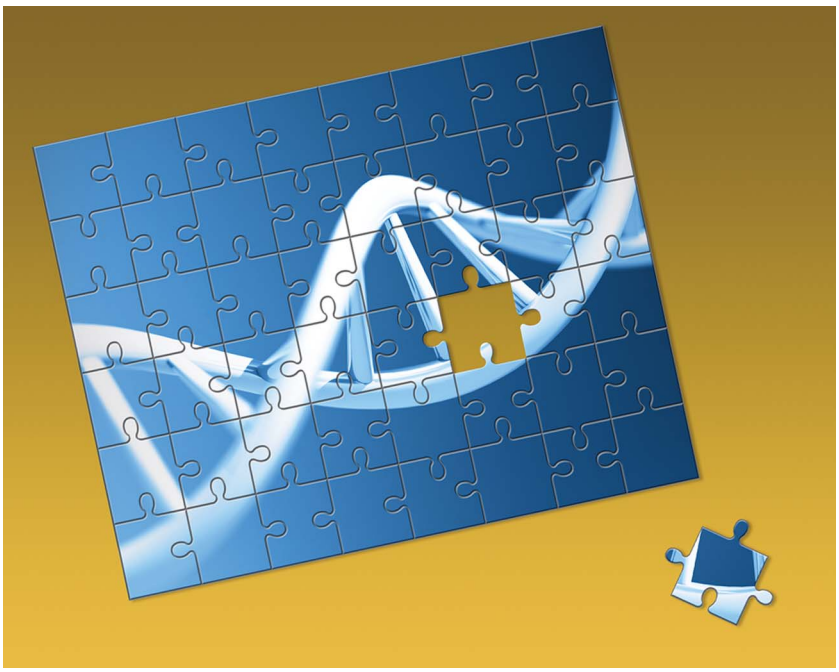
Solving Puzzles With Missing Pieces: The Power of Systems Biology

BY JAMES T. YURKOVICH
Department of Bioengineering,
University of California, San Diego, La Jolla, CA 92093, USA

BY BERNHARD O. PALSSON
Department of Bioengineering,
University of California, San Diego, La Jolla, CA 92093, USA



century of research in genetics, molecular biology, and biochemistry has enabled the genome-scale reconstruction of networks underlying well-studied cellular functions, such as metabolism. A quality-controlled reconstruction process effectively produces a circuit diagram of the metabolic network encoded in an organism's genome that can be modeled mathematically. Thus, a first principles "bottom-up" approach to systems biology rooted in fundamental mechanisms has arisen, and the quest to reveal the program that DNA encodes is underway. This article will familiarize you with some of the engineering concepts, methods, and applications in systems biology.

Life is a program written in DNA. Starting in 1995, genome sequences detailing this program have ushered in a new point of view in biology: a true systems-level, or "genome-scale," perspective. The genome sequence for an organism is analogous to having a component list for a circuit, except that many connections between components, and even some of the component functions themselves, are unknown. So how do we solve a puzzle when there are pieces missing? Enter systems biology. The combination of full genome sequences with over half a

I. GENOTYPE–PHENOTYPE RELATIONSHIP

The bottom-up approach to systems biology aligns with engineering thinking embodied in systems science. Systems biology aims to understand how all the molecules that make up a cell interact to form coherent physiological functions. Metabolic networks are made up of thousands of biochemical reactions that can now be "reconstructed" and converted into mathematical formats amenable to

Digital Object Identifier: 10.1109/JPROC.2015.2505338

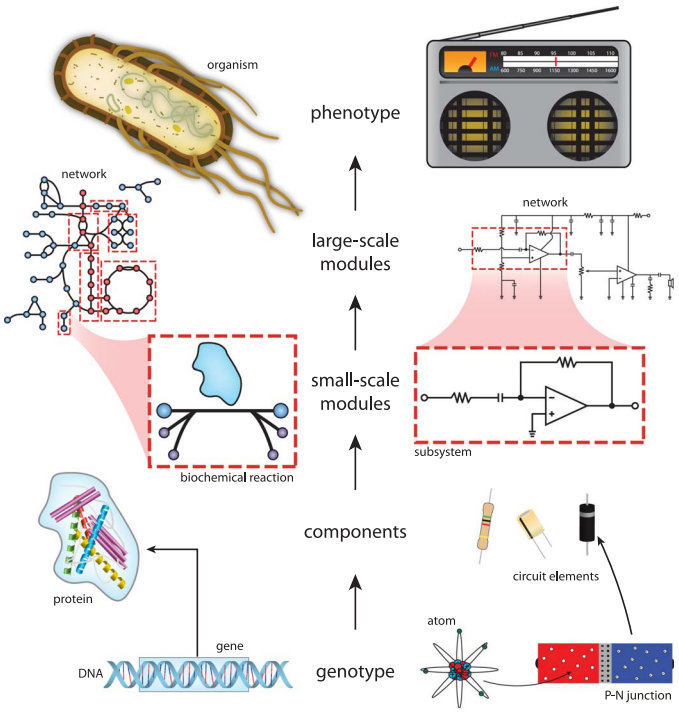


Fig. 1. *Genotype–phenotype relationship in biology (left) and an analogous view in an electrical system (right) allows for the modularization of a system over different complexity scales. This simplified representation omits several biological processes (e.g., transcription and translation) and engineering processes (e.g., software implementation).*

modeling. Because these models are built from first principles, they are able to describe the functional states of networks and therefore the systems-level behavior of the cell. Bottom-up systems biology is helping to unravel and understand the "genotype–phenotype relationship" on a genome-scale basis. The "genotype" of an organism (the collection of all genetic elements on a genome) contains the information that determines its form and function (the "phenotype"). Defining, understanding, and using this relationship is fundamental to systems biology.

The genotype–phenotype relationship is multiscale (see Fig. 1). At the smallest scale, molecular biology and biochemistry give us an understanding of DNA and how information in the form of genes and other genetic elements is encoded in it. These genetic elements have various structural and regulatory functions and encode the proteins that catalyze and facilitate biochemical reactions. At larger scales, these reactions form ever more complicated modules of biochemical functions in a network setting that together manifest an overall cellular behavior, or phenotype. Phenotypic states can thus be viewed as the result of running the "program" that is encoded in the DNA.

Taking a few liberties, we construct a simple analogy to relate the genotype–phenotype relationship to a familiar electrical engineering concept (see Fig. 1). In an electrical circuit, we can see a "genotype–phenotype" relationship emerge as we go from an atomic level to a component level to an engineering application. At the lowest scale of system complexity, we have the atoms that make up P-N junctions described by semiconductor physics. Using this information allows for the construction of both active and passive circuit elements. Together, these circuit elements can be arranged in a network from which the "phenotype" or systems function emerges: capturing and converting a signal to music through an amplifier and a speaker. As in an organism, the form and function of the radio is defined by the properties of its "genotype."

A fundamental paradigm for the implementation of systems biology on the genome scale has arisen [1], driven by the recent ability to generate data describing the many levels of biological complexity. First, all components within a cell (proteins, biochemical reactions, etc.) are enumerated and annotated. These components are then connected and used to reconstruct the network map. These reconstructed networks are translated into mathematical formats that describe the underlying biological knowledge. Finally, testable predictions are made using the mathematical models that describe the network. This process must be repeated for each new organism of interest. Well-studied model organisms such as *Escherichia coli (E. coli)* can serve as a Rosetta Stone for inferring the genetic content and function of poorly characterized organisms.

II. MODELING METHODS AND TECHNIQUES

Once a network is reconstructed, it is translated into a mathematical format using fundamental physical and chemical principles. As mentioned before, these reconstructed networks are incomplete. Finding the missing pieces and accounting for incomplete network structure requires the use of mathematics familiar to engineers to help construct a network that still carries biological meaning. Here, we describe some of the standard modeling tools and techniques, highlighting some of the differences that arise between biological systems and other engineering systems.

A. Reconstructing a Network

The reconstruction process is a system identification problem aimed at reverse engineering and inferring

biochemical network structure from first principles. Biologists and chemists study organisms on a molecular level to identify individual connections between molecules (reactions) and characterize the inputs and outputs. This data is stored in large repositories and is not always organism specific. Therefore, it must be manually curated to identify which reactions occur in a given network (i.e., some reactions may not be capable of occurring in a given organism and should not be included). This curation becomes a time-intensive process as even relatively simple bacteria such as *E. coli* have thousands of reactions. Like in electrical networks, connections between components are modular and can be linked to form the larger network (see Fig. 1).

The reconstruction process has been reduced to a standard operating procedure that can be followed to derive the network structure for new organisms [2]. The resulting networks are inherently incomplete because we simply do not have knowledge of all of the compounds and connections. Standard system identification techniques are used to expand models and infer missing content (referred to as "gap filling").

The major limiting factor for a reconstruction is the time it takes to complete—a reconstruction of the human metabolic network took a team of six people over two years! Further, the human aspect of the curation process leads to concerns regarding consistency and quality control among reconstructions. Thus, there is a strong need for tools that could automate the reverse engineering of biochemical networks to yield reconstructions; such tools would allow for the elucidation of the network structure of new organisms of interest without devoting hundreds of hours to the process.

B. Translating Into a Mathematical Format

Once the biochemical network has been reconstructed, it must be translated into a mathematical format that is amenable to modeling. As is common in simplified modeling of many engineering systems, the entire network can be captured in a matrix that represents the inputs and outputs (in this case, the stoichiometry of all reactions in the network). This stoichiometric matrix $\mathbf{S}$ is an incidence matrix with rows representing nodes and columns representing links (see Fig. 2). Because a given compound only participates in a handful of reactions, $\mathbf{S}$ is sparse. Further, almost all of the nonzero entries are either $+1$ or -1 (outputs are positive and inputs are negative by convention). The simple mathematical structure of $\mathbf{S}$ allows for manageable computation and compression of large networks. The formulation of a biochemical network as a connectivity matrix represents a huge leap forward because it enables the use of familiar systems engineering tools like loop analysis (see Fig. 2).

C. Dynamic Description of Biochemical Reactions

An important feature of $\mathbf{S}$ is the bilinearity of the reactions it represents. Two chemical components can react to produce a third, leading to more than two nonzero entries in the corresponding column of $\mathbf{S}$. The content of all the nodes in a system like metabolism (i.e., the concentrations of the corresponding compounds) can be represented by a system of ordinary differential equations (ODEs). These systems of ODEs can be numerically solved to provide an idea of the system's state at a given time [3]. Having a model that is able to predict the concentrations of metabolic nodes is powerful because the nodes in metabolism represent some of the primary targets for therapeutic drugs [4]. In a biological network, it is possible for individual nodes to accumulate or lose mass; that is, the flow of mass into the node may not match that of the outflow. The rates of intracellular node accumulation can impact steady-state models because of the timescales on which some reactions occur. This characteristic is often neglected in simplified modeling techniques, such as loop analysis.

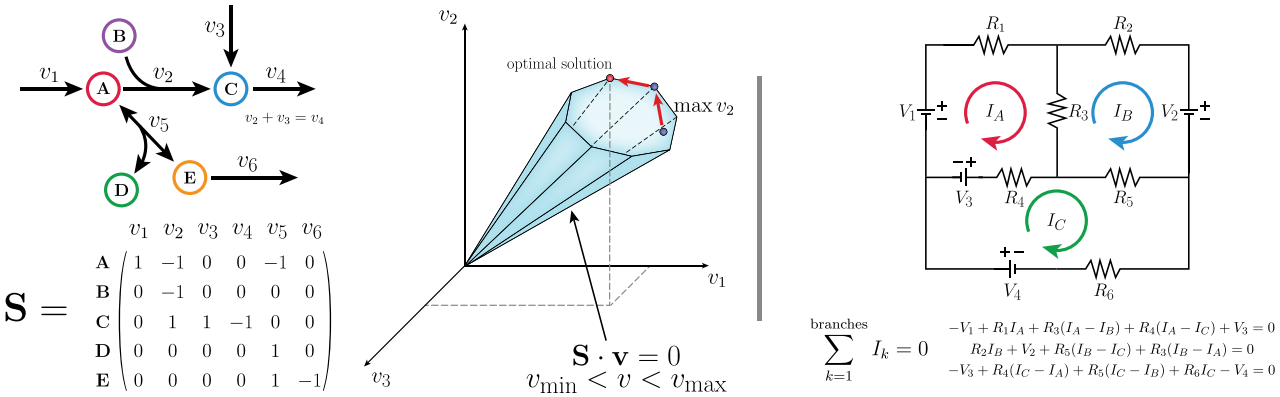


Fig. 2. *Loop and network analysis in systems biology (left) and electrical circuits (right). Some of the most powerful tools for analyzing systems-level models follow directly from Kirchhoff's laws.*

D. Modeling at the Genome Scale

In systems biology, it is generally impractical—if not impossible—to build true genome-scale ODE models. This impracticality is due in part to the lack of parameterization for individual reactions (initial conditions and rate constants are not known for every reaction). However, like in many engineering applications, modeling the system at steady state is often a good proxy for the most interesting biological state. Thus, other methods for genome-scale network analysis have been developed, benefitting from the application of well-developed systems tools such as hidden Markov processes [5] and convex optimization [6]. Some of these other modeling approaches do not require the extensive parameterization of ODE models and are therefore more amenable to modeling large, incompletely characterized systems.

One of the more successful methods for modeling genome-scale metabolic networks is to use constraint-based modeling to represent the flow of mass (the "flux") through every reaction in the network. Termed "flux balance analysis," this modeling strategy comes directly from the network structure and therefore bypasses the need for extensive parameterization [7]. With the network described in the form of a matrix, a simple matrix equation is used to model the system at steady state

$$\mathbf{S} \cdot \mathbf{v} = 0$$

where $\mathbf{S}$ is the stoichiometric matrix and $\mathbf{v}$ is a vector that represents the flux of each reaction in the network. Solving this simple matrix equation results in a solution space where each point in the space is a possible flux state of the steady-state system. Thus, a family of candidate solutions—rather than a single solution—is obtained. Constraints representing the properties of the biological machinery involved in each reaction are then imposed, resulting in a constrained solution space (see Fig. 2). The balanced network obeys Kirchhoff's laws: the flux around a metabolic loop must add to zero and the sum of the flux into a node must equal the sum of the fluxes leaving that node.

III. EXPLICITLY COMPUTING THE GENOTYPE–PHENOTYPE RELATIONSHIP

Constraint-based models make use of COnstraint-Based Reconstruction and Analysis (COBRA) methods [8] to simulate, analyze, and predict phenotypes. The variety of biological constraints applied to biochemical network reconstructions has grown from simply placing bounds on individual reaction fluxes to now include compartmentalization of molecules, mass conservation, and thermodynamic directionality of reactions. These additions have vastly increased the scope of biological questions that can be addressed using COBRA methods [9]. Applications often focus on perturbing individual components through gene deletion or addition and understanding how the effects propagate throughout the system. Such studies lead to novel predictions about the genotype–phenotype relationship that now can be tested experimentally on a large scale given the advances in genome editing [10].

The solution space of these constraint-based models is typically convex and can be characterized in several ways. One of the key capabilities of constraint-based models is the ability to construct an optimization problem to find the minimum or maximum flux for a reaction of interest (see Fig. 2). Systematically optimizing each reaction in the model to find the minimum and maximum feasible fluxes can therefore be used to characterize the solution space. This capability, for instance, allows for direct molecular engineering applications, such as coupling the production of valuable biomolecules with vital growth pathways of the organism. The suite of COBRA methods has led to a number of applications [4], the development of computational toolboxes [8], [11], and a series of scientific meetings focused on method development and applications [12].

The first-generation constraint-based models of metabolism incorporate what knowledge we have of how specialized proteins ("enzymes") facilitate and catalyze biochemical reactions. When constraints are placed on the flux through reactions in the network, these models provide useful and accurate phenotypic predictions [9]. The ability to optimize a model for a specific phenotype is useful for representing organisms possessing a clear biological objective; in many bacterial species, that objective is to maximize growth (the faster an organism can grow, the more successful it will be). To compute the growth rate, an objective function is defined and added to the model in the form of a reaction; that reaction can then be maximized as a function of important system inputs (see Fig. 3). These predictions can then be experimentally tested and have been successfully used in the design process of industrial production organisms [13].

While it is powerful to have the ability to find optimal states for industrial engineering applications (e.g., production of valuable biomolecules), defining an accurate objective function is very challenging. Life does not necessarily operate at an optimal growth state; instead, we often see that organisms operate at near-optimal growth states because they are trying to optimize for other functions (such as readiness for unforeseen environmental stresses). Thus, other engineering tools have been integrated into the suite of COBRA methods to expand the analytical capabilities so that additional states can be calculated. One such tool uses a Markov chain Monte Carlo method to sample the solution space, computing possible flux states of the network. There is currently high interest in computing these nonoptimal growth states, either through the development of new COBRA methods or by adding additional constraints.

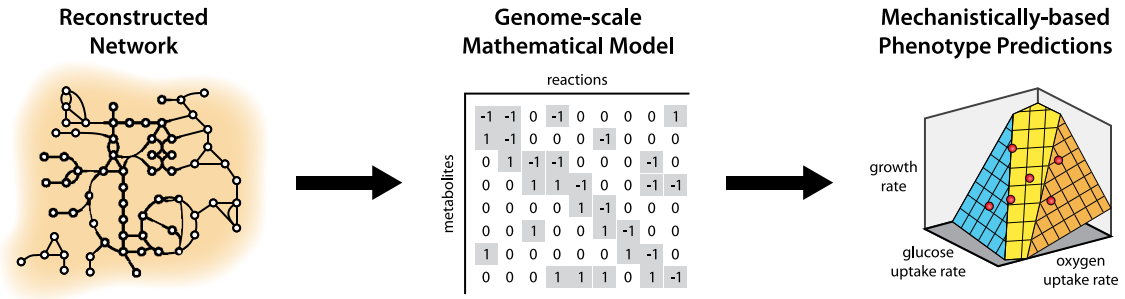


Fig. 3. *Genome-scale mathematical models allow for explicit computational modeling of the genotype–phenotype relationship. Optimizing ME models for an objective such as maximum growth rate provides mechanistic knowledge behind optimal functional states of both metabolic fluxes and expression levels as a function of important inputs such as glucose or oxygen [16].*

Recently, constraint-based models have been extended to account for the mechanistic detail of gene expression and protein synthesis (the processes of "transcription" and "translation," respectively). In other words, models can now account for more than just the network structure itself—they can compute both the cost of synthesizing all the machinery required for a particular state (the protein and enzyme demand) and the cost of running that particular state (how much flux is required for each reaction in the network). These next-generation "metabolism and expression models" (ME models) allow genome-scale reconciliation of molecular biology and biochemistry by explicitly accounting for the biological machinery responsible for gene expression and protein synthesis. The increasingly comprehensive ME models are able to provide better predictions for biological objectives, such as growth (see Fig. 3). As data for new organisms are generated and analyzed, ME models can be constructed for new organisms and systems, like human metabolism. Much exciting work lies ahead to automate the generation of ME models, to include new data types and biological knowledge, and to use the models to solve fundamental and applied problems in the life sciences.

IV. FUTURE OF SYSTEMS BIOLOGY

Systems biology moves away from studying biological systems on the scale of individual components and pathways and toward the scale of individual cells and, potentially, communities of cells. Studying biological systems on this scale elucidates the genotype–phenotype relationship in previously unattainable ways. Since the systems of interest are inherently multiscale, models must build upon layers of knowledge in a hierarchical fashion. In order to build a comprehensive genome-scale model, we must work to integrate as many disparate data types as possible. Several workflows designed to incorporate various data sets into computational models already exist [2], [14].

Reconstructions can grow markedly in scope based on modification of these workflows using new data streams. The field now has a diverse set of detailed models: from the simplest bacterium with 525 genes [15], to microbial communities such as the microbiome, to multicell and organ functions [16]. Models that include 3-D protein structures are emerging, allowing genome-scale models to include atomistic detail in protein structures and predictions of temperature sensitivity to growth [16]. More and more biological functions are likely to be built into large-scale cellular models as these workflows continue to develop.

One of the current challenges faced by the field of systems biology is a familiar one: How do we utilize the vast amounts of Big Data available to further our understanding of biology? Experimental data are being produced at a frantic pace, but the amount of knowledge gained from that data does not proceed at the same pace. The rate of knowledge accumulation (as measured by the doubling time of the number of COBRA publications) is half that of information processing [16]. Therefore, as with many other fields, there is great opportunity for the development of powerful computational methods to help extract the knowledge hidden in the data that has already been generated.

There is ample opportunity for experts in other fields to contribute to the advancement of systems biology. The mathematical language of systems science is natural to engineers; we routinely deal with system identification, build hierarchical models, solve optimization problems, and handle large data sets. Augmented with a clear understanding of biological characteristics such as those outlined above, engineers should be equipped to continue to impact the future development of systems biology and help figure out how to put the puzzle pieces back in place. ■

REFERENCES

[1] T. Ideker, T. Galitski, and L. Hood, "A new approach to decoding life: Systems biology," *Annu. Rev. Genomics Human Genetics*, vol. 2, pp. 343–372, 2001.

[2] I. Thiele and B. O. Palsson, "A protocol for generating a high-quality genome-scale metabolic reconstruction," *Nature Protocols*, vol. 5, no. 1, pp. 93–121, 2010.

[3] B. O. Palsson, *Systems Biology: Simulation of Dynamic Network States*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[4] A. Bordbar, J. M. Monk, Z. King, and B. O. Palsson, "Constraint-based models predict metabolic and associated cellular functions.," *Nature Rev. Genetics*, vol. 15, no. 2, pp. 107–120, 2014.

[5] M. Vidyasagar, *Hidden Markov Processes: Theory and Applications to Biology*. Princeton, NJ, USA: Princeton Univ. Press, 2014.

[6] J. R. Banga, "Optimization in computational systems biology," *BMC Syst. Biol.*, vol. 2, 2008, DOI:10.1186/1752-0509-2-47.

[7] J. D. Orth, I. Thiele, and B. Ø. Palsson, "What is flux balance analysis?" *Nature Biotechnol.*, vol. 28, no. 3, pp. 245–248, Mar. 2010.

[8] J. Schellenberger *et al.*, "Quantitative prediction of cellular metabolism with constraint-based models: The COBRA Toolbox v2.0," *Nature Protocols*, vol. 6, no. 9, pp. 1290–1307, Sep. 2011.

[9] N. E. Lewis, H. Nagarajan, and B. O. Palsson, "Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods," *Nature Rev. Microbiol.*, vol. 10, no. 4, pp. 291–305, 2012.

[10] O. Shalem, N. E. Sanjana, and F. Zhang, "High-throughput functional genomics using CRISPR-Cas9," *Nature Rev. Genetics*, vol. 16, no. 5, pp. 299–311, 2015.

[11] A. Ebrahim, J. A. Lerman, B. O. Palsson, and D. R. Hyduke, "COBRApy: COnstraints-based reconstruction and analysis for Python," *BMC Syst. Biol.*, vol. 7, no. 1, 2013, DOI:10.1186/1752-0509-7-74.

[12] Conference on Constraint-Based Reconstruction and Analysis (COBRA). [Online]. Available: http://www.aiche.org/sbe/conferences/conference-on-constraint-based-reconstruction-and-analysis-cobra/2015

[13] H. Yim *et al.*, "Metabolic engineering of Escherichia coli for direct production of 1,4-butanediol," *Nature Chem. Biol.*, vol. 7, no. 7, pp. 445–452, 2011.

[14] E. J. O'Brien, J. M. Monk, and B. O. Palsson, "Using genome-scale models to predict biological capabilities," *Cell*, vol. 161, no. 5, pp. 971–987, 2015.

[15] J. R. Karr *et al.*, "A whole-cell computational model predicts phenotype from genotype," *Cell*, vol. 150, no. 2, pp. 389–401, 2012.

[16] B. O. Palsson, *Systems Biology: Constraint-based Reconstruction and Analysis*, 2nd ed., Cambridge, U.K.: Cambridge Univ. Press, 2015.