OMICS A Journal of Integrative Biology
Volume 7, Number 3, 2003

From Annotated Genomes to Metabolic Flux Models and Kinetic Parameter Fitting

DANIEL SEGRÈ,[1] JEREMY ZUCKER,[2] JEREMY KATZ,[3] XIAOXIA LIN,[1] PATRIK D'HAESELEER,[1] WAYNE P. RINDONE,[1] PETER KHARCHENKO,[1] DAT H. NGUYEN,[1] MATTHEW A. WRIGHT,[1] and GEORGE M. CHURCH[1]

ABSTRACT

Significant advances in system-level modeling of cellular behavior can be achieved based on constraints derived from genomic information and on optimality hypotheses. For steady-state models of metabolic networks, mass conservation and reaction stoichiometry impose linear constraints on metabolic fluxes. Different objectives, such as maximization of growth rate or minimization of flux distance from a reference state, can be tested in different organisms and conditions. In particular, we have suggested that the metabolic properties of mutant bacterial strains are best described by an algorithm that performs a minimization of metabolic adjustment (MOMA) upon gene deletion. The increasing availability of many annotated genomes paves the way for a systematic application of these flux balance methods to a large variety of organisms. However, such a high throughput goal crucially depends on our capacity to build metabolic flux models in a fully automated fashion. Here we describe a pipeline for generating models from annotated genomes and discuss the current obstacles to full automation. In addition, we propose a framework for the integration of flux modeling results and high throughput proteomic data, which can potentially help in the inference of whole-cell kinetic parameters.

INTRODUCTION

THE ASSESSMENT OF OUR ABILITY to extract and understand relevant biological information from genomes lies in the capacity to build computational models and make predictions that can be tested. We describe here methods for utilizing annotated genomes to construct whole-cell models, as well as analytical tools to process the output of these models. The emphasis is on general strategies that can be applied to many different organisms, taking advantage of the increasing availability of whole genome sequences and databases with biochemical and functional information. Our philosophy is that models should use information piped automatically from public databases and provide feedback about those features of the data itself, or of the way it is stored, which make the model-generation process ambiguous or unfeasible. This approach com-

[1]Lipper Center for Computational Genetics and Department of Genetics, Harvard Medical School, Boston, Massachusetts.
[2]Research Computing Department, Dana-Farber Cancer Institute, Boston, Massachusetts.
[3]Harvard Division of Continuing Education, Cambridge, Massachusetts.

plements the strategy of manually building and curating specific models, which has a more immediate short-term payoff, but less long-range implications for the computational biology community. Such a policy is especially appropriate for the cell modeling effort undertaken in the BioSPICE program (DARPA, 2003), whose goals include the development of modular software tools based on a common modeling language and on transportable model definitions (www.biospice.org).

Computational models at a whole-cell level use a variety of mathematical methods and cover a wide range of resolutions and organisms (Bailey, 2001; Covert et al., 2001; Jamshidi et al., 2001; Tomita, 2001). Our efforts focus on constraint-based models whose specifications can be inferred almost entirely from the genomic, biochemical and structural information distributed in public databases (Kanehisa et al., 2002; Karp et al., 2002). While here we discuss flux balance models of metabolic networks, we are also developing models for chromosome structure-dynamics, based on genome- and experiment-derived distance constraints (Wright et al., 2002). Constraint-based models are useful in several ways: (i) the degrees of freedom of the constrained system provide an indication of our level of understanding, (ii) consistency tests of multiple constraints offer the possibility of critically revising data or our interpretations of data, (iii) computer simulations or optimization algorithms can be applied within constrained spaces to search for solutions that correspond to specific configurations; we are especially interested in optimal configurations as they may capture important properties of evolutionary adaptation.

The constraint-based models discussed here explore cellular metabolism at steady state, based on the framework of Flux Balance Analysis (FBA; Fig. 1) (Varma and Palsson, 1994; Bonarius et al., 1997; Edwards and Palsson, 1998). As opposed to approaches aimed at modeling dynamic behavior through differential equations and/or stochastic simulations, steady state metabolic flux analyses do not rely on the knowledge of kinetic parameters (Bailey, 2001). A metabolic network can be viewed as a system that processes "input metabolites" (nutrients), and produces "output metabolites" (e.g., essential biomass components, such as amino acids, lipids). Intracellular reactions convert these and many other metabolites into each other according to well defined and mostly known molecular proportions (stoichiometric coefficients). Steady state reaction rates constitute fluxes, which can be measured experimentally (Yanagimachi et al., 2001; Fischer and Sauer, 2003). Due to mass conservation for each metabolite, fluxes can be treated as unknowns that are constrained by linear relationships, and cannot vary independently of each other. Additional inequality constraints typically derive from nutrient limitations and thermodynamic considerations (such as irreversibility). The set of all fluxes compatible with these constraints constitutes the feasible space of the metabolic network (Schilling and Palsson, 1998). Optimization methods, such as linear and quadratic programming (Luenberger, 1989), can be used to find, within this space of allowed states, a single state that possibly reflects the actual flux distribution of the analyzed cell under a defined set of nutrient conditions (Covert et al., 2001). An optimization criterion that has been frequently used in modeling bacterial cells is the maximization of growth capacity, defined as a flux that produces biomass components in fixed proportions (Fig. 1) (Schilling et al., 1999)). Metabolic flux methods have proven successful in performing whole cell studies with explanatory and predictive capacity (Edwards et al., 2001; Ibarra et al., 2002; Segrè et al., 2002). Importantly, even in cases where FBA fails to explain experimental data, the ensuing formal treatment of a metabolic network constitutes a powerful tool for representing and refining knowledge. FBA models for different organisms mostly differ in the detailed lists of metabolic reactions for which an enzyme is present. Various microorganisms have been modeled with FBA, such as *Haemophilus influenzae* (Edwards and Palsson, 1999; Schilling and Palsson, 2000), *Escherichia coli* (Edwards and Palsson, 2000a,b) *Helicobacter pylori* (Schilling et al., 2002), *Saccaromices cerevisiae* (Forster et al., 2003), as well as organelles (Ramakrishna et al., 2001).

RESULTS

We discuss here flux balance modeling progress on several fronts. One of the most fascinating aspects of flux balance methods is the possibility of testing various objective functions, each of which could capture a different system-level property of the cell. Along these lines, we have developed the method of minimization of metabolic adjustment (MOMA), which is described in a simple example below. MOMA was

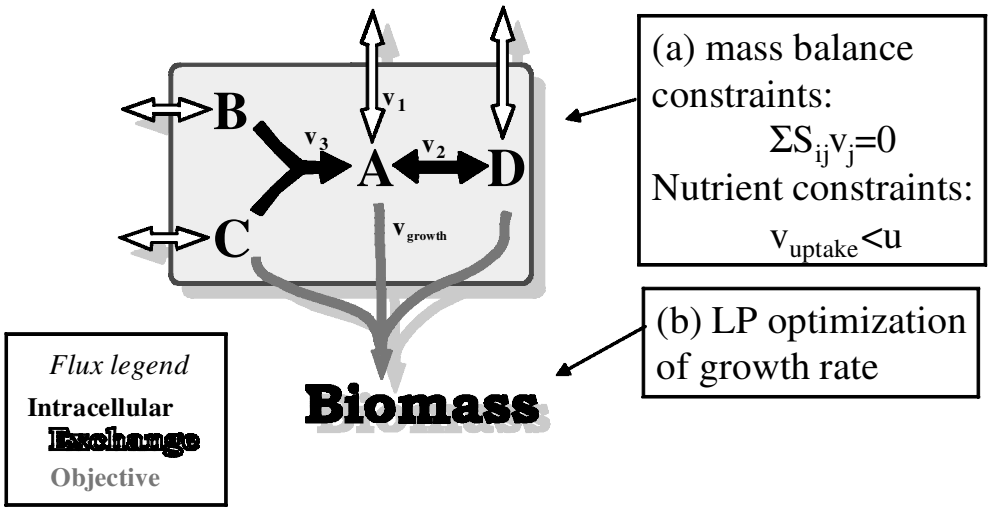


FIG. 1. A schematic description of how a Flux Balance Analysis (FBA) model works. The fundamental constraints derive from mass conservation of metabolites at steady state. One should note that in addition to intracellular fluxes, reactions include transport across the membrane and a growth flux that produces biomass. In general, the mass balance equations involve stoichiometric coefficients S_{ij}. The conditions for a specific *in silico* experiment are defined by choosing bounds for the nutrients' uptake rates. No kinetic parameters are used in FBA modeling. Metabolic phenotypes (i.e., growth rate and flux distributions) are computed using Linear Programming (LP). For example, one can find the flux state that maximizes growth, and test the hypothesis that a bacterium evolved towards the same goal. (For a more complete description of FBA, see Covert et al., 2001.)

initially developed and tested in *E. coli* (Segrè et al., 2002) for the purpose of analyzing the metabolic behavior of mutant strains. This method could be especially useful when comparing *in silico* knockouts with high throughput experimental results (Fischer and Sauer, 2003). In order to incorporate FBA and MOMA tools into BioSPICE (DARPA 2003), we have built appropriate Java Open Agent Architecture (OAA; Martin et al., 1999a) interfaces. A broad application of these tools, however, requires the existence of complete and consistent stoichiometric reconstructions for the organism analyzed. While various metabolic network reconstructions have been published, general criteria and implementations for automated generation of metabolic models from annotated genomes are among the next important steps. We present here preliminary results for an automatic metabolic model reconstruction pipeline. These will provide guidelines and suggestions for data formatting and model testing. Finally, in an attempt to show how the applicability of flux balance models goes beyond steady state predictions, we discuss the possibility of integrating flux balance results and enzyme level measurements to infer whole-cell metabolism kinetic parameters.

Perturbed metabolic networks and the minimization of metabolic adjustment (MOMA)

When biomass production is used as an objective function, the hypothesis tested with FBA is that the cell has evolved towards optimal growth capacity (Edwards et al., 2001; Ibarra et al., 2002). While natural selection is a valid rationale for this hypothesis in the case of wild type organisms, mutant organisms may be expected not to have reached their hypothetical optimal performance. MOMA (Segrè et al., 2002) addresses this issue by testing the hypothesis that a gene deletion causes a minimal flux redistribution with respect to the wild type metabolism, compatibly with the absence of the removed reaction. Quadratic programming (QP) is employed for solving the ensuing distance minimization problem in flux space. The method, illustrated below in a simple example, explores feasible metabolic flux states that are suboptimal with respect to growth. In addition to predicting the metabolic behavior of mutant strains, MOMA could be equally applied to organisms forced to grow under unnatural metabolic or environmental conditions.

To illustrate the differences between FBA and MOMA predictions for a metabolic network subject to a gene deletion, let us consider a simple reaction network as shown in Figure 2A. There are six metabolites

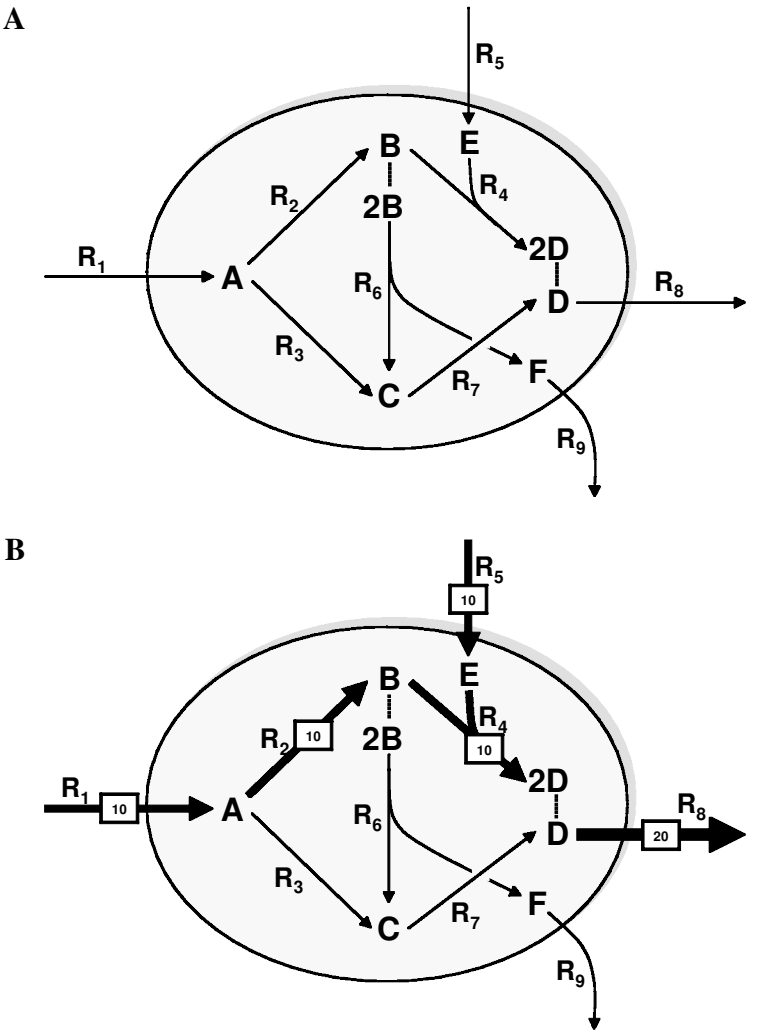


FIG. 2. The methods of Flux Balance Analysis (FBA) and Minimization of Metabolic Adjustment (MOMA) applied to a toy metabolic network. This simple example illustrates the basic principles and consequences of using MOMA for the prediction of the effects of gene knockouts. (**A**) The reaction network contains nine reactions and six metabolites. Arrows represent chemical reactions. Two metabolites (B and D) are represented twice in the network, because they are involved in different reactions with different stoichiometries. These multiple instances of metabolites in the network are connected by dotted lines. (**B**) A flux balance analysis solution for the above described network is superimposed to the network itself. Nonzero flux values are reflected both in the thickness of the arrows, and in the associated labels. The flux balance problem, which maximizes the flux for reaction R_8, is detailed in the text. (**C**) A different solution for the optimization problem, again using FBA. This corresponds to the optimal solution for a simple mutant network, for which the flux of R_4 is forced to zero. (**D**) Here the flux distribution of the mutant network is computed with MOMA, as described in the text and in (Segrè et al., 2002). The resulting fluxes use the same alternative pathway as in 2C, but less efficiently than in the FBA prediction. Resistance to change implies a tendency to keep using portions of the old pathway (reaction R2). A manifestation of this phenomenon has been suggested to appear in *E. coli* mutants.

involved: A, B, C, D, E, and F. A is the limiting input and D is the desired output. B and C are intermediate metabolites; E is an optional and unlimited input; while F is an optional byproduct of the network. The nine reactions are also listed in Table 1. By using LP for maximizing the flux of R_8 (output flux) we obtain the flux distribution shown in Figure 2B. No fluxes are forced to be zero in this case, and therefore we refer to this as the wild-type system. To maximize the production of metabolite D, the most efficient

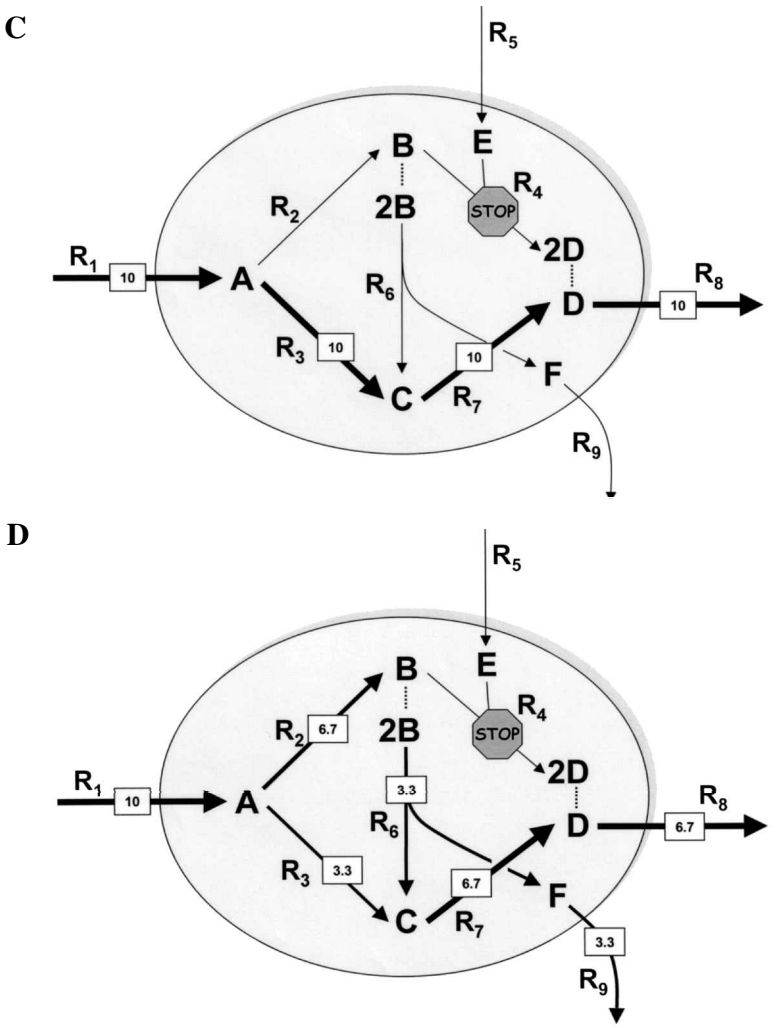


FIG. 2. *Continued.*

pathway goes through R_2 and R_4, and two moles of D are produced for each mole of A that is consumed, through the incorporation of E in R_4. In a simulation of a knockout experiment, we delete the gene that codes for the enzyme catalyzing R_4, so this reaction is removed from the network (its flux is set to zero). FBA still assumes that maximizing the flux of R_8 is the objective and the resulting flux distribution obtained with LP optimization is shown in Figure 2C. In this case, a completely different pathway is used compared to the wide type condition. A is converted into D through R_3 and R_7, and one mole of D is produced for each mole of A that is consumed. The MOMA method assumes that the minimum deviation from the wild type fluxes reflects the response of the cell to perturbations of the reaction network. The corresponding QP minimization leads to the flux distribution shown in Figure 2D. In contrast to the solution obtained from FBA calculation, the mutant system now utilizes (1) part of the pathway used in the wild type network that converts A to B (i.e., R_2); (2) a reaction converting B to C, which can be converted to D (i.e., R_6); and (3) direct conversion of A to C (i.e., R_3). In short, this solution includes a combination of the two pathways utilized in the wild type and in the FBA mutant solution. Note that for each mole of A that is consumed, only 0.67 mole of D is now produced, hence this solution is suboptimal. This reduction of the output flux or yield is a consequence of the low efficiency of pathway R_2-R_6 compared to pathway R_3 (R_6 has a byproduct F besides C).

By comparing FBA and MOMA predictions with experimental data for *E. coli*, we confirmed previous

Table 1. The List of Reactions Used in the Model of Figure 2

R_1: $\rightarrow$ A
R_2: A $\rightarrow$ B
R_3: A $\rightarrow$ C
R_4: B + E $\rightarrow$ 2D
R_5: $\rightarrow$ E
R_6: 2B $\rightarrow$ C + F
R_7: C $\rightarrow$ D
R_8: D $\rightarrow$
R_9: F $\rightarrow$

findings that *E. coli* wild type may have evolved towards maximal growth, and found that mutants are more compatible with the suboptimal performance prediction of MOMA, than with FBA (Segrè et al., 2002). In general, the real flux distribution may lie between the FBA and MOMA solutions, as well as in other unexplored regions of the feasible space. While the FBA solution is likely to overestimate the growth rate, the MOMA solution might be too conservative when considering the capability of the cell to make adjustments in response to perturbations. In general, determining which of the two solutions is more accurate depends on how well the regulatory system of the cell can respond to the perturbation, and on how efficiently alternative pathways in the cell can sustain an unusually high flux.

Flux balance models in BioSPICE

The algorithm implemented in the minimization of metabolic adjustment is described schematically in Figure 3. The BioSPICE (DARPA 2003) graphical user interface for MOMA is written in Java, and enables the user to select several model parameters, run the MOMA Perl script, and read the output, either in a short or in a more detailed version. The optimization software packages currently used by the MOMA Perl script are the GNU Linear Programming Kit (GLPK [Makhorin 2001]) for LP, and the Object Oriented Quadratic Programming (OOQP [Gertz and Wright 2001]) package for QP. The user interface window is shown in Figure 4A. Up to eight genes can be simultaneously knocked out (upper left part of the window). In the upper right portion of the user interface window, the user can choose the upper bounds for the uptake of glucose, nitrogen and oxygen, and decide whether or not isoenzymes of the selected genes for mutations should be knocked out as well. A typical output of the program is shown in Figure 4B. Future versions will include more comprehensive choices of nutrients and simulation options.

We have successfully completed an experiment in representing the most important fluxes in the *E. coli* flux balance model (Edwards and Palsson, 2000a) using a draft version of Level 2 of the Systems Biology Markup Language (SBML [Hucka et al., 2003]). Although SBML was initially developed for purposes of representing kinetic models, we found that by using the existing ListOfReactions structures developed for that purpose, the only information that could not be represented with existing fields was the flux limits associated with each flux enumerated in this ListOfReactions. We were able to present these as annotations for each flux reaction defined in our own unique name space. The version of this SBML level 2 model that has been included in BioSPICE 3.0 (DARPA, 2003) conformed completely to the SBML Level 2 draft that was current until the end of May of 2003. Since then a final revision to the Level 2 standard has been prepared that will require some modifications in our model file. The modified version will be made available at http://arep.med.harvard.edu/ecoliseparategenessbml2.xml. An example of SBML description of a reaction for FBA is shown in the Appendix.

Automated generation of metabolic network models

Currently, genomic data mining needs to be combined with manual curation in order to build *in silico* models of bacteria (Covert et al., 2001). However, a fully automatic *in silico* reconstruction based on an-

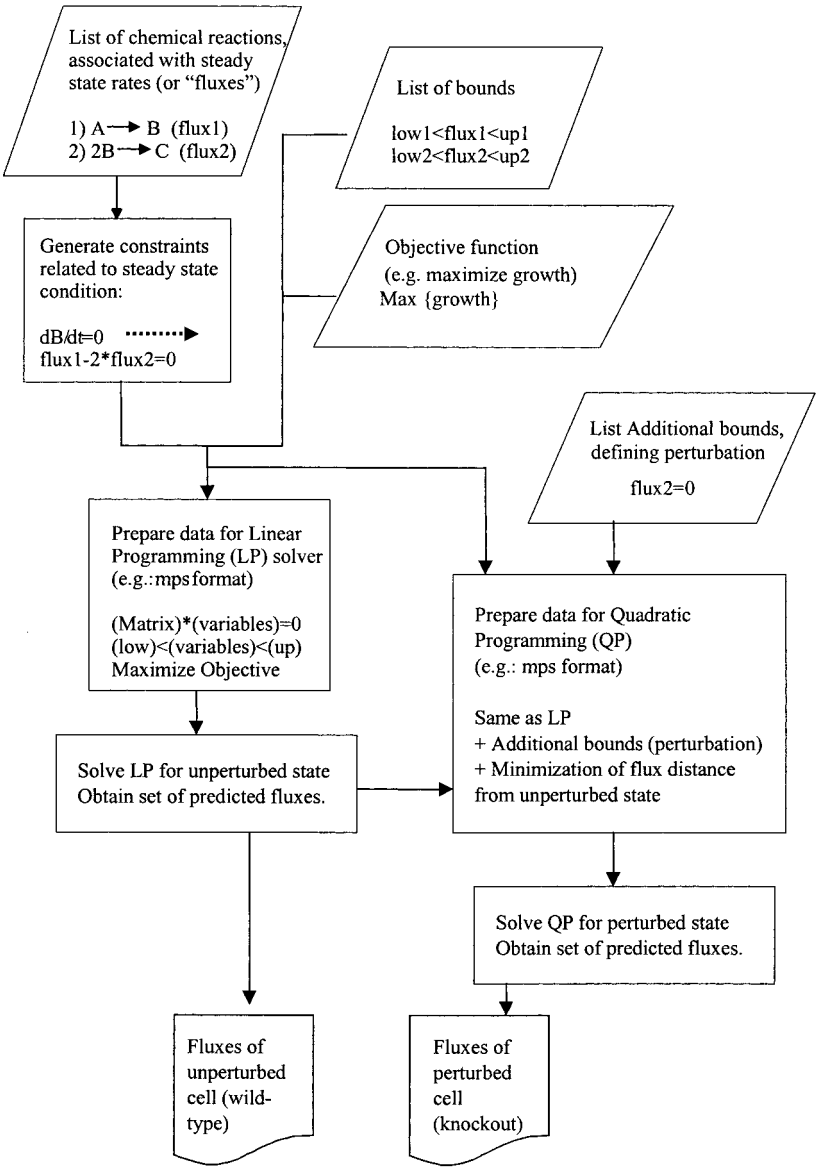
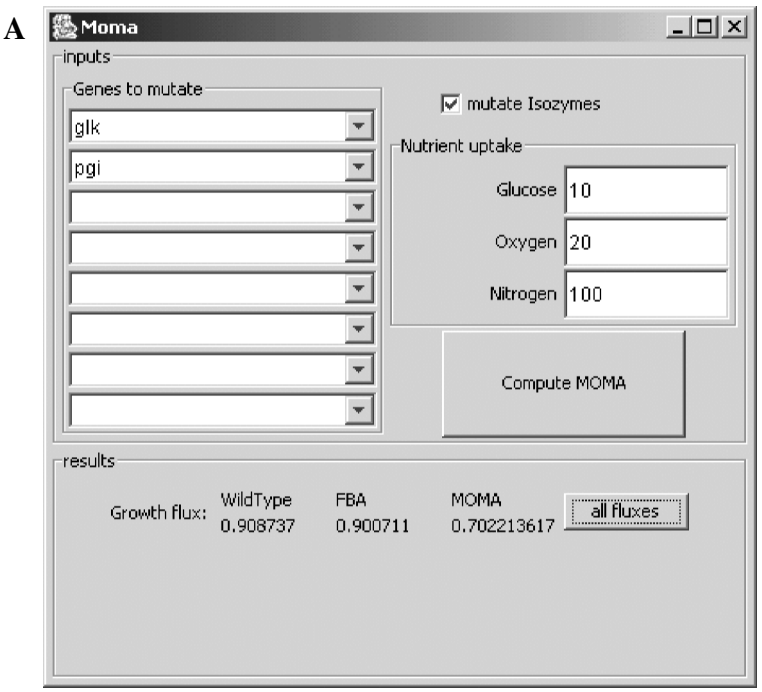


FIG. 3. A schematic flow chart representing the computational steps performed in order to generate predictions of fluxes and growth rates for metabolic networks perturbed by gene knockout. A combination of FBA and MOMA methods is utilized, using linear and quadratic programming, respectively.

notated genomes would make it possible to quickly build flux balance models for a large number of newly sequenced organisms. We assess here the feasibility of a computational stoichiometry reconstruction, and discuss the possibility of using current on-line databases to build a bioinformatics pipeline that enables one to go from a genome sequence to metabolic phenotype predictions. A preliminary implementation of this pipeline, as shown in Figure 5, involves the following steps: (i) start with an annotated genome, (ii) build an organism-specific metabolic database, (iii) query the database to produce the stoichiometric matrix, (iv) apply network debugging algorithms to check the completeness of the matrix, (v) integrate this data with nutrient uptake fluxes, essential biomass constituents, and other strain-specific parameters, and (vi) run an FBA and MOMA solver.

The first step in this pipeline is to convert an annotated genome into an organism specific metabolic data-



B

Solution Fluxes

Flux Name	Wild Type	FBA	MOMA
GLK1	0.0	0.0	0.0
PGI1R	4.49563	0.0	0.0
PGI2R	0.0	0.0	0.0
PGI3R	0.0	0.0	0.0
GALMR	0.0	0.0	0.0
AGP	0.0	0.0	0.0
PFKA	0.0	0.0	2.46335018E-9
PFKB1	7.24105	5.75067	6.41804357
FBP	0.0	0.0	0.9916305
FBAR	7.24105	5.75067	5.42641307
TPIAR	-7.13395	-5.64452	-5.3436572
GAPAR	15.6657	14.1892	14.5449335
GAPCR	0.0	0.0	-1.12506654
PGKR	15.6657	14.1892	13.4198669
GPMAR	14.0573	12.595	13.196837
GPMBR	0.0	0.0	-0.863163047
ENOR	14.0573	12.595	12.3336739
PPSA	0.0	0.0	0.381682441
PYKA	0.608874	0.0	4.43866452E-8
PYKF	0.0	0.0	2.99192635E-9
ACEE	8.2366	7.64868	6.75038565
GLGC	0.139945	0.138709	0.118398013

FIG. 4. A The user interface of the MOMA program in BioSPICE, allowing the simultaneous deletion of any combination of up to eight genes. The interface is extendable to include more deletions and more nutrient uptake options, as well as different organisms. (**B**) A sample output window from the BioSPICE version of MOMA.

base. This step is accomplished to a large degree by using the Pathologic software, part of the Pathway Tools Suite (Karp et al., 2002; Karp et al., 2002), which automatically produces a Pathway/Genome Database (PGDB) containing pathways, reactions, metabolites, enzymes, and genes that can be directly inferred

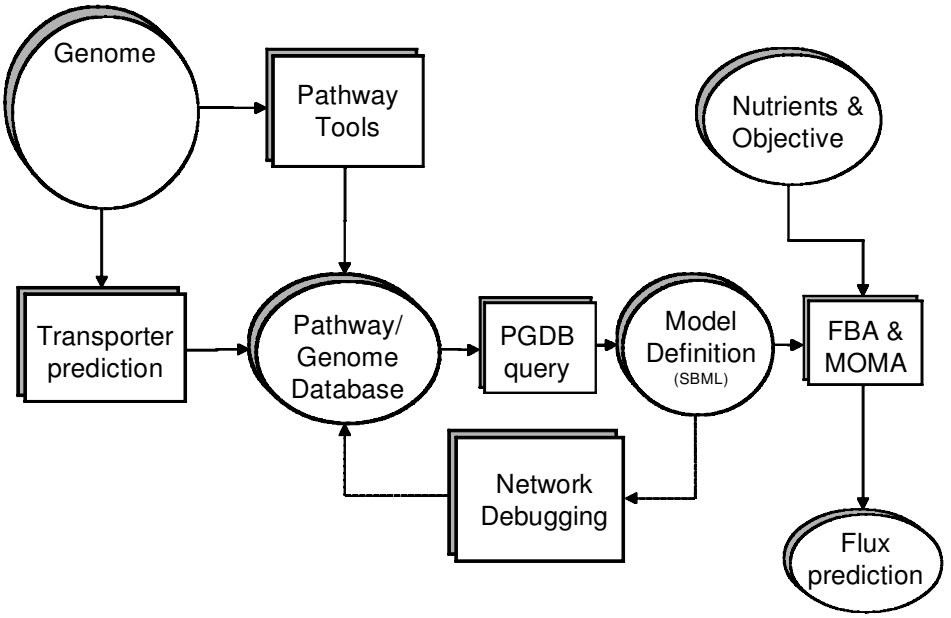


FIG. 5. A chart representing the pipeline for the automated generation of metabolic flux models, as described in the text.

from the genome annotations. However, a complete metabolic model construction often requires information that must be integrated from independent sources. For example, in EcoCyc, every known transport reaction is accounted for, but the Pathway/Genome databases that are computationally derived from annotated genomes using Pathologic do not yet contain this information (Karp et al., 2002; Karp et al., 2002). In order for a flux balance model to represent a cell faithfully, transport reactions must be included in the stoichiometric reconstruction. We are investigating the possibility of using the Membrane Transport Database (Paulsen, 2003) as a source of genome-derived transport reaction data. Additional information is required in order to perform *in silico* gene knockout experiments: a single reaction may be catalyzed by multiple enzymes; in turn, each enzyme may be a protein complex made up of multiple gene products. These associations can be described with Boolean relationships, which we call gene-reaction predicates (Fig. 7). For well-studied organisms (e.g., *E. coli*), gene-reaction predicates are partially known, and these annotations are available through the Pathway/Genome Database (Karp et al., 2002; Karp et al., 2002). For less studied organisms, the database information about these associations is limited by the genome annotation currently available. Cross-species homology searches and protein-protein interaction data could aid in solving this problem.

The next step in the pipeline is to use the Pathway Tools API to query Pathway/Genome Database and extract all the information necessary to define FBA and MOMA models. The data can be stored in SBML format (Hucka et al., 2003), which, in addition to containing the Enzyme Commission (EC) numbers, transport reactions, reaction directions, reaction stoichiometries, and metabolite names necessary for building *in silico* flux balance experiments, also represents the gene-enzyme-reaction relationships necessary for predicting the effects of a gene knockout. An example schema is displayed in the Appendix (see also Fig. 7).

Once the lists of reactions and metabolites have been compiled, the stoichiometric matrix is generated (Fig. 6). Such automatically generated stoichiometric matrices constitute an imperfect representation of the corresponding metabolic networks, which we found in general to be incomplete and contain some inconsistencies. In the past, these issues have prompted the use of manual curation. Our aim here, however is to explore further the possibility of automating this process as much as possible. Any problem that derives from an incomplete or ambiguous description of reactions and metabolites in public databases may consti-

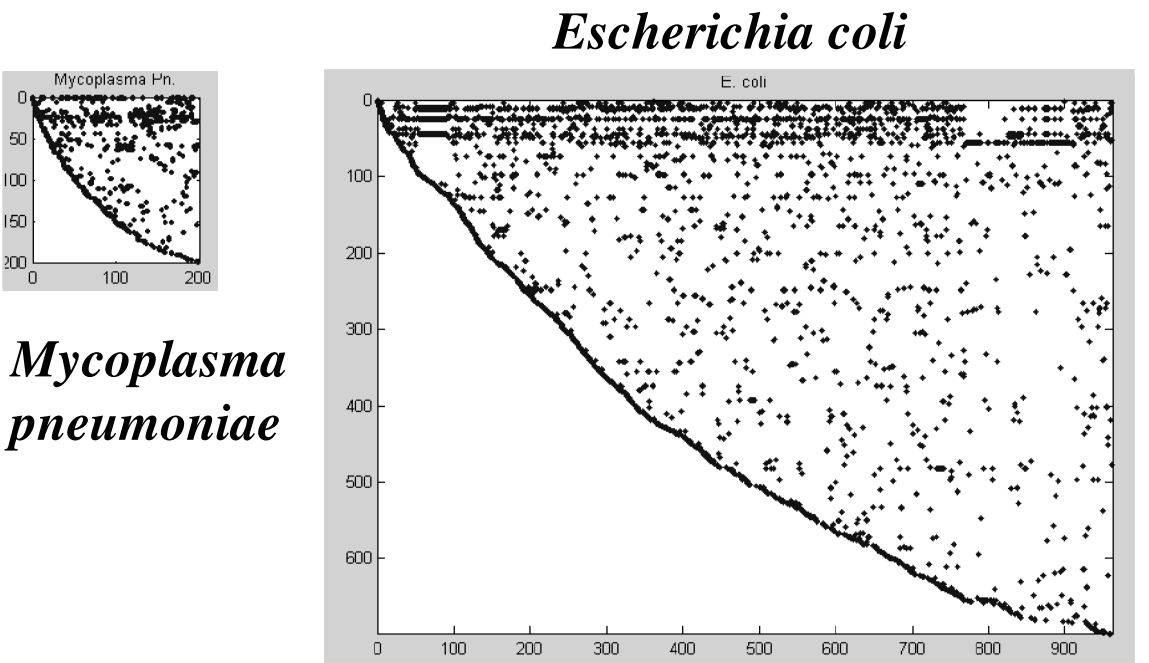


FIG. 6. Two examples of stoichiometric matrices obtained with the automatic pipeline. Rows correspond to metabolites, columns to reactions. The dots indicate nonzero entries in the matrices, showing the resulting sparsity pattern. The special triangular-like shape of these matrices is due to the sorting of metabolites according to the order by which they are added to the stoichiometry file, while parsing the metabolic reactions.

tute an obstacle for all future attempts of building large scale *in silico* models, and should therefore be addressed at its root. A diagnostic effort, aided by feedback from modelers, could help detect potential problems at the level of genome annotation (Brenner, 1999), and refine filters that would make sure all reactions are represented in a model-compatible way.

Incompleteness may be manifested as the existence of metabolites that cannot be produced by any reaction in the network. A network is considered incomplete if it cannot synthesize all essential biomass compounds given a known minimal nutrient set. To test for incompleteness in the network, we use the forward propagation algorithm described in (Romero and Karp, 2001) and a newly introduced backward propagation algorithm. Briefly, the Forward Propagation algorithm attempts to "fire" all reactions whose reactants belong to a set of producible metabolites, initialized with a nutrient set. At each step, the products of reactions that "fired" are added to the set of producible metabolites. This process continues until no more additional reactions can be fired. For biomass constituents that cannot be produced by the network, the backward propagation algorithm is run. This algorithm works by recursively descending through the precursors of the missing biomass constituents, and identifying unsynthesized intermediate metabolites that prevent the essential compound from being produced. These represent gaps in the network that must be resolved by introducing additional reactions or pathways. A caveat about interpreting the results of these algorithms is that the set of nutrients, biomass components and reactions may not be known with high confidence, so that it may be difficult to discriminate between missing reactions and missing nutrients.

To pass the test for consistency, every reaction must be elementally balanced. Anabolic polymerization reactions (Monomer + $\text{Polymer}_n \rightarrow \text{Polymer}_{n+1}$) require careful treatment. These reactions can be replaced with reactions of the kind Monomer → Polymer, in which Polymer represents the constituting unit of the class of polymers. In addition, many metabolites are superclasses (i.e., "an alcohol, or "a nucleotide"), and rules must be generated to identify and properly cope with these circumstances. Once the stoichiometric

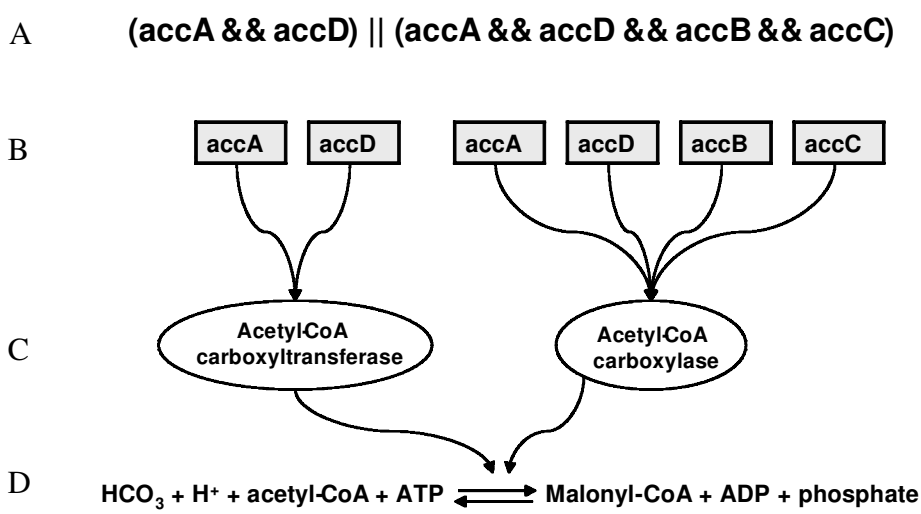


FIG. 7. An example of biochemical reaction for which we wrote a complete SBML representation (see Appendix), as part of an *E. coli* flux balance model. (**A**) The gene-reaction predicate, in Boolean logic, describes which genes have to be present in order for the reaction to be catalyzed, that is, for the flux to be potentially nonzero. (**B**) The genes themselves are the ultimate elementary units for performing simulations of knockouts (**C**) Enzymes are potentially composed of different protein subunits. (**D**) The reaction can be catalyzed by two different enzymes (as reflected in the OR of the gene-reaction predicate).

matrix passes the tests for completeness and consistency, the user may specify a nutrient set and an objective function which is used to generate a file in MPS format for the optimization software (Makhorin, 2001). FBA and MOMA can then be performed. We have applied our pipeline to generate stoichiometric matrices for several organisms, as exemplified in Fig. 5.

Inference of metabolic network kinetic parameters from flux balance methods and enzyme ratio measurements

Flux balance methods allow us to derive steady-state fluxes directly from the stoichiometry of the network and from condition-specific constraints. However, in order to achieve a complete understanding of whole cell metabolic networks, we will also need a complete characterization of their kinetic equations and dynamics, including all reaction parameters and concentration levels.

In FBA and MOMA, no kinetic equations are necessary to predict steady state flux distributions. The results could be used to reverse-engineer a system of equations compatible with the predicted steady states. However, in the absence of additional constraints infinitely many sets of equations could be chosen. In other words, given only the steady-state fluxes, the kinetic parameters of the metabolic equations are underdetermined. We show that, in principle, we can estimate all the kinetic parameters of the metabolic equations by applying FBA/MOMA in a number of different environmental conditions, provided we also have ratio measurements of the enzyme concentrations under those conditions.

Let us assume the reaction kinetics for each flux in the metabolic network can be described using Michaelis-Menten equations (Fell, 1996). Suppose that N is the number of reactions (and of fluxes) in the network, and M is the number of distinct metabolites. For illustration, consider a single substrate/single product irreversible reaction, $S_j \rightarrow P_j$, with rate v_i, and catalyzed by enzyme E_i. The classical form of the Michaelis-Menten equation can be written as follows:

$$v_i = \frac{V_i[S_j]}{K_i + [S_j]} \tag{1}$$

where $[S_j]$ represents the concentration of substrate S_j, K_i is the Michaelis-Menten constant, and V_i is the limiting (maximal) rate, which depends on the enzyme concentration $[E_i]$ and on the turnover number ki ($V_i = k_i\ [E_i]$). In general, more complex equations are required, which include additional kinetic parameters to account for possible multiple substrates/products, cross-inhibition and allosteric effects. For example, the reversible version of the single substrate/single product reaction (i.e., product inhibition) is characterized by the maximal forward rate and three kinetic constants (Fell, 1996).

For a single condition, there are N flux equations. The unknowns consist of M metabolite concentrations, N maximal rates and G kinetic constants ($G = N$ in the case where all equations are like Eq. 1). Hence, for a single condition, we can write:

$$\begin{aligned} \text{Equations} &= N \\ \text{Unknowns} &= M + N + G \end{aligned} \tag{2}$$

FBA (or MOMA for gene knockouts) allows us to determine the set of N fluxes under a number of different environmental conditions, which can be specified through the nutrient uptake rates. For every condition, there will be a steady state with a new set of M metabolite and N enzyme concentrations. However, all the kinetic constants remain unchanged. For example, for a single substrate/single product irreversible reaction i, we could write:

$$\begin{aligned} \text{Condition 1: } v_{i,1} &= \frac{V_{i,1}}{K_i + [S_j]_1}[S_j]_1 \\ \text{Condition 2: } v_{i,2} &= \frac{V_{i,2}}{K_i + [S_j]_2}[S_j]_2 \end{aligned} \tag{3}$$

where $V_{i,2}/V_{i,1} = E_{i,2}/E_{i,1}$. For each extra condition we get an additional N equations (for the new fluxes), but also $N+M$ unknowns (N maximal rates and M metabolite concentrations). Clearly, steady state fluxes alone are insufficient. As an additional constraint, we will assume that for each additional condition we can measure the ratio of enzyme concentrations with respect to the first condition (e.g., using quantitative mass spectrometry (Aebersold et al., 2000) or based on expression ratios). This gives us an additional N equations per condition, leading to the following general expression for the number of unknowns and equations under c different conditions:

$$\begin{aligned} \text{Equations}(c) &= N + 2N(c - 1) \\ \text{Unknowns}(c) &= c(N + M) + G \end{aligned} \tag{4}$$

An important parameter is the critical number of conditions c^* for which equations(c^*) = unknowns(c^*), that is, the number of conditions above which the system of equations becomes possibly overdetermined:

$$c^* = \frac{N + G}{N - M} \tag{5}$$

For a real metabolic network, we expect the number of kinetic constants to be between N and $8N$ (N for single substrate/single product irreversible reactions, up to $8N$ for bisubstrate/biproduct reactions [Savageau, 1976]). A fairly complete model of the *E. coli* metabolic network contains $M = 436$ metabolites and $N = 720$ fluxes (Edwards and Palsson, 2000a). Using an intermediate value of $G = 5N$ gives $c^* = 15$ conditions, indicating it may be possible to solve for the kinetic parameters using only a small number of conditions and/or knockouts. Fitting of the parameters to the predicted fluxes and observed enzyme ratios could be achieved using nonlinear optimization tools (Mendes and Kell, 1998).

In order for this method to work, the c conditions/knockouts should be independent from each other, and the FBA assumptions of optimality (suboptimality for MOMA) need to provide a good approximation. Measurements of absolute metabolite concentrations, or addition of known reaction rates could reduce the number of conditions that need to be explored significantly. Likewise, additional constraints based on stability of the network dynamics, optimal relationships between kinetic parameters (Heinrich and Schuster, 1996) and estimates of V_i based on $\max(v_i)$ could easily be incorporated into the parameter optimization. We en-

vision that the method described here may eventually be able to supply complete sets of putative in vivo binding and kinetic constants for whole metabolic networks. Systems likely to be adequate initial targets of this analysis are *E. coli* and the human red blood cell (Jamshidi et al., 2001).

DISCUSSION

Metabolic flux models represent a strong paradigm for the current integration efforts in systems biology. They can be inferred using annotated genomes available in public databases, and tools that associate genes and enzymes with explicit chemical reaction equations. Different biological conditions, as well as mutant strains and arbitrary metabolic objective functions can be chosen for *in silico* experiments. And finally, high-throughput experiments for detecting intracellular fluxes and growth phenotypes constitute a promising source of biological data against which hypotheses can be tested.

The coordination of these interdisciplinary efforts will benefit from the existence of a common language and of standards for model definitions and analyses. In the case of flux balance models, such standards constitute a developing issue (Hucka et al., 2003). Small adjustments to current representations of biochemical reactions could make it straightforward to generate automatically a whole-cell stoichiometric model. Network debugging algorithms could then immediately point out unresolved issues and potential missing pieces. One should not forget that part of the uncertainty that could be encountered derives from conflicting experimental results or from aspects of biochemical networks that are just unknown. But there is no reason for conflicting results and uncertainties not to be represented in machine-readable formats as well, so that updates could be incorporated immediately in databases and models. During our participation in the BioSPICE project, the SBML format has grown to include flux balance models. This will help both the growth of an automatic database-to-model pipeline, and the link between flux balance steady state models and full dynamic models involving chemical kinetics equations. Towards this integration, we have shown how flux balance models and high throughput experimental methods could potentially fuse with the world of whole-cell dynamic modeling. Ultimately, BioSPICE could serve as a platform on which cellular steady states obtained with flux balance models could be compared with steady states derived from kinetic equations. Flux balance–derived steady states would represent phenotypes associated with cellular objectives, while dynamic models could describe stability of different states and transitions between them.

ACKNOWLEDGMENTS

We are very grateful for the support of DARPA BioComp, AFRL contract number F30602-01-2-0586, and for the community-building efforts of the BioComp staff. Patrik D'haeseleer is a PhRMA/Harvard CEIGI fellow. Wayne Rindone is supported in part by an NIH glue grant (P.I. Dr. Ron Tompkins). Dat H. Nguyen is supported by the Alfred P. Sloan foundation and the U.S. Department of Energy. Special thanks to Peter Karp for useful comments on the manuscript.

REFERENCES

AEBERSOLD, R., RIST, B., and GYGI, S. P. (2000). Quantitative proteome analysis: methods and applications. Ann NY Acad Sci **919,** 33–47.

BAILEY, J.E. (2001). Complex biology with no parameters. Nat Biotechnol **19,** 503–504.

BONARIUS, H.P.J., SCHMID, G., and TRAMPER, J. (1997). Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. Trends Biotechnol **15,** 308–314.

BRENNER, S.E. (1999). Errors in genome annotation. Trends Genet **15,** 132–133.

COVERT, M.W., SCHILLING, C.H., FAMILI, I., et al. (2001). Metabolic modeling of microbial strains *in silico.* Trends Biochem Sci **26,** 179–186.

DARPA. (2003). BioSPICE [On-line]. Available: www.biospice.org.

EDWARDS, J.S., IBARRA, R.U., and PALSSON, B.O. (2001). *In silico* predictions of Escherichia coli metabolic capabilities are consistent with experimental data. Nat Biotechnol **19,** 125–130.

EDWARDS, J.S., and PALSSON, B. O. (1998). How will bioinformatics influence metabolic engineering? Biotechnol Bioeng **58,** 162–169.

EDWARDS, J.S., and PALSSON, B.O. (1999). Systems properties of the *Haemophilus influenzae* Rd metabolic genotype. J Biol Chem **274,** 17410–17416.

EDWARDS, J.S., and PALSSON, B.O. (2000a). The *Escherichia coli* MG1655 *in silico* metabolic genotype: its definition, characteristics, and capabilities. Proc Natl Acad Sci USA **97,** 5528–5533.

EDWARDS, J.S., and PALSSON, B.O. (2000b). Metabolic flux balance analysis and the *in silico* analysis of *Escherichia coli* K-12 gene deletions. BMC Bioinformatics **1,** 1.

FELL, D. (1996). *Understanding the Control of Metabolism* (London, Portland Press).

FISCHER, E., and SAUER, U. (2003). Metabolic flux profiling of *Escherichia coli* mutants in central carbon metabolism using GC-MS. Eur J Biochem **270,** 880–891.

FORSTER, J., FAMILI, I., PALSSON, B.O., et al. (2003). Large-scale evaluation of *in silico* gene deletions in *Saccharomyces cerevisiae.* OMICS **7,** 193–202.

GERTZ, E.M., and WRIGHT, S.J. (2001). Object-oriented software for quadrating programming in [On-line]. Available: *www.cs.wisc.edu/~swright/ooqp/.*

HEINRICH, R., and SCHUSTER, S. (1996). *The Regulation of Cellular Systems* (New York, Chapman & Hall).

HUCKA, M., FINNEY, A., SAURO, H.M., et al. (2003). The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics **19,** 524–531.

IBARRA, R.U., EDWARDS, J.S., and PALSSON, B.O. (2002). *Escherichia coli* K-12 undergoes adaptive evolution to achieve in silico predicted optimal growth. Nature **420,** 186–189.

JAMSHIDI, N., EDWARDS, J.S., FAHLAND, T., et al. (2001). Dynamic simulation of the human red blood cell metabolic network. Bioinformatics **17,** 286–287.

KANEHISA, M., GOTO, S., KAWASHIMA, S., et al. (2002). The KEGG databases at GenomeNet. Nucleic Acids Res **30,** 42–46.

KARP, P.D., PALEY, S., and ROMERO, P. (2002). The pathway tools software. Bioinformatics **18,** S1–S8.

KARP, P.D., RILEY, M., PALEY, S.M., et al. (2002). The MetaCyc Database. Nucleic Acids Research **30,** 59–61.

LUENBERGER, D.G. (1989). *Linear and Nonlinear Programming* (Reading, MA, Addison-Wesley).

MAKHORIN, A. (2001). GNU Linear Programming Kit (GLPK) [On-line]. Available: www.gnu.org/software/glpk/glpk.html.

MARTIN, D.L., CHEYER, A.J., and MORAN, D.B. (1999). The Open Agent Architecture: a framework for building distributed software systems. Applied Artificial Intelligence **13,** 91–128.

MENDES, P., and KELL, D. (1998). Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. Bioinformatics **14,** 869–883.

PAULSEN, I.T. (2003). TransportDB [On-line]. Available: www.membranetransport.org.

RAMAKRISHNA, R., EDWARDS, J.S., McCULLOCH, A. et al. (2001). Flux-balance analysis of mitochondrial energy metabolism: consequences of systemic stoichiometric constraints. Am J Physiol Regul Integr Comp Physiol **280,** R695–R704.

ROMERO, P., and KARP, P.D. (2001). Nutrient-related analysis of pathway/genome databases. Proceedings of the Pacific Symposium on Biocomputing, **6,** 471–482.

SAVAGEAU, M.A. (1976). *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology* (Reading, MA, Addison-Wesley).

SCHILLING, C.H., COVERT, M.W., FAMILI, I., et al. (2002). Genome-scale metabolic model of Helicobacter pylori 26695. J Bacteriol **184,** 4582–4593.

SCHILLING, C.H., EDWARDS, J.S. and PALSSON, B.O. (1999). Toward metabolic phenomics: analysis of genomic data using flux balances. Biotechnol Prog **15,** 288–295.

SCHILLING, C.H., and PALSSON, B.O. (1998). The underlying pathway structure of biochemical reaction networks. Proc Natl Acad Sci USA **95,** 4193–4198.

SCHILLING, C.H., and PALSSON, B.O. (2000). Assessment of the metabolic capabilities of *Haemophilus influenzae* Rd through a genome-scale pathway analysis. J Theor Biol **203,** 249–283.

SEGRÈ, D., VITKUP, D., and CHURCH, G.M. (2002). Analysis of optimality in natural and perturbed metabolic networks. Proc Natl Acad Sci USA **99,** 15112–15117.

TOMITA, M. (2001). Whole-cell simulation: a grand challenge of the 21st century. Trends Biotechnol **19,** 205–210.

VARMA, A., and PALSSON, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type *Escherichia coli* W3110. Appl Environ Microbiol **60,** 3724–3731.

WRIGHT, M.A., SEGRÈ, D., and CHURCH, G.M. (2002). 4-D model of bacterial chromosome structure. Presented at the International Conference on Systems Biology 2002.

YANAGIMACHI, K.S., STAFFORD, D.E., DEXTER, A.F., et al. (2001). Application of radiolabeled tracers to biocatalytic flux analysis. Eur J Biochem **268,** 4950–4960.

Address reprint requests to:
Dr. George M. Church
Lipper Center
Harvard Medical School
Boston, MA 02115

URL: http://arep.med.harvard.edu

APPENDIX

```
<?xml version="1.0"?>
<sbml xmlns=http://www.sbml.org/sbml/level2 version="1" level="2">
<model id="ECOLI" name="Generated from EcoCyc Pathway/Genome Database">
<listOfCompartments>
  <compartment id="cytoplasm"/>
  <compartment id="periplasm"/>
</listOfCompartments>
<listOfSpecies>
  <species id="HCO3" name="HCO3-" initialAmount="0"
compartment="cytoplasm" boundaryCondition="false"/>
  <species id="PROTON" name="H+" initialAmount="0"
compartment="cytoplasm" boundaryCondition="false"/>
  <species id="ACETYL_COA" name="acetyl-CoA" initialAmount="0"
compartment="cytoplasm" boundaryCondition="false"/>
  <species id="ATP" name="ATP" initialAmount="0" compartment="cytoplasm"
boundaryCondition="false"/>
  <species id="MALONYL_COA" name="malonyl-CoA" initialAmount="0"
compartment="cytoplasm" boundaryCondition="false"/>
  <species id="Pi" name="phosphate" initialAmount="0"
compartment="cytoplasm" boundaryCondition="false"/>
  <species id="="ACETYL_COA_CARBOXYLTRANSFER_CPLX "name="acetyl-CoA
carboxyltransferase" compartment="cytoplasm" boundaryCondition="false">
    <annotation xmlns:fbml="http://arep.med.harvard.edu/fbml">
      <fbml:gene id="EG11647" name="accA"/>
      <fbml:gene id="EG10217" name="accD"/>
    </annotation>
  </species>
  <species id="ACETYL_COA_CARBOXYLMULTI_CPLX" name="acetyl CoA
carboxylase" compartment="cytoplasm" boundaryCondition="false">
    <annotation xmlns:fbml="http://arep.med.harvard.edu/fbml">
      <fbml:gene id="EG10217" name="accD"/>
      <fbml:gene id="EG11647" name="accA"/>
      <fbml:gene id="EG10276" name="accC"/>
      <fbml:gene id="EG10275" name="accB"/>
    </annotation>
  </species>
```

```
</listOfSpecies>
<listOfReactions>
  <reaction id="ACETYL_COA_CARBOXYLTRANSFER_RXN" name="EC# 6.4.1.2"
reversible="true" >
    <annotation xmlns:flux="http://arep.med.harvard.edu/fluxns">
    <flux:limitation lower="-INF"/>
    <flux:limitation upper="INF"/>
    </annotation>
    <listOfReactants>
      <speciesReference species="HCO3" stoichiometry="1"/>
      <speciesReference species="PROTON" stoichiometry="1"/>
      <speciesReference species="ACETYL_COA" stoichiometry="1"/>
      <speciesReference species="ATP" stoichiometry="1"/>
    </listOfReactants>
    <listOfModifiers>
      <modifierSpeciesReference species="ACETYL_COA_CARBOXYLTRANSFER_CPLX"/>
      <modifierSpeciesReference species="ACETYL_COA_CARBOXYMULTI_CPLX"/>
    </listOfModifiers>
    <listOfProducts>
      <speciesReference species="MALONYL_COA" stoichiometry="1"/>
      <speciesReference species="Pi" stoichiometry="1"/>
      <speciesReference species="ADP" stoichiometry="1"/>
    </listOfProducts>
</listOfReactions>
</model>
</sbml>
```