Use of Randomized Sampling for Analysis of Metabolic Networks*

Published, JBC Papers in Press, October 20, 2008, DOI 10.1074/jbc.R800048200

Jan Schellenberger^ and Bernhard Ø. Palsson^1

From the ^1Bioinformatics Program and the ^2Bioengineering Department, University of California, San Diego, La Jolla, California 92093-0412

Genome-scale metabolic network reconstructions in microorganisms have been formulated and studied for about 8 years. The constraint-based approach has shown great promise in analyzing the systemic properties of these network reconstructions. Notably, constraint-based models have been used successfully to predict the phenotypic effects of knock-outs and for metabolic engineering. The inherent uncertainty in both parameters and variables of large-scale models is significant and is well suited to study by Monte Carlo sampling of the solution space. These techniques have been applied extensively to the reaction rate (flux) space of networks, with more recent work focusing on dynamic/kinetic properties. Monte Carlo sampling as an analysis tool has many advantages, including the ability to work with missing data, the ability to apply post-processing techniques, and the ability to quantify uncertainty and to optimize experiments to reduce uncertainty. We present an overview of this emerging area of research in systems biology.

The advent of whole genome sequencing has led to the curation of many genome-scale metabolic reconstructions (1, 2). These reconstructions are mathematically structured knowledge bases containing biochemical, genetic, and genomic information about a metabolic network. Whereas the content of these network reconstructions can be fairly complete, the functional (i.e. physiological) states that these networks can achieve are more difficult to determine, and such determination is an active area of research. The analysis of whole cell metabolism comes with unknown quantities. Even a description of a steady-state condition may require knowledge of a large number of metabolite concentrations and reaction rates (fluxes). Dynamic analysis adds additional complexity.

Although significant effort has gone toward measuring each of these quantities through various high-throughput omics methods, obtaining all the needed numerical values is a significant challenge that may not be met in full for a long time. This incompleteness of data has created the need for analysis methods that are able to give meaningful results with only partial measurements. One approach that has been successfully used is so-called constraint-based reconstruction and analysis (COBRA). This approach emphasizes describing the constraints that a system must satisfy rather than computing an explicit solution. Some of the variables and constraints that have been applied in the past are listed in Table 1. This approach leads to the definition of a “solution space,” which contains the set of feasible solutions that satisfy all imposed constraints. Fig. 1 shows a schematic of this approach and how it compares with the traditional simulation approach. If the system equations are set up correctly, the “true state” of the network will lie within the imposed constraints and may then be further analyzed.

Many approaches exist for studying this solution space (3, 4). One of the more recent approaches is randomized sampling of candidate solutions. To study the space of solutions, a random set of points is chosen from it to act as a surrogate for the entire space. Many of the properties that can be calculated for one candidate solution can then be calculated for point throughout the entire space, and the properties of this set of solutions can be evaluated in a statistical fashion. This procedure gives information about the how limiting the imposed constraints are, and the results can be used to design further experiments to shrink the size of the solution space.

The workflow associated with the constraint-based paradigm is outlined in Fig. 2. The scope of the biological system is defined as variables (fluxes, concentrations, pressure, etc.) and parameters (kinetic constants, thermodynamic values, etc.). Experimental measurements, biophysical constraints, and other known constraints are imposed on the system, yielding a solution space. It can then be probed with a variety of methods, including optimization and Monte Carlo sampling. The results from such studies are used to design further experiments to further constrain the system.

The ultimate goal of large-scale network analysis is to provide a framework for understanding whole cell metabolism. This includes interpreting data from various high-throughput omics experiments, creating predictive models of how the cell works, and ultimately being able to manipulate cells for medical or industrial purposes.

Steady-state Flux Balance Analysis

A metabolic network can be concisely described in matrix format using the stoichiometric matrix, S (sometimes called N). Every row in this matrix represents one metabolite, and each column represents a reaction. A non-zero entry in S indicates participation of a metabolite (i) in a reaction (j). The rates of every reaction (flux) can be written as a vector (v) forming the fundamental mass balance equation (Equation 1).

$$\frac{dx}{dt} = S \cdot v(x)$$

(Eq. 1)

Integrating this equation over time yields the time course of concentrations (x(t)). In general, the rates of reaction (v) are a function of the concentrations (x) as well as enzyme kinetic constants and other parameters. Because it is hard to measure all the kinetic constants needed to simulate dynamic responses, the steady states of the system are often studied. Steady state implies $\frac{dx}{dt} = 0$ and therefore $Sv = 0$. A different way of stating this is that the (right) null space of S contains the steady-state flux distributions. This solution space is finite in size given...
enzyme capacity constraints \( v_{\text{min}} < v < v_{\text{max}} \); it has been studied extensively, and its properties can be determined by a variety of approaches (3). It has recently been realized that randomized sampling of solutions in the solution space is an effective way to characterize its contents.

Example Uses of Sampling in Small Spaces

A number of questions about solution spaces can be addressed using randomized sampling. Three notable examples of studies of small solution spaces are now described.

Designing Experiments—The existence of a solution space means that many conceivable flux distributions satisfy the steady-state condition. Even with all uptake and secretion rates (inputs and outputs) measured, the internal flux rates are usually not uniquely specified. If there are two parallel pathways for the same process, it will be impossible to distinguish which is being used. This non-uniqueness of flux states was realized early (5), and thus, additional measurements are needed to further eliminate candidate solutions. Sampling of the candidate solutions can be used to find the most informative measurements to make (6, 7). Many possible experiments were simulated, and Monte Carlo sampling was used to simulate random experimental noise and to propagate it through the network. The ratio of measurement noise to computed noise was statistically quantified, thus rejecting experiments with poor design in favor of more informative ones.

Determining the Shape of Solution Spaces—Smaller networks (<40 reactions) have flux solution spaces that can be studied directly using techniques from convex analysis. Wiback et al. (8) defined the flux space of the core red blood cell model and computed its volume using a mathematical technique called vertex enumeration. Monte Carlo elimination sampling was used to describe the “shape” of the space by plotting the distribution of points as a function of flux through each reaction. The shape flux space contains important information about the likelihood of finding the true solution at any particular flux value. A “narrow” space indicates low likelihood, whereas a “wide” space indicates greater likelihood. An updated method for computing the volume of the space can be found in Ref. 9.

Consequences of Genetic Variation—Price et al. (10) developed a refined sampling approach that scaled to slightly larger networks. Enzymopathies of the red blood cell were studied by decreasing the \( v_{\text{max}} \) of the reaction catalyzed by enzymes with known biochemical deficiencies and clinically observable altered phenotypes. It was shown that the enzymopathies that decreased the volume of the flux space most significantly were more likely to have a clinical effect \( \textit{in vivo} \). This work first considered looking at the correlations between points as another way of describing the shape of the space and noted that these values tend to shift while simulating enzymopathies.

Determining Global Network Properties

With the availability of genome-scale network reconstructions, there has been significant interest in characterizing them in an unbiased fashion. Two unbiased approaches have emerged (3). One approach has been the development of network-based pathways as convex basis vectors such as extreme
pathways (11) and elementary modes (12). Monte Carlo sampling provides an alternate unbiased way. Both have significant numerical challenges at the genome scale, with the determination of convex pathways being potentially impractical (13), whereas randomized sampling can be achieved at this scale.

Sampling Methods—With the advent of Monte Carlo methods for the study of the flux space of much larger networks, the dominant algorithm of choice has been the Markov chain Monte Carlo (MCMC), also known as the “hit-and-run” sampler. Unlike the elimination algorithm usable for smaller networks, the MCMC algorithm produces a valid solution point at every iteration. An initial valid point is moved repeatedly inside the space according to probabilistic rules. The trail of valid generated points becomes the sample. One key disadvantage of this algorithm is that there is no guarantee that these points cover the entire space in a finite time. This behavior is known as “slow mixing.” One improvement to the MCMC algorithm is artificial centering hit-and-run (ACHR) (14). Essentially all publications of sampling of the flux space have used this algorithm.

Elucidation of a High Flux Backbone—Barabási and co-workers published two studies that used Monte Carlo sampling of the flux space to look at global network properties. In Ref. 15, they demonstrated that *Escherichia coli* flux space is dominated by a few high capacity reactions (the “high flux backbone”) that are robust between different simulated media conditions. In Ref. 16, they compared the metabolic reconstructions of *Helicobacter pylori*, *E. coli*, and *Saccharomyces cerevisiae* under randomly generated media conditions and used flux balance analysis to compute the flux shifts between them. An “activity core” was defined based on reactions that are always required for growth. This activity core varies in size between the three reconstructions and is a reflection of the redundancy and robustness built into these systems.

Definition of “Modules”—Modules can be determined based on correlations between reactions. When two reactions are correlated either perfectly ($r^2 = 1$) or nearly perfectly ($r^2 \sim 1$), the flux through them must be linearly related at steady state. An example of this is two reactions in a linear pathway. This concept was reviewed by Papin et al. (17) as one technique of network classification and applied to the *Mycobacterium tuberculosis* network (18). Although correlated sets based on extreme pathways and uniform random samples are distinct, a comparison showed the global network properties to be conserved (19).

Applications of Sampling to Study Disease States

Whereas global network properties are of academic interest, sampling has also been used to study clinical issues.

*Human Mitochondria*—The human cardiac mitochondrial network (20) was analyzed using the sampling approach to characterize network capabilities under different conditions, including various diets and simulating diabetic conditions. A particularly striking result was the observation that the pyruvate dehydrogenase flux in diabetic patients is constrained to be lower than in non-diabetic patients due to mass conservation constraints alone. This result was surprising because although it had been known that pyruvate dehydrogenase has a decreased flux in diabetics, it was believed to be a consequence of unknown regulatory mechanisms. The sampling approach thus demonstrated how bottom-up reconstructions can describe real biochemical and physiological conditions and provide mechanistic insights into the cause and effect relationships.

Co-sets and Their Applications—Another result of this study came from the use of comprehensively sampled points in the flux space to calculate statistically perfectly correlated sets of reactions, termed co-sets (17). A non-zero flux in one member of a co-set implies a non-zero flux in all other members of the set and vice versa. Hence, co-sets define reactions that are used together as a result of mass conservation determined by network topology. These reaction co-sets can be used to simplify the network and to enable analysis of disease states and alternative drug targets in terms of causal pathways rather than individual reactions. They can also be used to correlate the causality of single nucleotide polymorphisms that appear in the enzymes participating in a co-set (21).

*Human Neural Reconstruction*—Occhipinti et al. (22) studied a five-compartment model of human neuron/astrocyte metabolism with Monte Carlo sampling and statistical inference. The method was based on a previous study (23) and assumes a prior distribution on each flux in the network. Experimental measurements are treated as observations that are used to update the distributions of the fluxes. In this way, it is possible to test specific experimental hypotheses *in silico*. The study investigated which metabolic pathways are more active in brain metabolism at steady states with different neural activities and was able to provide quantitative answers.

Extensions to Analysis of Dynamic States

Moving beyond a steady-state framework to dynamic analysis presents a new set of challenges. Kinetic parameters needed for dynamic analysis are numerous and often unknown. *In vitro* techniques often fail to give numerical values for kinetic param-
MINIREVIEW: Random Sampling of Metabolic Networks

eters that are consistent with those in vivo (24). Obtaining such a consistent set is expected to require whole system approaches.

There is a strong need for developing methods that allow for global kinetic analysis because many of the important systemic properties lie in the dynamic domain. With the recent development of metabolomic measurements, we may be able bridge the gap between steady-state and fully dynamic descriptions. Because of the inherent uncertainty in measurements, Monte Carlo sampling can be used to further this aim.

**k-Cone Analysis**—Famili et al. (25) proposed defining a space of feasibility within which kinetic parameters must be found. This was termed the k-cone. It is analogous to the flux space in that every point represents a feasible set of kinetic parameters. This construction assumes that the concentrations of metabolites are known and that the system is at steady state. Famili et al. were able to show that 1) feasible kinetic parameters exist; 2) it is possible to find the set of kinetic parameters that most closely match a measured set (by optimization); and 3) by repeating this procedure under different conditions, the set of kinetic parameters can be narrowed down.

**Structural Kinetic Modeling**—Steuer et al. (26–28) proposed a different way of looking at whole system kinetics. Structural kinetic modeling makes no assumptions about kinetic parameters and instead focuses on the feedback parameters in a network. Equation 1 is linearized about a hypothesized steady-state value \( \mathbf{x}_{ss} \) with a steady-state flux of \( \mathbf{v}_{ss} \). The linearization results in a Jacobian matrix \( \mathbf{J} \), which contains dynamic properties of the system. For example, the eigenvalues of \( \mathbf{J} \) indicate whether the system is stable or unstable near the steady state. Many of the feedback parameters to build \( \mathbf{J} \) are unknown. To study their effects, Monte Carlo sampling was used, and statistical properties were determined (28). A ranking of a feedback site's stability was obtained.

**Bayesian Statistical Approaches**—Liebermeister and Klipp (29) used Bayesian statistics to describe the distribution of kinetic parameters. Starting with a stoichiometric model, various experimental data types (metabolic concentrations, thermodynamic, kinetic) are integrated in a statistical fashion to produce a set of kinetic models. As a test example, a small threonine metabolism model was used to simulate noisy experimental measurements. The Bayesian approach was able to narrow the possible range of kinetic parameters from a very large range initially to a much smaller feasible range.

**Challenges for the Future**

The challenges in this area are several. The main categories are as follows.

**Computational Challenges**—Sampling a solution space is difficult as the number of variables increases. This is known as the “curse of dimensionality.” Various tricks have been employed to get around this limitation (for example, ACHR), and this is an area of active research. The flux space and k-cone formalism have the advantage that the commonly used constraints form a convex space, which has some “nice” properties. Writing a sampling procedure capable of sampling a generic set of constraints may be quite difficult. Instead, a careful tradeoff between feasibility of sampling and expressiveness of the formalism and constraints is required. The COBRA toolbox (31) contains tools for Monte Carlo sampling as well as other COBRA techniques.

**Extension of Formalism to Other Parameters**—As shown, Monte Carlo sampling has been used to sample the flux space as well as the kinetic space. However, it is conceivable to extend the formalism to many other areas. There have been recent attempts to describe regulatory networks in a stoichiometric fashion (32). The formalism uses an R (for “regulatory”) matrix, which is analogous to the S matrix. An R matrix has been constructed based on the model used by Covert et al. (33) and interrogated using uniform random sampling. The simulation results exhibited strong concordance for two environments for which microarray experiments have been conducted. In addition, an analysis of the row and null spaces demonstrated that these spaces describe all possible gene expression states for E. coli in a given environment.

**Thermodynamic Constraints**—Kümmel et al. (34) used thermodynamic considerations to describe a set of constraints on the concentrations of metabolites. With the increasing availability of metabolomic data as well as computational estimates of thermodynamic properties, constraints of this type are sure to become more prevalent. Table 1 illustrates the variables, parameters, and constraints that have been defined for metabolic networks. In all cases, it is not possible to obtain all measurements for a full description with current high-throughput techniques. However, there are constraints between variables that suggest that not every variable must be explicitly measured. For some of these applications, Monte Carlo sampling has been applied (fluxes and kinetics). For others, this has not been done yet.

One simplification to full thermodynamic considerations is the inclusion of the so-called “loop law.” This law places additional constraints on internal metabolic fluxes to avoid circular flux distributions that are thermodynamically infeasible. Price et al. (30) studied this feasible space by sampling the space of a H. pylori reconstruction and eliminating points that violate the loop law. Although this method was shown not to scale to larger networks, it has the advantage of not requiring knowledge of thermodynamic constants.

**Conclusions**

Monte Carlo sampling is emerging as an approach to deal with the analysis of genome-scale metabolic networks. Ideally, we would like to know the true internal state of a cell at all times. As this goal is currently intractable, much effort has been focused on giving an accurate description of the solution space within which the true state must lie. Unbiased analysis of this space through randomized sampling has yielded many novel results and provides a framework by which further experiments can be designed.

**REFERENCES**

1. Reed, J. L., Famili, I., Thiele, I., and Palsson, B. Ø. (2006) Nat. Rev. Genet. 7, 130–141
2. Kim, H. U., Kim, T. Y., and Lee, S. Y. (2008) Mol. BioSyst. 4, 113–120
3. Price, N. D., Reed, J. L., and Palsson, B. Ø. (2004) Nat. Rev. Microbiol. 2, 152–160
4. E. P. Gianchandani, A. R. Joyce, B. Ø. Palsson, and J. A. Papin, submitted for publication.
886–897

4. Palsson, B. Ø. (2006) Systems Biology: Properties of Reconstructed Networks, Cambridge University Press, New York

5. Varma, A., and Palsson, B. Ø. (1993) J. Theor. Biol. 165, 477–502

6. Savinell, J. M., and Palsson, B. Ø. (1992) J. Theor. Biol. 155, 201–214

7. Savinell, J. M., and Palsson, B. Ø. (1992) J. Theor. Biol. 155, 215–242

8. Wiback, S. J., Famili, I., Greenberg, H. J., and Palsson, B. Ø. (2004) J. Theor. Biol. 228, 437–447

9. Braunstein, A., Mulet, R., and Pagnani, A. (2008) BMC Bioinformatics 9, 240

10. Price, N. D., Schellenberger, J., and Palsson, B. Ø. (2004) Biophys. J. 87, 2172–2186

11. Schilling, C. H., Letscher, D., and Palsson, B. Ø. (2000) J. Theor. Biol. 203, 229–248

12. Schuster, S., and Hilgetag, C. (1994) J. Biol. Syst. 2, 165–182

13. Yeung, M., Thiele, I., and Palsson, B. Ø. (2007) BMC Bioinformatics 8, 363

14. Kaufman, D. E., and Smith, R. L. (1998) Operations Res. 46, 84–96

15. Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N., and Barabási, A.-L. (2004) Nature 427, 839–843

16. Almaas, E., Oltvai, Z. N., and Barabási, A.-L. (2005) PLoS Comput. Biol. 1, e68

17. Papin, J. A., Reed, J. L., and Palsson, B. Ø. (2004) Trends Biochem. Sci. 29, 641–647

18. Jamshidi, N., and Palsson, B. Ø. (2007) BMC Syst. Biol. 1, 26

19. Barrett, C. L., Price, N. D., and Palsson, B. Ø. (2006) BMC Bioinformatics 7, 132

20. Thiele, I., Price, N. D., Vo, T. D., and Palsson, B. Ø. (2005) J. Biol. Chem. 280, 11683–11695

21. Jamshidi, N., and Palsson, B. Ø. (2006) Mol. Syst. Biol. 2, 38

22. Occhipinti, R., Puchowicz, M. A., LaManna, J. C., Somersalo, E., and Calvetti, D. (2007) Ann. Biomed. Eng. 35, 886–902

23. Daniela Calvetti, E. S. (2006) SIAM Multiscale Model. Simul. 5, 1333–1366

24. Teusink, B., Passarge, I., Reijenga, C. A., Eschalho, E., van der Weijden, C. C., Scheper, M., Walsh, M. C., Bakker, B. M., van Dam, K., Westerhoff, H. V., and Snoep, J. L. (2000) Eur. J. Biochem. 267, 5313–5329

25. Famili, I., Mahadevan, R., and Palsson, B. Ø. (2005) Biophys. J. 88, 1616–1625

26. Steuer, R., Gross, T., Selbig, J., and Blasius, B. (2006) Proc. Natl. Acad. Sci. U. S. A. 103, 11868–11873

27. Steuer, R. (2007) Phytochemistry 68, 2139–2151

28. Grimbs, S., Selbig, J., Bulik, S., Holzhütter, H.-G., and Steuer, R. (2007) Mol. Syst. Biol. 3, 146

29. Liebermeister, W., and Klipp, E. (2006) Theor. Biol. Med. Model. 3, 42

30. Price, N. D., Thiele, I., and Palsson, B. Ø. (2006) Biophys. J. 90, 3919–3928

31. Becker, S. A., Feist, A. M., Mo, M. L., Hannum, G., Palsson, B. Ø., and Herrgard, M. J. (2007) Nat. Protocols 2, 727–738

32. Gianchandani, E. P., Papin, J. A., Price, N. D., Joyce, A. R., and Palsson, B. Ø. (2006) PLoS Comput. Biol. 2, e101

33. Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J., and Palsson, B. Ø. (2004) Nature 429, 92–96

34. Kümmel, A., Panke, S., and Heinemann, M. (2006) Mol. Syst. Biol. 2, 2006.0034