Reviews

Large-Scale Metabolic Models: From Reconstruction to Differential Equations

Kieran Smallbone[1] and Pedro Mendes[1,2]

[1]Manchester Centre for Integrative Systems Biology and School of Computer Science, University of Manchester, Manchester Institute of Biotechnology, Manchester, UK
[2]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, VA

Abstract

Genome-scale kinetic models of metabolism are important for rational design of the metabolic engineering required for industrial biotechnology applications. They allow one to predict the alterations needed to optimize the flux or yield of the compounds of interest, while keeping the other functions of the host organism to a minimal, but essential, level. We define a pipeline for the generation of genome-scale kinetic models from reconstruction data. To build such a model, inputs of all concentrations, fluxes, rate laws, and kinetic parameters are required. However, we propose typical estimates for these numbers when experimental data are not available. While little data are required to produce the model, the pipeline ensures consistency with any known flux or concentration data, or any kinetic constants. We apply the method to create genome-scale models of Escherichia coli *and* Saccharomyces cerevisiae*. We go on to show how these may be used to expand a detailed model of yeast glycolysis to the genome level.*

Introduction

In recent years, two major (and divergent) modeling methodologies have been adopted to increase our understanding of metabolism and its regulation. Models contain either a large set of reactions with no kinetic detail (known as constraint-based models), or a few reactions described to high kinetic detail (kinetic models).

Constraint-based modeling uses physicochemical constraints such as mass balance, energy balance, and flux limitations to describe the potential behavior of an organism.[1,2] The consensus metabolic network of the model organism *S. cerevisiae* contains thousands of reactions and metabolites.[3–5] From the steady state solution space of all possible fluxes, a number of techniques have been proposed to deduce network behavior, including flux balance and extreme pathway or elementary mode analysis. In particular, flux balance analysis (FBA) highlights the most effective and efficient paths through the network in order to achieve a particular objective function.[6] The key benefit of FBA lies in the minimal amount of biological knowledge and data required to make quantitative inferences about network behavior. However, constraint-based modeling is concerned only with fluxes through the system and does not make any inferences or predictions about cellular metabolite concentrations.

By contrast, kinetic modeling aims to characterize fully the mechanics of each enzymatic reaction in terms of how changes in metabolite concentrations affect local reaction rates. Many metabolic models are available at BioModels.net.[7] However, they typically do not extend beyond central carbon metabolism and contain only tens of reactions, which is insufficient to deduce global metabolic behavior. Moreover, a considerable amount of data are required to parameterize a mechanistic model; if complex reactions like phosphofructokinase are involved, a single enzyme kinetic formula may have ten or more kinetic parameters.[8] The determination of such parameters is costly and time consuming and, moreover, many may be difficult or impossible to determine experimentally.

Attempts have been made to combine the two research paradigms to create large-scale kinetic models.[9–11] Starting with a network stoichiometry, they typically define generic rate laws such as linlog or Michaelis-Menten-like kinetics, before estimating those kinetic constants for which no experimental data are available.[12–16] However such methods do not take into account known steady state flux or concentration data, nor do they ensure thermodynamic constraints.[17]

Here we propose a pipeline for generation of thermodynamically consistent kinetic models using limited steady-state concentration and flux data, applied to *E. coli* and yeast, two major host organisms for industrial biotechnology. We go on to ask how these models may be improved with the availability of time-course data.

Materials and Methods

The pipeline is displayed schematically in *Fig. 1*, and described in detail below. The five models developed below are available at BioModels.net:[7]

- *E. coli*, linlog: http://identifiers.org/biomodels.db/MODEL1302140001
- *E. coli*, modular: http://identifiers.org/biomodels.db/MODEL1302140002
- yeast, linlog: http://identifiers.org/biomodels.db/MODEL1302140003
- yeast, modular: http://identifiers.org/biomodels.db/MODEL1302140004
- yeast, Pritchard: http://identifiers.org/biomodels.db/MODEL1302140005

DOI: 10.1089/ind.2013.0003

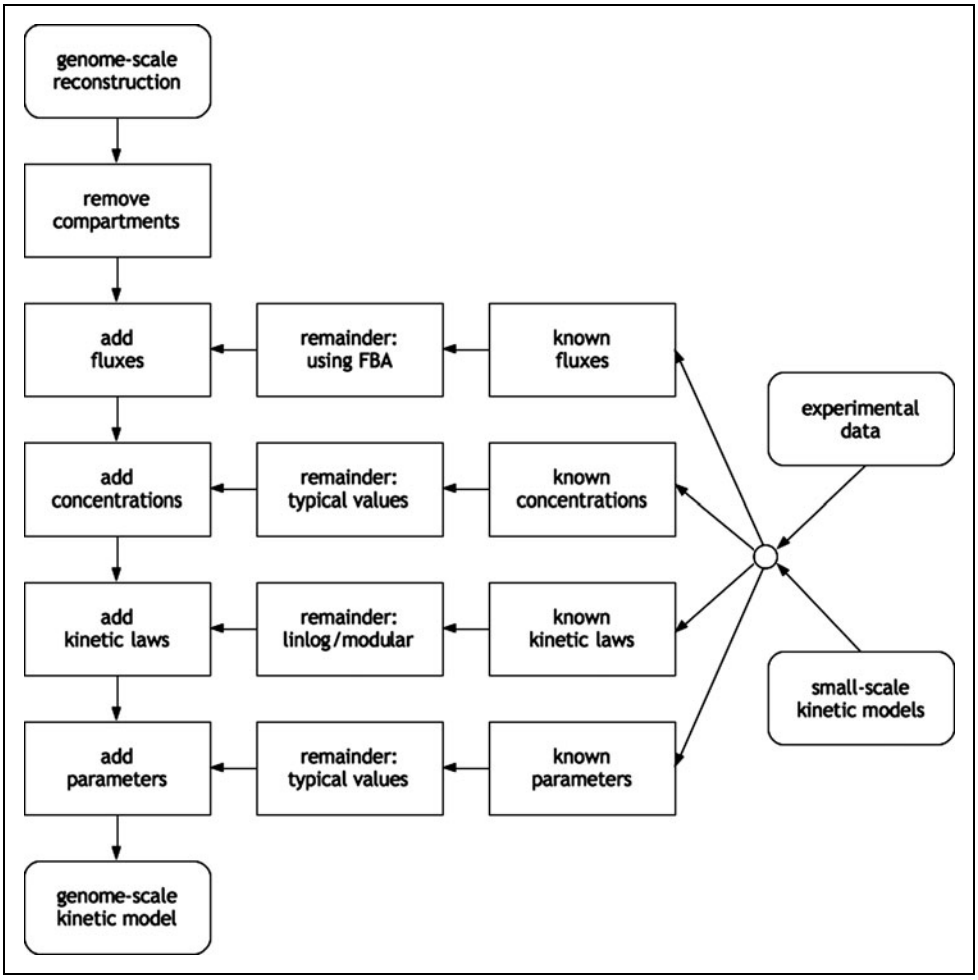

Fig. 1. Pipeline for the generation of a genome-scale kinetic model from a genome-scale reconstruction. The model may be populated with data from experiments or existing, small-scale models. Initial estimates may be provided for unknown entities by assigning them typical values.

NETWORK

Version 6 of the yeast consensus network was taken from http://yeast.sf.net/.[3–5] This is a comprehensive reconstruction of yeast metabolism made available in Systems Biology Markup Language (SBML).[18] We map all variables to either the intra- or extracellular space. This reduces the size of the network (*Table 1*); moreover, since all variables are now in one compartment, this transformation removes the need to know the (relative) compartment sizes.

Version 1 of the *E. coli* network was taken from http://ecoli.sf.net/. This is structured identically to the yeast model and derives from a recently published reconstruction.[19]

Table 1. Size of Models Used in This Study

ORGANISM	TYPE	VARIABLES	REACTIONS	COMPARTMENTS
E. coli	Reconstruction	1,806	2,583	6
E. coli	No comps	1,132	1,801	2
E. coli	Model	402	399	2
Yeast	Reconstruction	1,456	1,887	16
Yeast	No comps	764	1,159	2
Yeast	Model	303	282	2

FLUXES

A number of methods exist to measure the fluxes through the network. Experimentally, one may use isotope labeling, for example, to measure some fluxes. In the absence of such data, we can estimate unknown system fluxes using FBA, which allows the identification of an optimal path through the network in order to achieve a particular objective.[6] The techniques may be combined, with computational techniques used to choose a specific flux from the space of all solutions consistent with experimental data.

Here, we use geometric FBA to identify a unique reference flux from the space of all solutions to the FBA problem.[20] This algorithm, in particular the minimization of total flux, produces a flux distribution that is free from cycles and is thus thermodynamically feasible.[11]

Reactions with zero fluxes are removed from the network; the reaction directionality of negative fluxes is reversed, so that the predicted flux distribution is strictly positive. The resultant *E. coli* model has 402 reactions and 399 variables, while the yeast model has 303 reactions and 282 variables.

CONCENTRATIONS

To build a kinetic model, concentration values must be provided for all metabolites. Typically these will come from metabolomics measurements, or databases such as The Human Metabolome Database (HMDB).[21] In the absence of such data, typical values may be used. Here, extracellular nutrients were set to initial concentrations of 1 mM and intracellular metabolites were set to 0.1 mM (*Table 2*); these values are typical orders of magnitude.[10,11]

KINETIC RATE LAWS

Kinetic rate laws may be derived from knowledge of the mechanism underlying the enzymatic process, or taken from databases such as Sabio-RK.[22] In many cases

Table 2. Estimated Parameter Values for the Example Reaction A + B <-> 2 C.

PARAMETER	DESCRIPTION	VALUE
A_0, B_0, C_0	Initial concentration	0.1 mM (intracellular) 1 mM (extracellular)
V_0	Initial flux	Calculated in step 2
ε_A, ε_B	Elasticity	-1x stoichiometry
V_{max}	Maximum flux	Calculated from V_0
K_A, K_B, K_C	Michaelis constant	$K_A = A_0$; $K_B = B_0$; $K_C = C_0$
$K_{eq,0}$	Equilibrium constant	2

these are not known, particularly at the relevant physiological pH; we must instead resort to approximations such as the linlog or common modular rate laws that can be applied to any reaction stoichiometry.[11–13,16]

Consider the reaction $A + B \rightleftharpoons 2C$. Drawing ideas from metabolic control analysis, linlog defines the rate[15]

$$v = v_0\ (1 + \varepsilon_A \log(A/A_0) + \varepsilon_B \log(B/B_0) + \varepsilon_C \log(C/C_0)),$$

where v_0 is the initial flux; A_0, B_0, and C_0 are the initial concentrations; and ε_A, ε_B, ε_C are the elasticities.

The common modular rate instead defines

$$v = (V_{max}/K_A K_B)(A\ B - C^2/K_{eq})/((1 + A/K_A)(1 + B/K_B) + (1 + C/K_C)^2 - 1)$$

where V_{max} is the maximum flux; K_A, K_B, and K_C are the Michaelis constants; and K_{eq} is the equilibrium constant.

The linlog rate law suffers from a lack of saturation when its substrates tend to infinity. By contrast, the modular rate law is saturative and, moreover, includes thermodynamic properties via the equilibrium constant. Nonetheless, it may be preferable to use linlog kinetics, as systems of linlog equations contain fewer parameters and may be more numerically robust.[13]

PARAMETERIZATION

The kinetic formulae above lead to a range of kinetic parameters. Ideally, these would be measured under the conditions of interest using an enzymatic assay; when this is not possible, databases such as Brenda, Sabio-RK, or TECRdb (for equilibrium constants) can be used.[22–24] Again, in the absence of such data, first estimates must be provided. These are summarized in *Table 2*.

For the linlog rate law and example reaction $A + B \rightleftharpoons 2C$, the initial flux v_0 is calculated in step 2, and the initial concentrations A_0, B_0, and C_0 are calculated in step 3. Elasticities are estimated following the tendency modeling approach and taken to be the negative of their corresponding stoichiometric coefficient; thus $\varepsilon_A = \varepsilon_B = 1$ and $\varepsilon_C = -2$.[12]

Michaelis constants K_A are typically of the same order of magnitude as the metabolite concentration to which they refer.[11] Thus, for the common modular rate law, we take as initial estimates $K_A = A_0$, $K_B = B_0$, and $KC = C_0$. To ensure that the reaction is initially thermodynamically favorable in the forward direction, we take $K_{eq} = K_{eq,0}\ C_0^2/A_0\ B_0$ for some $K_{eq,0} > 1$. Following Ao[9], we take $K_{eq,0} = 2$, though alternatives are explored below. The only remaining unknown is V_{max} which may be simply calculated by equating the flux calculated in step 2 with the kinetic rate law, whose terms are all strictly positive.

With these choices of v_0 and V_{max} for the linlog and common modular rate laws, respectively, we ensure that the initial state is a steady state (though this state is not necessarily stable).

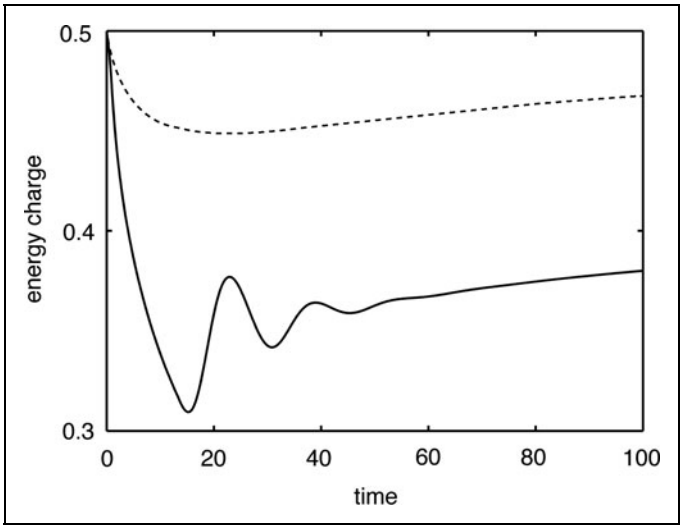

Fig. 2. Change in energy charge in the *E. coli* models, following a perturbation in extracellular glucose from 1 mM to 2 mM. Shown are the models with both modular **(solid line)** and linlog **(dashed line)** rate laws.

COMPUTATIONAL RESOURCES

At each stage of the pipeline, SBML files are manipulated using libSBML; this library may be accessed from a variety of programming languages.[25] Geometric FBA is performed using the Cobra toolbox, which is available for Matlab and python.[26] Steady state solutions, stiffness, and metabolic control analysis calculations are carried out using COPASI.[27]

Results

The methodology outlined above defines a pipeline for generation of genome-scale kinetic models from reconstruction data. As an example, we produce four models that do not

Table 3. Model Robustness Characteristics

ORGANISM	TYPE	λ_{max} (10^{-8}/s)	STIFFNESS (10^{11})
E. coli	Linlog	−3.91	1.21
E. coli	Modular	−2.75	1.89
Yeast	Linlog	−4.14	0.367
Yeast	Modular	540	7.59

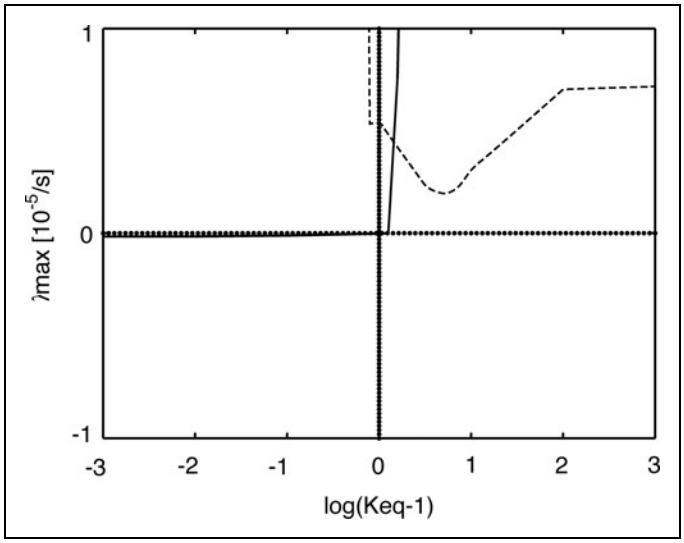

Fig. 3. Variation in the maximum eigenvalue with $\log(K_{eq,o}-1)$. Shown are *E.coli* **(solid line)** and yeast **(dashed line)**.

incorporate any data beyond what are available from the metabolic reconstructions, using two organisms (*E. coli* and yeast) and two generic rate laws (linlog and common modular). Once such a genome-scale metabolic model has been produced, *in silico* experiments may be performed and compared to experimental data. For example, *Fig. 2* illustrates the temporal evolution of *E. coli* energy charge [adenosine triphosphate (ATP); adenosine diphosphate (ADP); adenosine monophosphate (AMP)]:

$$([ATP]+[ADP]/2)/([ATP]+[ADP]+[AMP])$$

The energy charge, which reflects the extent to which there are anhydride-bound phosphate groups per adenosine moiety, varies between 0 (all AMP) and 1 (all ATP).[28] Atkinson argued that the energy charge is the main effector of enzymes that are sensitive to the energy status of the cell.[28] Following a change in extracellular glucose from 1 mM to 2 mM, we see that both *E. coli* models demonstrate a drop in energy charge in response to the perturbation. The modular (solid line) exhibits an oscillatory response, while the linlog (dashed line) model does not oscillate. Hypotheses such as this may be validated through and improved by comparison to extant experimental time-course data.

While linlog is a less accurate representation of saturative enzyme kinetics at the single reaction level, their relative simplicity (being linear in log-space) means they are better behaved at the genome scale. *Table 3* presents the maximum eigenvalue λ_{max} and stiffness of the four models as calculated using COPASI.[27] The linlog models are more stable and less stiff than their common modular counterparts; indeed the modular yeast model is unstable. Numerical robustness may be an important consideration when dealing with models of this size. It should be noted, however, that for all four models λ_{max} is very small. Its ''non-zeroness'' may be a numerical artifact, in which case linear theory cannot inform us about the stability of the system.

We may explore how characteristics such as model stability are affected by choice of parameter estimate. In *Fig. 3*, we present the effect of changing the estimated equilibrium constant on λ_{max}. For the *E. coli* model (solid line), the model becomes less stable as the equilibrium constants increase. This could be expected intuitively: an increase in K_{eq} leads to less product inhibition. However, the yeast model is most stable at $K_{eq} \approx 6$; it is unstable for all choices of equilibrium constant.

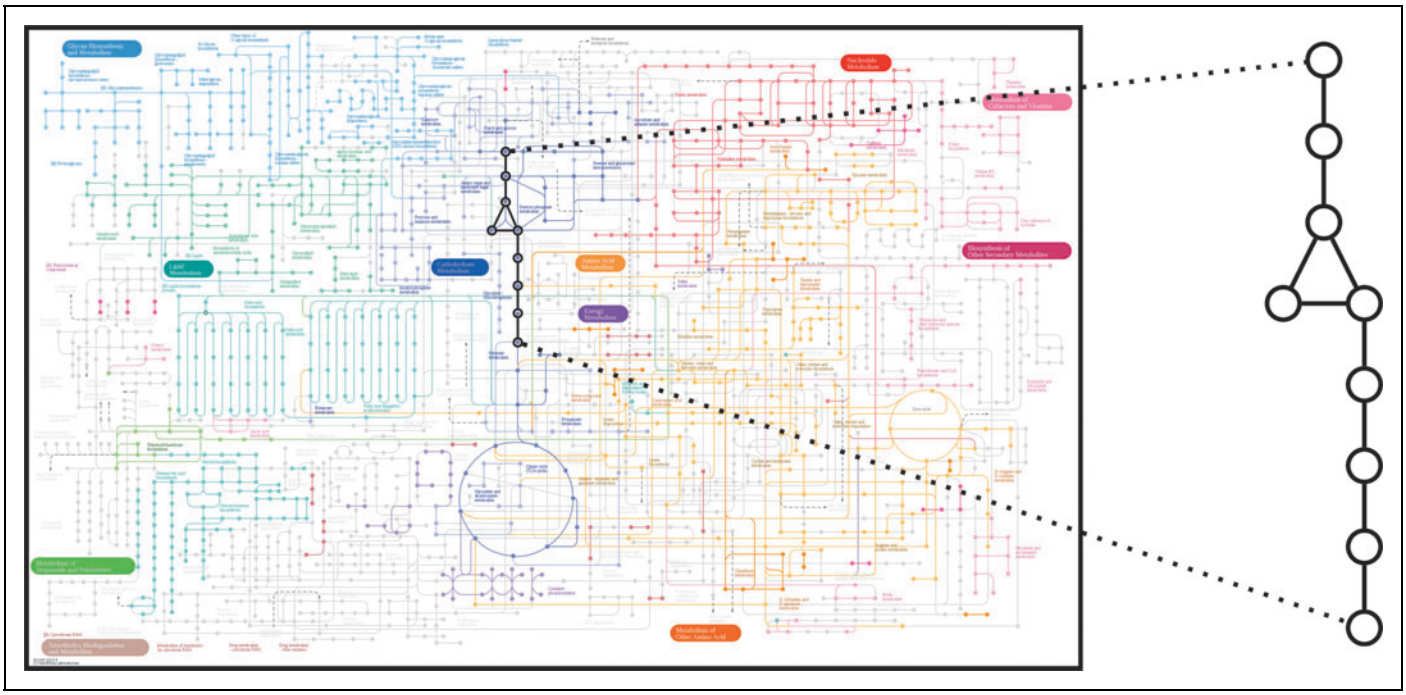

Fig. 4. Pathway level models with detailed kinetics, such as glycolysis, may be expanded to the genome level.

Table 4. Scaled Flux Control Coefficients for the Small-Scale Yeast Glycolysis Model and the Genome-Scale Model with Embedded Glycolysis

REACTION	GLYCOLYSIS	GENOME-SCALE
HXT	0.928	0.685
HXK	0.137	0.101
ATPase	−0.065	−0.032
PFK	0.013	0.057
FBA	0.009	0.038
TDH	0.006	0.009
ADH	0.006	−0.035
ENO	0.004	0.006
PGI	0.001	0.006
PYK	0.001	−0.056
PGK	0.000	0.001
GPM	0.000	−0.001
PDC	0.000	−0.135
TPI	0.000	0.000
AK	0.000	0.000

Up to this point, we have only considered the characteristics of models built without data. One natural way to add data is to embed existing small-scale kinetic models into larger genome-scale models. By running the smaller model to steady state, its fluxes, concentrations, and kinetics may be used as part of the above pipeline. Moreover, through use of semantic annotations, the embedding may be automated. We apply this idea by embedding a model of yeast glycolysis within a genome-scale model (*Fig. 4*).[29] By construction, the small-scale and large-scale models must share the same concentrations and fluxes at steady state.

However, the system-level behavior also remains similar. *Table 4* compares the flux control coefficients over glucose consumption for the small-scale model of glycolysis, and the large-scale model of glycolysis embedded within the whole network.[15] The control distributions are similar: they are highly correlated ($r = 0.973$) with those reactions with high control in the small model having similar control in the big model. The largest discrepancy is found in pyruvate decarboxylase; its substrate, pyruvate, forms a branch-point to the tricarboxylic acid cycle in the metabolic network. However, this important branch is not considered in the small glycolysis model; hence the reaction has much less control.

Discussion

The methodology outlined in this paper defines a pipeline for generation of genome-scale kinetic models from reconstruction data. To build such a model, inputs of all concentrations, fluxes, rate laws, and kinetic parameters are required. However, we propose typical estimates for these numbers when experimental data are not available. The pipeline ensures consistency with a sparse data set; while no data are required to produce the model, it can incorporate any known flux or concentration data or any kinetic constants. For example, the pipeline may be used to expand pathway level models with detailed kinetics—such as glycolysis—to investigate its interaction at the genome level. It can also allow one to make predictions of the effect of cloning a new pathway into a host, as is often desired in industrial biotechnology.

Genome-scale kinetic models offer possibilities not afforded by other methodologies. Genome-scale constraint-based modeling is concerned only with fluxes through the system and does not make any inferences or predictions about cellular metabolite concentrations. Nor does constraint-based modeling predict control properties of the network. Small-scale kinetic models can make predictions about concentrations and control, but only within their (small) remit. We have seen that genome-scale kinetic models can encompass the smaller-scale models while also allowing investigation of long-distance interactions, as shown in *Fig. 4*.

Two issues have arisen in this review. The first is the inherent stiffness of genome-scale models. Since cells need to produce some metabolites (such as ATP) at a much higher rate than others (such as zymosterol), metabolic processes will necessarily be taking place at different timescales. As such, systems biology tools are needed that can robustly simulate models of this size and with these numerical instabilities.

The second issue is that the pipeline above is built using data from a single steady state and, further, is built in such a way that the model exactly matches these data. However, time-course data such as that presented in *Fig. 2* contains more information and could be used to further constrain any unknown model parameters. Repositories of such dynamic data for use in model parameterization and validation would be of great benefit to the systems biology community.

Acknowledgments

The authors would like to acknowledge the financial support of the EU FP7 (KBBE) grant 289434 ''BioPreDyn: New Bioinformatics Methods and Tools for Data-Driven Predictive Dynamic Modelling in Biotechnological Applications.'' Pedro Mendes also thanks the BBSRC (BB/J019259/1) and NIH (R01 GM080219) for financial support. The authors also thank Natalie Stanford and Neil Swainston for valuable discussions.

Author Disclosure Statement

No competing financial interests exist.

REFERENCES

1. Covert MW, Famili I, Palsson BØ. Identifying constraints that govern cell behavior: A key to converting conceptual to computational models in biology? *Biotechnol Bioeng* 2003;84:763–772.

2. Price ND, Reed JL, Palsson BØ. Genome-scale models of microbial cells: Evaluating the consequences of constraints. *Nat Rev Microbiol* 2004;2:886–897.

3. Herrgård MJ, Swainston N, Dobson P, et al. A consensus yeast metabolic network obtained from a community approach to systems biology. *Nature Biotechnol* 2008;26:1155–1160.

4. Dobson PD, Jameson D, Simeonidis E, et al. Further developments towards a genome-scale metabolic model of yeast. *BMC Syst Biol* 2010;4:145.

5. Heavner BD, Smallbone K, Barker B, et al. Yeast 5—an expanded reconstruction of the *Saccharomyces cerevisiae* metabolic network. *BMC Syst Biol* 2012;6:55.

6. Kauffman KJ, Prakash P, Edwards JS. Advances in flux balance analysis. *Curr Opin Biotechnol* 2003;14:491–496.

7. Li C, Donizelli M, Rodriguez N, et al. BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol* 2010;4:92.

8. Teusink B, Passarge J, Reijenga CA, et al. Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *Eur J Biochem* 2000;267:53113–5329.

9. Ao P, Lee LW, Lidstrom M, et al. Towards kinetic modeling of global metabolic networks: Methylobacterium extorquens AM1 growth as validation. *Chinese J Biotechnol* 2008;24:980–994.

10. Smallbone K, Simeonidis E, Swainston N, Mendes P. Towards a genome-scale kinetic model of cellular metabolism. *BMC Syst Biol* 2010;4:6.

11. Stanford NJ. "Towards a full genome-scale model of yeast metabolism." Ph.D. thesis, University of Manchester, UK, 2011: www.escholar.manchester.ac.uk/uk-ac-man-scw:138720

12. Visser D, Heijnen JJ. Dynamic simulation and metabolic re-design of a branched pathway using linlog kinetics. *Metab Eng* 2003;5:164–176.

13. Smallbone K, Simeonidis E, Broomhead DS, Kell DB. Something from nothing: Bridging the gap between constraint-based and kinetic modelling. *FEBS J* 2007;274:5576–5585.

14. Liebermeister W, Klipp E. Bringing metabolic networks to life: Convenience rate law and thermodynamic constraints. *Theor Biol Med Model* 2006;3:1–11.

15. Kacser H, Burns JAE. The control of flux. *Symp Soc Exp Biol* 1973;27:65–104.

16. Liebermeister W, Uhlendorf J, Klipp E. Modular rate laws for enzymatic reactions: Thermodynamics, elasticities and Implementation. *Bioinformatics* 2010;26:1528–1534.

17. Lubitz T, Schulz M, Klipp E, Liebermeister W. Parameter balancing in kinetic models of cell metabolism. *J Phys Chem B* 2010;114:16298–16303.

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

19. Orth JD, Conrad TM, Na J, et al. A comprehensive genome-scale reconstruction of *Escherichia coli* metabolism—2011. *Mol Syst Biol* 2011;7:535.

20. Smallbone K, Simeonidis E. flux balance analysis: A geometric perspective. *J Theor Biol* 2009;258:311–315.

21. Wishart DS, Jewison T, Guo AC, et al. HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Res* 2013;41:D801–D807.

22. Wittig U, Kania R, Golebiewski M, et al. SABIO-RK—database for biochemical reaction kinetics. *Nucleic Acids Res* 2012;40:D790–D796.

23. Scheer M, Grote A, Chang A, et al. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res* 2011;39:D670–D676.

24. Goldberg RN, Tewari YB, Bhat TN. Thermodynamics of enzyme-catalyzed reactions—a database for quantitative biochemistry. *Bioinformatics* 2004;20:2874–2877.

25. Bornstein BJ, Keating SM Jouraku A, Hucka M. LibSBML: An API library for SBML. *Bioinformatics* 2008;24:880–881.

26. Schellenberger J, Que R, fleming RM, et al. Quantitative prediction of cellular metabolism with constraint-based models: the COBRA Toolbox v2.0 *Nat Protoc* 2011;6:1290–1307.

27. Hoops S, Sahle S, Gauges R, et al. COPASI: a COmplex PAthway SImulator. *Bioinformatics* 2006;22:3067–3074.

28. Atkinson DE, Walton GM. Adenosine triphosphate conservation in metabolic regulation. Rat liver citrate cleavage enzyme. *J Biol Chem* 1967;242:3239–3241.

29. Pritchard L, Kell DB. Schemes of flux control in a model of *Saccharomyces cerevisiae* glycolysis. *Eur J Biochem* 2001;269:3894–3904.

Address correspondence to:
Kieran Smallbone, PhD
Manchester Centre for Integrative Systems Biology
and School of Computer Science
University of Manchester
Manchester Institute of Biotechnology
131 Princess Street
Manchester M1 7DN
United Kingdom
Phone: 44 161 306 5146
Fax: 44 161 275 6204

E-mail: kieran.smallbone@manchester.ac.uk