method that adds a minimum-sized set of reactions from a universal database that allows for a specified level of biomass production in the resulting model. MetaFlux [9] is an automated approach to find missing reactions, exchange reactions, and biomass metabolites. OptStrain [10] determines the maximum possible yield of a desired product based on the inclusion of all reactions in a universal database and then finds the minimum number of reactions from the database needed to achieve the optimal yield. Segrè et al. [11] use the Forward Propagation and Backward Propagation/Backtracking algorithms [12] to first determine the metabolites that can be produced in a model, and then find the precursors of essential nonproducible metabolites that cannot be produced.

Several investigators have proposed methods for filling gaps in metabolic networks outside of the FBA paradigm, including searching through a network of metabolites and reactions for logically possible paths [13, 14] and using logic programming to construct pathways [15]. These methods do not ensure that the mass balancing constraints of FBA models are satisfied, nor do they consider the effects of generated pathways on the production of biomass. Thus, the application of these methods does not guarantee the generation of a constraint-based model that produces biomass when FBA is applied.

A fundamental assumption of FBA modeling is that metabolites remain at constant concentration within the cell. Throughout this paper, we use the term *metabolite* to refer to any molecule whose concentration is of interest, including byproducts of metabolism, coenzymes, and protons. Let $v_j$ be the flux through reaction $j$, for each $j \in R$, which is the number of times that a reaction occurs per unit time. Let $S_{ij}$ be the stoichiometric coefficient for metabolite $i$ in reaction $j$, for each $i \in M$ and $j \in R$, with the convention that $S_{ij}$ is negative for molecules $i$ that are reactants for reaction $j$, positive for metabolites $i$ that are products for reaction $j$, and 0 otherwise. Metabolites may participate in a unidirectional or reversible *exchange reaction*. For our purposes, it will be helpful to distinguish *source reactions* from *escape reactions* and assign variables $b_i^{\text{src}}$ and $b_i^{\text{esc}}$ for the fluxes through these reactions. We wish to restrict transport fluxes to zero for any metabolite unless its concentration changes in the cell due to transport processes. The conservation of mass for metabolite $i$ may be stated as follows:

$$\sum_{j \in R} S_{ij} v_j + b_i^{\text{src}} - b_i^{\text{esc}} = 0. \tag{1}$$

The set of reactions $R$ may include a (potentially artificial) *biomass* reaction which reflects the objective of the cell in terms of which metabolites are emphasized for production or consumption by other processes. The objective

$$\max v_{\text{biomass}} \tag{2}$$

can be added to the model, reflecting the desire to maximize flux through the biomass reaction. Maximizing flux through the biomass reaction is one of several possible objectives that one could assign to a cell. FBA models with this particular objective have been shown to reflect the behavior of single-celled organisms during cell growth. Assessing whether positive biomass production is possible is an effective method for testing the completeness of a metabolic reconstruction. If an FBA model is incapable of producing biomass, then there is likely a gap in the reaction network.

Upper and lower bounds on each reaction flux are specified. If possible, these bounds are based on experimentally observed fluxes and free energy considerations, as for the *S. cerevisiae* and *E. coli* models [16–18]. If not, then a common lower and upper bound for all reactions can be assigned, and the fluxes returned by FBA give the investigator an idea of the relative activity of the reactions in the network for a given biomass reaction; the actual flux values in this latter case are less important than the ratios. For example if $v_j/v_k \geq 4$, the model indicates that a mechanism for maximizing biomass production exists wherein reaction $j$ is at least 4 times as active as reaction $k$. If we generate (1) for metabolites and reactions within a cell and add the flux bounds, we obtain the linear programming-based FBA model. The general model can be expressed compactly as follows:

$$\begin{gathered}\max v_{\text{biomass}} \\ Sv + b^{\text{src}} - b^{\text{esc}} = 0, \\ L \leq v \leq U, \\ L^{\text{src}} \leq b^{\text{src}} \leq U^{\text{src}}, \\ L^{\text{esc}} \leq b^{\text{esc}} \leq U^{\text{esc}}.\end{gathered} \tag{3}$$

In this paper, we propose a new approach to address the challenges of building FBA models called FBA-Gap. The procedure identifies gaps in the metabolic network that are preventing flux through a specified objective, which in our case is the biomass reaction that represents cellular growth. Given a metabolic reconstruction and a biomass reaction, the goal is to find the most plausible modification of the metabolic reconstruction so that the model is capable of sending flux through the biomass reaction. FBA-Gap uses mathematical optimization to determine a minimum cost set of additional exchange reactions needed such that the flux through the biomass reaction can exceed a given threshold. Costs are assigned to source and escape reactions *a priori* based on their plausibility and distance to the biomass reaction. In general, exchange reactions for metabolites that exist in the extracellular compartment are given a low cost, while exchange reactions for metabolites that exist only in cytosolic and intracellular compartments are given a high cost. The output is a minimum cost set of exchange reactions and a flux distribution for the expanded reaction network. If the model is robust and has no detrimental gaps, the selected exchange reactions will correspond to missing transport reactions for uptake of metabolites from *in silico* culture medium or for discharge of byproducts into the extracellular space. However, if the model has internal gaps in the reaction network, exchange reactions will be added for internal metabolites that are furthest from the biomass reaction.

Our method is a departure from previous gap-filling methods in that we place an increased emphasis on the accuracy of the final model. The approach is to preserve the set of reactions in the initial model and to direct the model builder to a set of reactions that lead to a biomass-producing model and can be added with high confidence. In the GapFind/GapFill framework, reactions are added until *every* metabolite in the model is produced, and many additional reactions may be added to a model that are not required for the production of biomass. We will demonstrate that the proposed method is less computationally intensive than GapFind/GapFill. In the method described in [8], hereafter referred to as GapReed, reactions may be added to the model which are downstream/upstream of the actual gap. In other words, there is no attempt to ensure that modifications address gaps in the "backbone" of the network; the gaps may be masked by implausible exchange reactions or secondary pathways. The emphasis in our method is directing the modeler to the gaps in the backbone of the network that can be addressed by adding high-confidence reactions to the model.

The cost structure in FBA-Gap for the artificial exchange reactions is crucial to the proper identification of gaps in the metabolic network. Our approach is to identify the gaps that are furthest distance from the biomass reaction, utilizing as much of the existing network as possible. A trivial "fix" to any constraint-based model would be to add exchange reactions for every component of the biomass reaction, which would always result in a solution that has no biological relevance. Measuring distance in a metabolic network is a well-studied problem. Distances between metabolites in a metabolic network have been used to establish and refute scale-freeness [19, 20]. Investigators have noted difficulties associated with the inclusion of coenzymes in distance calculations, not the least of which is specifying which metabolites are coenzymes [13]. Some of these coenzymes are ubiquitous so that every metabolite appears near every other metabolite. Solutions to these difficulties include the introduction of compartments [13], excluding the most common metabolites from distance calculations [21], and using the Euclidean distance of attribute vectors for metabolites [14]. In FBA-Gap, the length of a path in the metabolic network is based on the number of reactions in which each metabolite occurs, penalizing paths that pass through often-occurring metabolites. Gaps where coenzymes play a prominent role can be discovered, but preference is given to other gaps.

In the remainder of the paper, we describe the FBA-Gap method for building metabolic reaction networks and demonstrate its effectiveness in computational experiments. The method is used to help create a new metabolic reconstruction for a cellular organism based on a partial reconstruction. We compare the accuracy and computation time of FBA-Gap to existing gap-filling methods for this model. We then remove the exchange reactions from several existing models of organisms and apply FBA-Gap, yielding a hypothesis for minimal media for each organism. Finally, we delete a portion of the internal reactions of a working model, and apply FBA-Gap to detect the resulting gaps in the network.

## 2. Materials and Methods

FBA-Gap takes as input an FBA model and a lower bound for the flux through the artificial biomass reaction (to ensure growth). Whereas FBA can be considered a generalized maximum flow on a hypergraph, consider an analogy with maximum flows on graphs (Figure 1(a)). Intuitively, a gap corresponds to a missing arc. The main idea behind FBA-Gap is to find a minimum-cost set of artificial exchange reactions so that biomass may be produced. Note that for the graph in Figure 1(b), artificially adding flow to any of nodes $C$, $D$, or $E$ will ensure positive flow along the artificial arc. Given that we would like to fill the gap, we would benefit the most by knowing the needed exchange reaction that is furthest from the biomass reaction. This desire leads us to define a notion of *distance from the biomass reaction* and a corresponding cost structure that will lead us to the gaps.

*Integer Programming Model.* Let

$$x_i = \begin{cases} 1 & \text{if a source reaction is added for metabolite } i \\ 0 & \text{o.w.} \end{cases}$$
$$y_i = \begin{cases} 1 & \text{if an escape reaction is added for metabolite } i \\ 0 & \text{o.w.} \end{cases} \quad (4)$$

for $i \in R$. Then a minimum-cost set of exchange reactions for which a minimum threshold of flux through the biomass is attained can be determined by solving the following mixed-integer program:

$$\begin{aligned} \min \quad & (c^{\text{src}})^T x + (c^{\text{esc}})^T y, \\ \text{s.t.} \quad & Sv + b^{\text{src}} - b^{\text{esc}} = 0, \\ & L \le v \le U, \\ & (L^{\text{src}})^T x \le b^{\text{src}} \le (U^{\text{src}})^T x, \\ & (L^{\text{esc}})^T y \le b^{\text{esc}} \le (U^{\text{esc}})^T y. \end{aligned} \quad (5)$$

Note that a positive lower bound for the biomass reaction, $L_{\text{biomass}}$, is specified in the set of flux lower bounds. The first constraint ensures that a valid flux distribution is derived, that is, the mass balance constraints are satisfied. The last two constraints ensure that if the flux along a exchange reaction is positive, then an appropriate cost is enforced. The remaining constraint contains bounds for the reactions fluxes. Solving (5) is shown to be *NP-Complete* (see in the Supplementary Material available online at doi:10.1155/2012/323472). The selection of exchange metabolites that are most biologically plausible and/or furthest from the biomass reaction is ensured by a cost structure that is described in the next section.

*Cost Structure for Exchange Reactions.* First, we assign costs to extracellular metabolites. Extracellular metabolites are

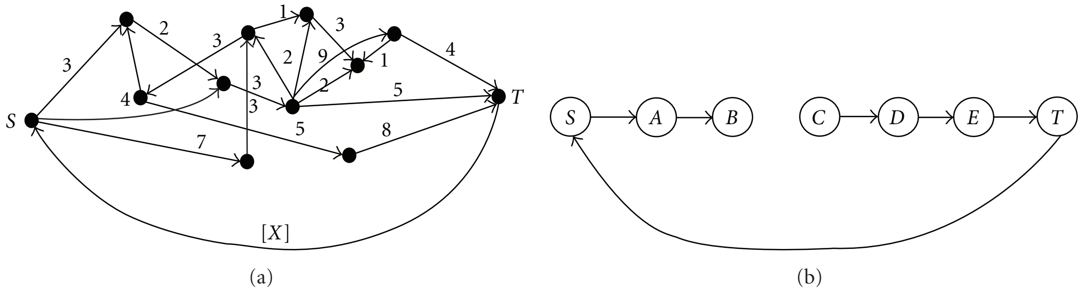


Figure 1: (a) An illustration of a maximum flow problem on a graph. The numbers above the arcs are capacities, and we wish to maximize flow from the source $S$ to the sink $T$; equivalently, we wish to maximize flow along the artificial arc $(T, S)$ such that the flow at each node is balanced. (b) An example of a small maximum flow problem with a gap such that no flow along arc $(T, S)$ is possible.

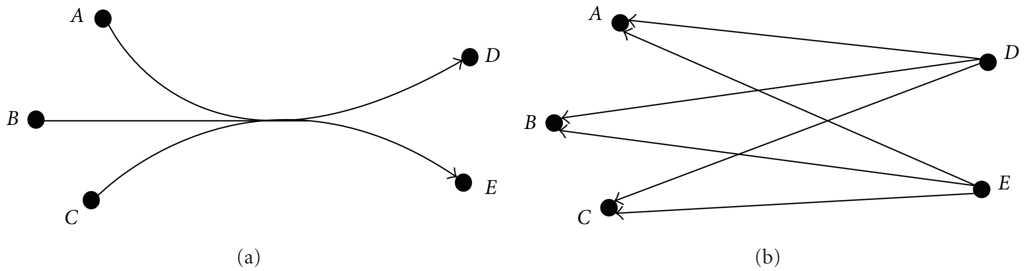


Figure 2: (a) An example of a hyperarc in the hypergraph H corresponding to a reaction $A + B + C \rightarrow D + E$, and (b) the corresponding arcs in $G$, the graph used for calculating distances.

substances that are either postulated to exist in the culture medium or are secreted by the cell. Adding a source reaction for such a metabolite is plausible if experimental culture media that support growth are likely to contain the substance, and adding an escape reaction is plausible if the cell likely secretes the metabolite. A low cost of 1 is assigned for biologically plausible exchange reactions for extracellular metabolites, and a cost of 20 is assigned for implausible exchange reactions for extracellular metabolites (Table S1). Therefore, up to 20 plausible exchange reactions will be selected before 1 implausible exchange reaction. The costs for artificial exchange reactions for internal metabolites are assigned based on the distance of a metabolite to the biomass reaction. Distance to the biomass reaction is defined as follows. Assume for the moment that all stoichiometric coefficients are 1. Let $H = (M, R)$ where $R \subseteq 2^{|M|} \times 2^{|M|}$ is the directed hypergraph associated with the reaction network for an organism, where each reaction corresponds to a hyperarc (Figure 2(a)). Define a directed graph $G = (M, \mathscr{R})$ as follows. For every hyperarc $r \in R$ with tail nodes $T_r$ and head nodes $H_r$, and for every $i \in T_r$ and $j \in H_r$, there is an arc $(j, i) \in \mathscr{R}$ (Figure 2(b)). An intuitive definition for the distance of a metabolite to the biomass reactants (products) is the minimum length of a directed path in $G$ from the metabolite to a biomass reactant (product). This distance measure does not work well because, for example, a large proportion of reactions in a cell involve cofactors such as ATP. Every metabolite is either involved in a reaction where ATP is produced or consumed or will be near such a reaction by this distance measure. Therefore, every metabolite will appear to be near the biomass reaction.

To remedy this effect, we penalize paths that pass through these often-occurring cofactor metabolites. Instead of measuring graph distance by the number of arcs, we define the distance along an arc $(i, j)$ in $\mathscr{R}$ by $d(i, j) = \deg(i)$, where $\deg(i)$ is the degree of $i$. Note that the degree of node $i$ in $G$ is precisely the number of reactions in which metabolite $i$ participates. The distance of a metabolite to the biomass reactants $d_i^{\text{src}}$ is the length of the shortest directed path in $G$ to a biomass reactant, which can be determined by applying Dijkstra's algorithm [22]. The analogous distance to the biomass products is denoted $d_i^{\text{esc}}$. Let $d_{\max}^{\text{src}}$ ($d_{\max}^{\text{esc}}$) be the maximum distance among all metabolites with a directed path to the biomass reactants (products) in $G$. To penalize the source transport reactions that are near the biomass reactants, we define the cost for internal metabolite $i$ to be $d_{\max}^{\text{src}} - d_i^{\text{src}} + 20$. The penalty of 20 in the cost formula ensures that the cost of an artificial exchange reaction for an internal metabolite is at least as high as the cost for an exchange reaction for an extracellular metabolite. Escape reactions are penalized in an analogous fashion. Dijkstra's shortest path algorithm is polynomial time and is computationally easy for the networks considered here. The computational complexity of the proposed method is dominated by solving instances of (5). We note here that our cost structure is less likely to find gaps involving ubiquitous but important backbone metabolites. However, the ubiquity of these metabolites in reactions that are already in the draft model indicates that they are unlikely to be responsible for a lack of biomass production in the *in silico* organism. We choose to penalize the inclusion of cofactors rather than simply removing them from the directed graph because determining which metabolites are cofactors can present a challenge [13]. The focus of the proposed method is on creating a high-confidence model that produces biomass, even if "secondary" pathways are involved; subsequent analyses with a working FBA model can help to identify remaining gaps in primary pathways.

*Applying FBA-Gap to Broken Models.* The trivial solution of zero flux on all reactions, including the biomass reaction, is always feasible for (3). A broken model is one for which the optimal objective value for (3) is lower than desired. The process of applying FBA-Gap to a broken model involves three stages: calculating distances to the biomass reaction, reviewing the output of FBA-Gap, and systematically adding reactions from a universal database. In the first stage, Dijkstra's algorithm is applied as described in the previous section to determine the distances of metabolites to biomass products and reactants. The distances are initialized to be infinite. If after application of the shortest path algorithm, the distance of a metabolite to biomass reactants, is still infinite, then there is no path in $G$ to the biomass reactants for that metabolite, and the metabolite will never be selected by solving (5) as having a source reaction. By construction of $G$, there is no sequence of reactions in $H$ beginning with a reaction that produces $i$ and ending with a reaction that produces a biomass reactant. Therefore, adding a source reaction for $i$ will only increase the objective value of (5) without helping to increase biomass production. Similarly, a metabolite $i$ with $d_i^{\text{esc}} = \infty$ after application of the shortest path algorithm will never be selected by solving (5) as having an escape reaction. The corresponding binary variables in (5) for these metabolites can be fixed to zero to reduce computation time. Further, the knowledgeable modeler can review this list to find metabolites that are known to be involved in the production of biomass and fill in gaps along known pathways.

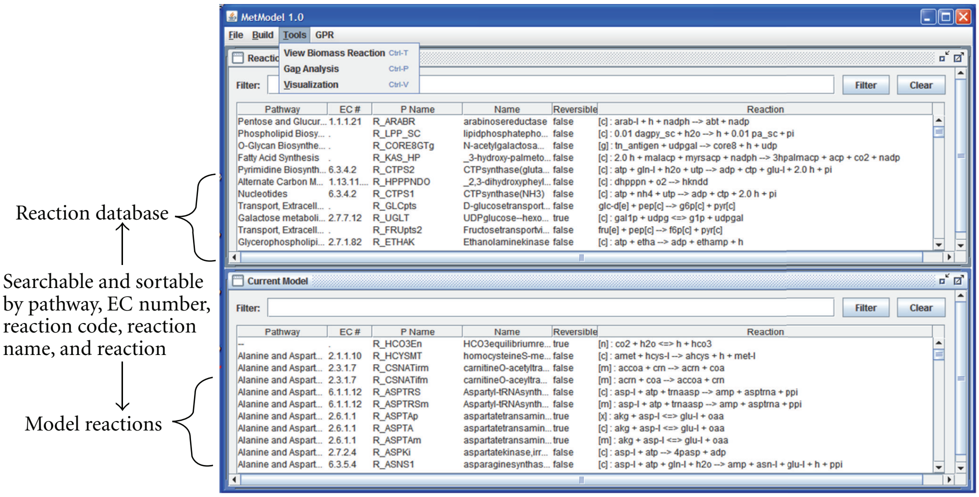
The next stage includes the solution of the integer program (5). The problem is NP-hard and is related to the closed hemisphere problem (see Supplementary Material), suggesting that heuristics for the latter may be adapted to solve challenging instances. Action can be taken to reduce the computational time of solving the integer program directly. The modeler can fix the binary variables to 0 or 1 corresponding to exchange reactions that should not be eligible for selection and exchange reactions that should be selected, respectively. This feature can be used to allow the modeler to specify a particular carbon source for the cell or facilitate discovery of solutions corresponding to secretion of a particular substance. Specifying certain exchange reactions is analogous to determining the list of exchanges to be "tried" as in the MetaFlux procedure [9]. In FBA-Gap, if too many binary variables are fixed to zero, there is a risk that the integer program becomes infeasible. If a feasible solution exists, the output includes a set of exchange reactions, fluxes on those exchange reactions, and fluxes for all other reactions in the network that will provide the desired flux through the biomass reaction. If no feasible solution exists, then the minimum biomass flux must be reduced and/or the bounds on reaction fluxes in the network need to be expanded.

If a feasible solution contains only biologically plausible exchange reactions for extracellular metabolites, then the source reactions can indicate components of a culture medium for the organism. Biologically plausible exchange reactions are those that are for metabolites likely to exist in the culture medium and transportable across the cell membrane. If a feasible solution includes exchange reactions for internal metabolites (e.g., metabolites in the cytoplasm), then the selected exchange reactions give an indication of the location of gaps in the reaction network. The modeler can then consult the appropriate diagram in a publicly available biochemical pathway database, for example, KEGG [23], BioCyc [24], or Reactome [25]. The search for missing reactions is facilitated by the authors' software MetModel GUI (Figure 3). The software includes a searchable and sortable database of metabolic reactions that can be added to a model, as well as capabilities for searching the reactions in a user's model. After adding new reactions to the model, the integer program (5) is resolved and the process of adding reactions is repeated until the model uses only low-cost exchange reactions. A flowchart of the steps in FBA-GAP is depicted in Figure 4.

*Small Example.* Consider the reaction network depicted in Figure 5. Table 1 contains the distance-to-biomass calculation. Metabolites $B, F, G, H$, and $I$ will not be selected as having source reactions, and metabolites $A, B, C, D, E$, and $H$ will not be selected as having escape reactions because they are infinite distance from the biomass reactants and products, respectively. The instance of (5) would be

$$
\begin{aligned}
\min \quad & x_A + 20x_C + 21x_D + 21x_E + 21y_F + 20y_G + y_I, \\
\text{s.t.} \quad & -v_{A\to E} + b_A^{\text{src}} = 0, \\
& -v_{C\to D} + b_C^{\text{src}} = 0, \\
& -v_{DE\to F} + v_{C\to D} + b_D^{\text{src}} = 0, \\
& -v_{DE\to F} + v_{A\to E} + b_E^{\text{src}} = 0, \\
& -v_{F\to GI} + v_{DE\to F} - b_F^{\text{esc}} = 0, \\
& v_{F\to GI} - b_G^{\text{esc}} = 0, \\
& v_{F\to GI} - b_I^{\text{esc}} = 0, \\
& 0 \le b_i^{\text{src}} \le U_i^{\text{src}} x_i, \quad i \in \{A, C, D, E\}, \\
& 0 \le b_i^{\text{esc}} \le U_i^{\text{esc}} y_i, \quad i \in \{F, G, I\}, \\
& L \le v \le U,
\end{aligned}
\tag{6}
$$

with the additional restriction that the variables $x_i$ and $y_i$ are binary. Included in the last set of constraints is anonzero



| Pathway | EC # | P Name | Name | Reversible | Reaction |
|---|---|---|---|---|---|
| Pentose and Glucur... | 1.1.1.21 | R_ARABR | arabinosereductase | false | [c] : arab-l + h + nadph --> abt + nadp |
| Phospholipid Biosy... | . | R_LPP_SC | lipidphosphatepho... | false | [c] : 0.01 dagpy_sc + h2o --> h + 0.01 pa_sc + pi |
| O-Glycan Biosynthe... | . | R_CORE8GTg | N-acetylgalactosa... | false | [g] : tn_antigen + udpgal --> core8 + h + udp |
| Fatty Acid Synthesis | . | R_KAS_HP | _3-hydroxy-palmeto... | false | [c] : 2.0 h + malacp + myrsacp + nadph --> 3hpalmacp + acp + co2 + nadp |
| Pyrimidine Biosynth... | 6.3.4.2 | R_CTPS2 | CTPsynthase(gluta... | false | [c] : atp + gln-l + h2o + utp --> adp + ctp + glu-l + 2.0 h + pi |
| Alternate Carbon M... | 1.13.11.... | R_HPPPNDO | _2,3-dihydroxypheyl... | false | [c] : dhpppn + o2 --> hkndd |
| Nucleotides | 6.3.4.2 | R_CTPS1 | CTPsynthase(NH3) | false | [c] : atp + nh4 + utp --> adp + ctp + 2.0 h + pi |
| Transport, Extracell... | . | R_GLCpts | D-glucosetransport... | false | glc-d[e] + pep[c] --> g6p[c] + pyr[c] |
| Galactose metaboli... | 2.7.7.12 | R_UGLT | UDPglucose--hexo... | true | [c] : gal1p + udpg <=> g1p + udpgal |
| Transport, Extracell... | . | R_FRUpts2 | Fructosetransportvi... | false | fru[e] + pep[c] --> f6p[c] + pyr[c] |
| Glycerophospholipi... | 2.7.1.82 | R_ETHAK | Ethanolaminekinase | false | [c] : atp + etha --> adp + ethamp + h |

| Pathway | EC # | P Name | Name | Reversible | Reaction |
|---|---|---|---|---|---|
| -- | . | R_HCO3En | HCO3equilibriumre... | true | [n] : co2 + h2o <=> h + hco3 |
| Alanine and Aspart... | 2.1.1.10 | R_HCYSMT | homocysteineS-me... | false | [c] : amet + hcys-l --> ahcys + h + met-l |
| Alanine and Aspart... | 2.3.1.7 | R_CSNATirm | carnitineO-acetyltra... | false | [m] : accoa + crn --> acrn + coa |
| Alanine and Aspart... | 2.3.1.7 | R_CSNATifm | carnitineO-aceyltra... | false | [m] : acrn + coa --> accoa + crn |
| Alanine and Aspart... | 6.1.1.12 | R_ASPTRS | Aspartyl-tRNAsynth... | false | [c] : asp-l + atp + trnaasp --> amp + asptrna + ppi |
| Alanine and Aspart... | 6.1.1.12 | R_ASPTRSm | Aspartyl-tRNAsynth... | false | [m] : asp-l + atp + trnaasp --> amp + asptrna + ppi |
| Alanine and Aspart... | 2.6.1.1 | R_ASPTAp | aspartatetransamin... | true | [x] : akg + asp-l <=> glu-l + oaa |
| Alanine and Aspart... | 2.6.1.1 | R_ASPTA | aspartatetransamin... | true | [c] : akg + asp-l <=> glu-l + oaa |
| Alanine and Aspart... | 2.6.1.1 | R_ASPTAm | aspartatetransamin... | true | [m] : akg + asp-l <=> glu-l + oaa |
| Alanine and Aspart... | 2.7.2.4 | R_ASPKi | aspartatekinase,irr... | false | [c] : asp-l + atp --> 4pasp + adp |
| Alanine and Aspart... | 6.3.5.4 | R_ASNS1 | asparaginesynthas... | false | [c] : asp-l + atp + gln-l + h2o --> amp + asn-l + glu-l + h + ppi |

FIGURE 3: Screenshot of MetModel GUI, software for building FBA models. The top frame contains the universal reaction database, and the bottom frame contains the set of reactions in the current working model.

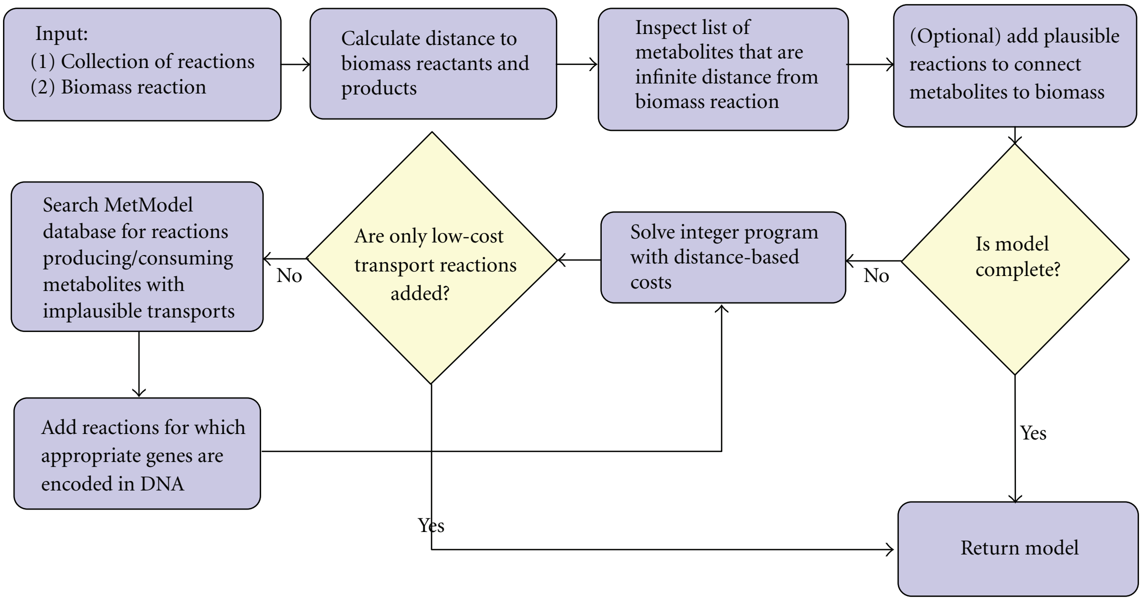

Input:
(1) Collection of reactions
(2) Biomass reaction
Calculate distance to biomass reactants and products
Inspect list of metabolites that are infinite distance from biomass reaction
(Optional) add plausible reactions to connect metabolites to biomass
Is model complete?
No
Solve integer program with distance-based costs
Are only low-cost transport reactions added?
No
Search MetModel database for reactions producing/consuming metabolites with implausible transports
Add reactions for which appropriate genes are encoded in DNA
Yes
Yes
Return model

FIGURE 4: Flowchart of steps in building metabolic reconstructions using FBA-Gap. Determining if a model is complete involves checking if biomass is produced using only biologically plausible exchange reactions.

lower bound on $v_{DE \to F}$ which requires biomass production. The algorithm selects an artificial source reaction for metabolite $C$ and an artificial escape for metabolite $G$. A source reaction for $C$ is selected rather than a reaction for $D$, because $C$ is further from the biomass reaction and therefore the cost is less. The selected source reactions will indicate to the modeler that reactions $B \leftrightarrow C$ and $G \leftrightarrow H$ are missing from the model. The reactions can be found hypothetically by searching through the database in MetModel GUI (Figure 3) or by searching the relevant pathway in another database. Adding these reactions and solving the new instance of (5) produces a solution that indicates that biologically plausible exchange reactions can be added for $A$, $B$, $I$, and $H$ in order to produce biomass.

## 3. Results

*Application to a Partial Metabolic Reconstruction.* To illustrate the ability of FBA-Gap to aid in the construction of new FBA models, we apply the methodology to a new multicompartment model for *Cryptococcus neoformans*. *C. neoformans* is a fungus that can cause meningitis in humans.

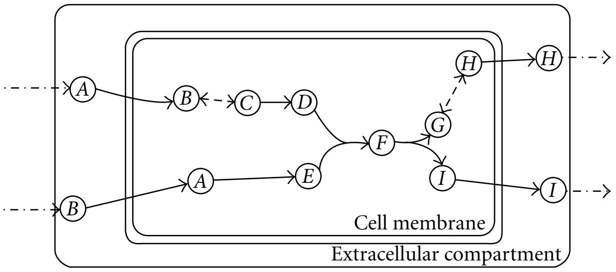


Figure 5: An example of a "broken" FBA model. The biomass reaction is indicated by a bold line, reactions included in the model are indicated by solid lines, reactions omitted from the model (gaps) are indicated by dashed lines, and plausible exchange reactions omitted from the model are indicated by dotted/dashed lines.

Table 1: Distances of metabolites to biomass reactants and biomass products for the network depicted in Figure 5.

| Distance to biomass reactants | Distance to biomass products |
|---|---|
| $d_A^{\text{src}} = 1$ | $d_A^{\text{esc}} = \infty$ |
| $d_B^{\text{src}} = \infty$ | $d_B^{\text{esc}} = \infty$ |
| $d_C^{\text{src}} = 1$ | $d_C^{\text{esc}} = \infty$ |
| $d_D^{\text{src}} = 0$ | $d_D^{\text{esc}} = \infty$ |
| $d_E^{\text{src}} = 0$ | $d_E^{\text{esc}} = \infty$ |
| $d_F^{\text{src}} = \infty$ | $d_F^{\text{esc}} = 0$ |
| $d_G^{\text{src}} = \infty$ | $d_G^{\text{esc}} = 1$ |
| $d_H^{\text{src}} = \infty$ | $d_H^{\text{esc}} = \infty$ |
| $d_I^{\text{src}} = \infty$ | $d_I^{\text{esc}} = 1$ |

Because no metabolic reconstruction of *C. neoformans* has been previously carried out, we assign a generic biomass reaction previously used for *B. subtilis* [26] using only central metabolites that occur in the cytosol:

$$\begin{aligned}
&1.241\,3\,\text{pg} + 2.097\,\text{AcCoa} + 1.236\,\text{Akg} + 35.115\,\text{ATP} \\
&\quad + 0.397\,\text{e4p} + 0.428\,\text{g3p} + 0.712\,\text{g6p} + 0.542\,\text{Gly} \\
&\quad + 14.405\,\text{NADPH} + 8.066\,\text{NH}_4 + 1.785\,\text{oaa} + 0.642\,\text{pep} \\
&\quad + 1.640\,\text{Pi} + 2.994\,\text{Pyr} + 0.445\,\text{r5p} \\
&\quad + 0.262\,\text{Ser-l} + 0.195\,\text{SO}_4 \\
&\quad \longrightarrow 2.852\,\text{CO}_2 + 3.015\,\text{NADH}.
\end{aligned} \tag{7}$$

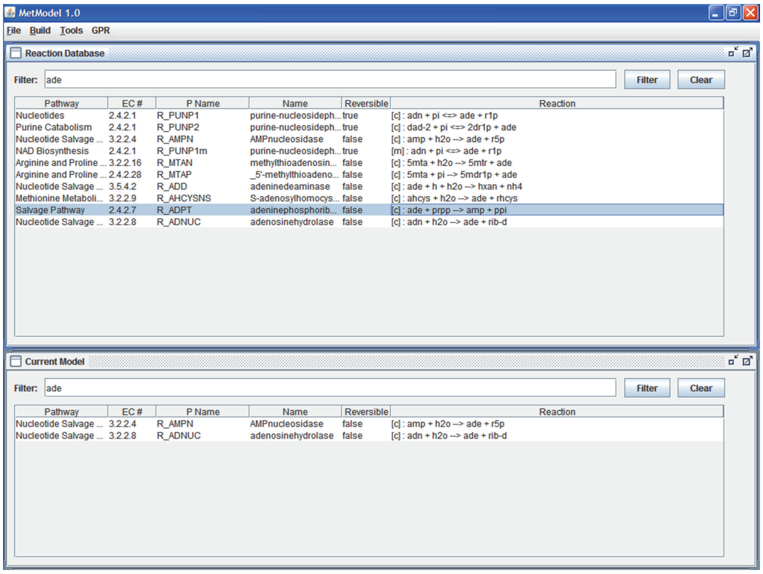
We begin with a partial reconstruction based on evidence from genome annotations and scientific literature. The ability of the model to produce biomass is not considered during this step. The initial model consists of 576 reactions and 712 metabolites with compartments corresponding to the cytosol, mitochondria, and peroxisome. This initial curation was carried out over several weeks. We then solve (5) to find gaps in the model with a time limit of 120 seconds. The MetModel GUI database and KEGG are explored to find reactions that fill the gaps by producing and consuming metabolites with artificial exchange reactions. An internal reaction is added to the model only if it is present in the MetModel GUI database and if KEGG specifies that a gene encodes an enzyme that catalyzes the reaction in the organism. After adding reactions, (5) is solved again, and additional reactions are added. The process is repeated as long as plausible reactions can be added.

Four rounds of solving (5) and manually adding reactions from the MetModel GUI database/KEGG are conducted (Tables S4–S7). Figure 6 illustrates how the searchable and sortable reaction database in MetModel GUI facilitates adding high-confidence reactions. After the first round of solving (5), cytoplasmic cyclic AMP is selected for an artificial source reaction. Browsing the MetModel GUI database reveals that the only reaction producing cyclic AMP, R_ADNCYC, is already included in the model. Upon inspection of the pathway containing R_ADNCYC, we discover an adjacent reaction, R_ADNK1, that is missing from the model and may be added because the corresponding enzyme is encoded in the *C. neoformans* genome. Cytoplasmic coenzyme A is also selected for an artificial source reaction. Inspection of the initial partial reconstruction reveals that much of fatty acid metabolism is omitted. Rather than adding the new pathways, we leave the artificial transport reaction as a placeholder. The artificial and plausible exchange reactions in Table S7 are sufficient to create a working model. In the final round of gap analysis, nine plausible source reactions, seven plausible escape reactions, and an artificial transport reaction for cytoplasmic coenzyme A are added to the model. The first round of solving (5) is terminated at the time limit of 120 seconds with a feasible solution. The remaining three rounds take less than one second to find a provably optimal solution. For a given metabolite, MetModel GUI instantly returns a list of potential reactions. For metabolites involved in reactions that form the backbone of the metabolic network, the list is typically short. Combined with searching the KEGG database, a round of FBA-Gap takes around ten minutes, and the completion of the partial reconstruction for *C. neoformans* takes around one hour.

*Comparison to Other Gap-Filling Algorithms.* In this section, FBA-Gap is compared to GapReed [8] and GapFind/GapFill [6] to demonstrate differences in results and computation time using the *C. neoformans* partial reconstruction. GapReed and GapFind/GapFill are implemented with a universal database curated from existing metabolic reconstructions.

When using GapFind/GapFill in these experiments, GapFind is applied to find all nonproducible metabolites in a model, and GapFill is applied to each nonproducible metabolite to find reactions to add so that the metabolite is produced. If GapFill is infeasible, an exchange reaction is added for that metabolite. Integer programming instances for all methods are solved using Gurobi (http://www.gurobi.com/) with a time limit of 600 seconds.

GapFind determines that only nine metabolites are producible and therefore there are 703 downstream unproducible metabolites. Solving the integer program for



Reaction Database

Filter: ade

| Pathway | EC # | P Name | Name | Reversible | Reaction |
|---|---|---|---|---|---|
| Nucleotides | 2.4.2.1 | R_PUNP1 | purine-nucleosideph... | true | [c] : adn + pi <=> ade + r1p |
| Purine Catabolism | 2.4.2.1 | R_PUNP2 | purine-nucleosideph... | true | [c] : dad-2 + pi <=> 2dr1p + ade |
| Nucleotide Salvage ... | 3.2.2.4 | R_AMPN | AMPnucleosidase | false | [c] : amp + h2o --> ade + r5p |
| NAD Biosynthesis | 2.4.2.1 | R_PUNP1m | purine-nucleosideph... | true | [m] : adn + pi <=> ade + r1p |
| Arginine and Proline ... | 3.2.2.16 | R_MTAN | methylthioadenosin... | false | [c] : 5mta + h2o --> 5mtr + ade |
| Arginine and Proline ... | 2.4.2.28 | R_MTAP | _5'-methylthioadeno... | false | [c] : 5mta + pi --> 5mdr1p + ade |
| Nucleotide Salvage ... | 3.5.4.2 | R_ADD | adeninedeaminase | false | [c] : ade + h + h2o --> hxan + nh4 |
| Methionine Metaboli... | 3.2.2.9 | R_AHCYSNS | S-adenosylhomocys... | false | [c] : ahcys + h2o --> ade + rhcys |
| Salvage Pathway | 2.4.2.7 | R_ADPT | adeninephosphorib... | false | [c] : ade + prpp --> amp + ppi |
| Nucleotide Salvage ... | 3.2.2.8 | R_ADNUC | adenosinehydrolase | false | [c] : adn + h2o --> ade + rib-d |

Current Model

Filter: ade

| Pathway | EC # | P Name | Name | Reversible | Reaction |
|---|---|---|---|---|---|
| Nucleotide Salvage ... | 3.2.2.4 | R_AMPN | AMPnucleosidase | false | [c] : amp + h2o --> ade + r5p |
| Nucleotide Salvage ... | 3.2.2.8 | R_ADNUC | adenosinehydrolase | false | [c] : adn + h2o --> ade + rib-d |

Figure 6: Searching for a gap-filling reaction is facilitated by a searchable and sortable database in MetModel GUI software. The initial model is unable to process $\text{ade}^c$. The reaction database in MetModel GUI has 10 reactions that involve $\text{ade}^e$. The KEGG database indicates that an enzyme for R_ADPT is encoded in the genome for *C. neoformans* and can be added to the model with confidence.

GapFind and the 703 integer programs for GapFill takes 50,169 seconds (about 14 hours). GapFill adds 550 reactions from the reaction database and 182 exchange reactions for metabolites.

The integer program for GapReed terminates at the time limit of 600 seconds. Source exchange reactions are added for nine cytosolic metabolites, one escape exchange reaction is added, and one reaction from the reaction database is added. The source exchange reactions are for cytoplasmic Gln-L, $SO_4$, FDP, $O_2$, NADPH, Gly, PRPP, ATP, and Acetyl-CoA. The escape reaction is for cytoplasmic $CO_2$, and the added reaction is the peroxidative reaction catalyzed by catalase ($2H_2O_2 \rightarrow 2H_2O + O_2$).

GapFind/GapFill and GapReed are more conducive to an automated implementation than FBA-Gap, but in this example, one can see some of the pitfalls of an automated approach. GapFind/GapFill adds many internal reactions and exchange reactions for cytosolic metabolites so that there is a high probability that implausible reactions are present in the final model. Further, GapFind/GapFill requires significantly more computation time. GapReed adds exchange reactions for more implausible cytosolic metabolites than FBA-Gap. A hybrid computational/manual curation approach such as FBA-Gap is able to derive a biomass-producing model with higher-confidence reactions for our partial reconstruction for *C. neoformans* than these two established methods.

*Application to Existing Models.* FBA-Gap is applied to four existing models with exchange reactions removed. The metabolic reconstructions used in this experiment are for *Trypanosoma cruzi* [27], *Bacillus subtilis* [26], *Heliocbacter pylori* (iIT341 GSM/GPR) [2], and *Escherichia coli* (iJR904 GSM/GPR) [28]. The results of applying the procedure provide a hypothesis for a defined culture medium for each organism.

The hypothesized culture media are summarized in Table S2. *T. cruzi* is a protozoan parasite of humans that causes Chagas disease. The reconstruction is a multi-compartment model of central metabolism for *T. cruzi*. FBA-Gap selects biologically realistic source and escape reactions. The source reactions correspond to the transport of extracellular metabolites that are plausible constituents of a culture medium for sustaining *T. cruzi* and the escape reactions correspond to metabolites that are likely secreted by *T. cruzi*. *B. subtilis* is a Gram-positive bacterium found in soil. As with the *T. cruzi* model, FBA-Gap selects biologically realistic exchange reactions for production of biomass. For the *E. coli* reconstruction, one undesirable escape reaction is selected ($\text{clpn\_ec}^c$) that is unique to the *E. coli* model. This metabolite occurs in only one reaction in the model and is a reactant in the biomass reaction. Therefore, a simple remedy is to increase the stoichiometric coefficient in the biomass reaction. For the *h. pylori* reconstruction, a single implausible escape reaction is selected for $\text{rhcys}^c$. Upon investigation of the network around this metabolite, we discover that there are reactions converting $\text{rhcys}^c$ to $\text{dhptd}^c$ and $\text{dhptd}^c$ to $\text{hmfurn}^c$, but there are no reactions consuming $\text{hmfurn}^c$. There are no reactions consuming $\text{hmfurn}^c$ in our universal database, so we can either add it as a reactant in the biomass reaction or add an escape reaction to remove it from the cell.

*Recovering Deleted Reactions from an Existing Model.* FBA-Gap is applied to an existing model with internal reactions deleted to evaluate the ability to find the resulting gaps in the network. In this experiment, we deleted a random sample of 222 internal reactions (15% of all reactions) from the *B. subtilis* model (Table S9). Solving (5) for the resulting model takes two seconds. FBA-Gap suggests 15 exchange reactions, all of which are source reactions (Table S10). Because we know which reactions are deleted from the model, we cannot properly evaluate how many of the deleted reactions a modeler would have added based on the suggested exchange reactions. Of the 15 metabolites in the selected artificial exchange reactions, 14 occur in at least one of the deleted reactions (93%), indicating that FBA-Gap does find the backbone metabolites that are next to the gaps. Of the 222 deleted reactions, 17 of them contain metabolites that are selected for artificial exchange reactions. Not all of the 222 deleted reactions are required to produce biomass, so it is likely that adding a subset of the deleted reactions would be sufficient for the resulting FBA model to produce biomass.

## 4. Discussion

This paper presents an optimization-based method for "debugging" metabolic reconstructions called FBA-Gap. We demonstrate the effectiveness of the procedure in helping to find gaps in a model for *C. neoformans.* FBA-Gap produces a more accurate reconstruction than an application of existing methods for filling gaps and requires less computation time. However, in contrast to other methods, FBA-Gap also involves manually selecting and approving which reactions to add to a model so that the overall time may be longer. As noted by Latendresse et al. [9], a fully automated gap-filling procedure likely leads to significant errors. The motivation behind FBA-Gap is to reduce the manual effort required by allowing the modeler to select from among a few suggested modifications to a model. The distance measure used in pricing artificial exchange reactions helps to indicate the location of gaps; these weights could also be incorporated into a procedure like MetaFlux [9], a more automated procedure that also has the capability of suggesting modifications to the biomass reaction. The FBA-Gap procedure provides hypotheses for defined culture media for organisms based on previously published models. One weakness of FBA-Gap is the computational complexity of solving (5). Finding optimal solutions to these integer programs is NP-Complete, but specialized solution methods may facilitate the computation of good solutions.

## List of Abbreviations

| | |
|---|---|
| FBA: | flux balance analysis |
| LP: | linear program. |

## Acknowledgments

This research is partially supported by NASA Award NNX09AR44A. The authors would like to thank the VCU Center for High Performance Computing for computational infrastructure and support.

## References

[1] N. D. Price, J. L. Reed, and B. Ø. Palsson, "Genome-scale models of microbial cells: evaluating the consequences of constraints," *Nature Reviews Microbiology*, vol. 2, no. 11, pp. 886–897, 2004.

[2] I. Thiele, T. D. Vo, N. D. Price, and B. Ø. Palsson, "Expanded metabolic reconstruction of *Helicobacter pylori* (iIT341 GSM/GPR): an in silico genome-scale characterization of single- and double-deletion mutants," *Journal of Bacteriology*, vol. 187, no. 16, pp. 5818–5830, 2005.

[3] J. S. Edwards, R. U. Ibarra, and B. O. Palsson, "In silico predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data," *Nature Biotechnology*, vol. 19, no. 2, pp. 125–130, 2001.

[4] A. R. Joyce, J. L. Reed, A. White et al., "Experimental and computational assessment of conditionally essential genes in *Escherichia coli*," *Journal of Bacteriology*, vol. 188, no. 23, pp. 8259–8271, 2006.

[5] J. Förster, I. Famili, B. Ø. Palsson, and J. Nielsen, "Large-scale evaluation of in silico gene deletions in *Saccharomyces cerevisiae*," *OMICS A Journal of Integrative Biology*, vol. 7, no. 2, pp. 193–202, 2003.

[6] V. S. Kumar, M. S. Dasika, and C. D. Maranas, "Optimization based automated curation of metabolic reconstructions," *BMC Bioinformatics*, vol. 8, article 212, 2007.

[7] C. S. Henry, M. Dejongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens, "High-throughput generation, optimization and analysis of genome-scale metabolic models," *Nature Biotechnology*, vol. 28, no. 9, pp. 977–982, 2010.

[8] J. L. Reed, T. R. Patel, K. H. Chen et al., "Systems approach to refining genome annotation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 46, pp. 17480–17484, 2006.

[9] M. Latendresse, M. Krummenacker, M. Trupp, and P. D. Karp, "Construction and completion of flux balance models from pathway databases," *Bioinformatics*, vol. 28, pp. 388–396, 2012.

[10] P. Pharkya, A. P. Burgard, and C. D. Maranas, "OptStrain: a computational framework for redesign of microbial production systems," *Genome Research*, vol. 14, no. 11, pp. 2367–2376, 2004.

[11] D. Segrè, J. Zucker, J. Katz et al., "From annotated genomes to metabolic flux models and kinetic parameter fitting," *OMICS A Journal of Integrative Biology*, vol. 7, no. 3, pp. 301–316, 2003.

[12] P. R. Romero and P. Karp, "Nutrient-related analysis of pathway/genome databases," *Pacific Symposium on Biocomputing*, pp. 471–482, 2001.

[13] M. Arita, "Metabolic reconstruction using shortest paths," *Simulation Practice and Theory*, vol. 8, no. 1-2, pp. 109–125, 2000.

[14] D. C. McShan, S. Rao, and I. Shah, "PathMiner: predicting metabolic pathways by heuristic search," *Bioinformatics*, vol. 19, no. 13, pp. 1692–1698, 2003.

[15] T. Gaasterland and E. Selkov, "Reconstruction of metabolic networks using incomplete information," in *Proceedings of the 3rd International Conference on Intelligent Systems in Molecular Biology*, pp. 127–135, 1995.

[16] A. M. Feist and B. Ø. Palsson, "The growing scope of applications of genome-scale metabolic reconstructions using *Escherichia coli*," *Nature Biotechnology*, vol. 26, no. 6, pp. 659–667, 2008.

[17] E. P. Gianchandani, M. A. Oberhardt, A. P. Burgard, C. D. Maranas, and J. A. Papin, "Predicting biological system objectives de novo from internal state measurements," *BMC Bioinformatics*, vol. 9, article 43, 2008.

[18] M. J. Herrgård, S. S. Fong, and B. Ø. Palsson, "Identification of genome-scale metabolic network models using experimentally

measured flux profiles," *PLoS Computational Biology*, vol. 2, no. 7, article e72, 2006.

[19] H. Jeong, B. Tombor, R. Albert, Z. N. Oltval, and A. L. Barabásl, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[20] M. Arita, "The metabolic world of *Escherichia coli* is not small," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 6, pp. 1543–1547, 2004.

[21] P. Kharchenko, L. Chen, Y. Freund, D. Vitkup, and G. M. Church, "Identifying metabolic enzymes with multiple types of association evidence," *BMC Bioinformatics*, vol. 7, article 177, 2006.

[22] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[23] M. Kanehisa and S. Goto, "KEGG: kyoto encyclopedia of genes and genomes," *Nucleic Acids Research*, vol. 28, no. 1, pp. 27–30, 2000.

[24] R. Caspi and P. D. Karp, "Using the MetaCyc pathway database and the BioCyc database collection," *Current Protocols in Bioinformatics*, chapter 1, unit 1.17, 2007.

[25] I. Vastrik, P. D'Eustachio, E. Schmidt et al., "Reactome: a knowledge base of biologic pathways and processes," *Genome Biology*, vol. 8, no. 3, article R39, 2007.

[26] M. Dauner and U. Sauer, "Stoichiometric growth model for riboflavin-producing bacillus subtilis," *Biotechnology and Bioengineering*, vol. 76, no. 2, pp. 132–143, 2001.

[27] S. B. Roberts, J. L. Robichaux, A. K. Chavali et al., "Proteomic and network analysis characterize stage-specific metabolism in *Trypanosoma cruzi*," *BMC Systems Biology*, vol. 3, article 52, 2009.

[28] J. L. Reed, T. D. Vo, C. H. Schilling, and B. O. Palsson, "An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR)," *Genome Biology*, vol. 4, no. 9, p. R54, 2003.