Large-Scale Metabolic Models: From Reconstruction to Differential Equations

Kieran Smallbone¹ and Pedro Mendes¹,²

¹Manchester Centre for Integrative Systems Biology and School of Computer Science, University of Manchester, Manchester Institute of Biotechnology, Manchester, UK
²Virginia 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 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 © MARY ANN LIEBERT, INC. • VOL. 9 NO. 4 • AUGUST 2013 INDUSTRIAL BIOTECHNOLOGY 179
Version 6 of the yeast consensus network was taken from http://yeast.sf.net/. This is a comprehensive reconstruction of yeast metabolism made available in Systems Biology Markup Language (SBML). 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.

**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. 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. This algorithm, in particular the minimization of total flux, produces a flux distribution that is free from cycles and is thus thermodynamically feasible.

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). 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.

**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. In many cases

---

**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.

**Table 1. Size of Models Used in This Study**

<table>
<thead>
<tr>
<th>ORGANISM</th>
<th>TYPE</th>
<th>VARIABLES</th>
<th>REACTIONS</th>
<th>COMPARTMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>Reconstruction</td>
<td>1,806</td>
<td>2,583</td>
<td>6</td>
</tr>
<tr>
<td>E. coli</td>
<td>No comps</td>
<td>1,132</td>
<td>1,801</td>
<td>2</td>
</tr>
<tr>
<td>E. coli</td>
<td>Model</td>
<td>402</td>
<td>399</td>
<td>2</td>
</tr>
<tr>
<td>Yeast</td>
<td>Reconstruction</td>
<td>1,456</td>
<td>1,887</td>
<td>16</td>
</tr>
<tr>
<td>Yeast</td>
<td>No comps</td>
<td>764</td>
<td>1,159</td>
<td>2</td>
</tr>
<tr>
<td>Yeast</td>
<td>Model</td>
<td>303</td>
<td>282</td>
<td>2</td>
</tr>
</tbody>
</table>
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 rate15

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

where $v_0$ is the initial flux; $A_0$, $B_0$, and $C_0$ are the initial concentrations; and $\varepsilon_A$, $\varepsilon_B$, $\varepsilon_C$ are the elasticities.

The common modular rate instead defines

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

where $V_{\text{max}}$ 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 $\varepsilon_A = -1$ 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 $K_C = C_0$. 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 $V_{\text{max}}$, 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 $V_{\text{max}}$, 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).

### 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>
<thead>
<tr>
<th>ORGANISM</th>
<th>TYPE</th>
<th>$\lambda_{\text{max}}$ (10^{-9}/s)</th>
<th>STIFFNESS (10^{-11})</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. coli</td>
<td>Linlog</td>
<td>-3.91</td>
<td>1.21</td>
</tr>
<tr>
<td>E. coli</td>
<td>Modular</td>
<td>-2.75</td>
<td>1.89</td>
</tr>
<tr>
<td>Yeast</td>
<td>Linlog</td>
<td>-4.14</td>
<td>0.367</td>
</tr>
<tr>
<td>Yeast</td>
<td>Modular</td>
<td>540</td>
<td>7.59</td>
</tr>
</tbody>
</table>

---

**Table 2. Estimated Parameter Values for the Example Reaction $A + B \rightleftharpoons 2C$.**

<table>
<thead>
<tr>
<th>PARAMETER</th>
<th>DESCRIPTION</th>
<th>VALUE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_0$, $B_0$, $C_0$</td>
<td>Initial concentration</td>
<td>0.1 mM (intracellular) 1 mM (extracellular)</td>
</tr>
<tr>
<td>$v_0$</td>
<td>Initial flux</td>
<td>Calculated in step 2</td>
</tr>
<tr>
<td>$\varepsilon_A$, $\varepsilon_B$</td>
<td>Elasticity</td>
<td>$-1 \times$ stoichiometry</td>
</tr>
<tr>
<td>$V_{\text{max}}$</td>
<td>Maximum flux</td>
<td>Calculated from $v_0$</td>
</tr>
<tr>
<td>$K_A$, $K_B$, $K_C$</td>
<td>Michaelis constant</td>
<td>$K_A = A_0; K_B = B_0; K_C = C_0$</td>
</tr>
<tr>
<td>$K_{eq,0}$</td>
<td>Equilibrium constant</td>
<td>2</td>
</tr>
</tbody>
</table>

---

**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.
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)]:

\[
\frac{[\text{ATP}] + [\text{ADP}]/2}{[\text{ATP}] + [\text{ADP}] + [\text{AMP}]}
\]

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 \(\lambda_{\text{max}}\) and stiffness of the four models as calculated using CO-PASI.\(^{27}\) 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 \(\lambda_{\text{max}}\) 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 \(\lambda_{\text{max}}\). 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 \(K_{\text{eq}}\) leads to less product inhibition. However, the yeast model is most stable at \(K_{\text{eq}} \approx 6\); it is unstable for all choices of equilibrium constant.

![Fig. 3. Variation in the maximum eigenvalue with log(Keq-1). Shown are E.coli (solid line) and yeast (dashed line).](image)

![Fig. 4. Pathway level models with detailed kinetics, such as glycolysis, may be expanded to the genome level.](image)
Table 4. Scaled Flux Control Coefficients for the Small-Scale Yeast Glycolysis Model and the Genome-Scale Model with Embedded Glycolysis

<table>
<thead>
<tr>
<th>Reaction</th>
<th>Glycolysis</th>
<th>Genome-Scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>HXT</td>
<td>0.928</td>
<td>0.685</td>
</tr>
<tr>
<td>HXK</td>
<td>0.137</td>
<td>0.101</td>
</tr>
<tr>
<td>ATPase</td>
<td>-0.065</td>
<td>-0.032</td>
</tr>
<tr>
<td>PFK</td>
<td>0.013</td>
<td>0.057</td>
</tr>
<tr>
<td>FBA</td>
<td>0.009</td>
<td>0.038</td>
</tr>
<tr>
<td>TDH</td>
<td>0.006</td>
<td>0.009</td>
</tr>
<tr>
<td>ADH</td>
<td>0.006</td>
<td>-0.035</td>
</tr>
<tr>
<td>ENO</td>
<td>0.004</td>
<td>0.006</td>
</tr>
<tr>
<td>PGK</td>
<td>0.000</td>
<td>0.001</td>
</tr>
<tr>
<td>GPM</td>
<td>0.000</td>
<td>-0.001</td>
</tr>
<tr>
<td>PDC</td>
<td>0.000</td>
<td>-0.135</td>
</tr>
<tr>
<td>TPI</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>AK</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

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). 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. 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


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

This work is licensed under a Creative Commons Attribution 3.0 United States License. You are free to copy, distribute, transmit and adapt this work, but you must attribute this work as “Industrial Biotechnology. Copyright 2013 Mary Ann Liebert, Inc. http://liebertpub.com/ind, used under a Creative Commons Attribution License: http://creativecommons.org/licenses/by/3.0/us/”