Mechanistic modeling confronts the complexity of molecular cell biology

Robert D. Phair
Integrative Bioinformatics, Inc., Mountain View, CA 94041

ABSTRACT **Mechanistic modeling has the potential to transform how cell biologists contend with the inescapable complexity of modern biology. I am a physiologist–electrical engineer–systems biologist who has been working at the level of cell biology for the past 24 years. This perspective aims 1) to convey why we build models, 2) to enumerate the major approaches to modeling and their philosophical differences, 3) to address some recurrent concerns raised by experimentalists, and then 4) to imagine a future in which teams of experimentalists and modelers build—and subject to exhaustive experimental tests—models covering the entire spectrum from molecular cell biology to human pathophysiology. There is, in my view, no technical obstacle to this future, but it will require some plasticity in the biological research mind-set.**

Monitoring Editor
Jennifer Lippincott-Schwartz
National Institutes of Health

Received: Aug 29, 2014
Revised: Sep 15, 2014
Accepted: Sep 16, 2014

INTRODUCTION: WHY BIOLOGISTS BUILD MODELS

The word "model" means different things to different scientists—even to different modelers. My focus here is mechanistic mathematical models whose complexity and nonlinearity is sufficient to render classical mathematical analysis helpless and computation essential. Just as it was for physics in the 17th century and engineering in the 19th century, *complexity* is the inescapable reality that is driving cell biology toward modeling. We build models because the human brain struggles when 7 ± 2 processes interact (Miller, 1956). We build models because the universal scientific remit is accurate prediction despite incomplete knowledge and because we have found that well-tested mechanistic models are our best defense against the counterintuitive behavior of complex systems (Forrester, 1971).

Unambiguous communication is another important and underappreciated motivation for modeling. When we read prose descriptions of a working model toward the end of a scientific paper, it is unlikely we perceive exactly the idea the author intended. Diagrams are better than prose. Diagrams are, I think, the natural common language linking modelers and experimentalists, but diagrams are most effective when drawn using a standard notation (Kitano *et al.*, 2005). Otherwise, a diagram can become as imprecise as prose. When diagrams are drawn using a standard notation, such as Systems Biology Graphical Notation (SBGN; Le Novère *et al.*, 2009), each symbol represents a quantifiable phenomenon, and the corresponding differential equations can be constructed automatically. The next step is choosing appropriate rate laws for the biochemical reactions, transport processes, and binding interactions represented in the diagram. For example, an enzyme-catalyzed process might have a rapid-equilibrium, reversible Michaelis-Menten rate law incorporating the Haldane relationship to enforce the appropriate thermodynamic constraint. This rate law can incorporate inhibitors and activators and even posttranslational modifications of the enzyme (if V_{max} is written as $k_{cat}E$ and E is the solution of its own differential equation). Other processes will be characterized by binding constants or rate constants. The power of modeling arises from its ability to take all these into account simultaneously and make testable predictions.

Precise communication is so important to modelers and systems biologists that there are already curated international repositories of biological models (Le Novère *et al.*, 2006; Lloyd *et al.*, 2008; Yu *et al.*, 2011), nascent standards for encoding biological models (Lloyd *et al.*, 2004; Hucka *et al.*, 2010), and even early efforts to standardize graphical notation for model diagrams (Le Novère *et al.*, 2009; Wimalaratne *et al.*, 2009).

APPROACH: HOW BIOLOGISTS USE MODELS

Modelers distinguish between models of data and models of mechanism. Statistical models, like correlation and multiple regression and cluster analysis, as well as any effort to fit data directly to functions, like polynomials or Fourier series or sums of exponentials, are models of data. Bruce Alberts often reminds us, Data is not understanding.

DOI:10.1091/mbc.E14-08-1333

Address correspondence to: Robert D. Phair (rphair@integrativebioinformatics.com; http://integrativebioinformatics.com).

Abbreviation used: NIH, National Institutes of Health.

Likewise, models of data are not understanding; it remains notoriously difficult to extract cause and effect from statistical models. Mechanistic models, on the other hand, almost always comprise dynamic systems of ordinary or partial differential equations. Why is that?

Mechanistic models purport to represent causality. Time derivatives represent changes in biological variables. These changes can be seen as effects of the processes on the other side of the differential equation. In turn, the processes (binding, transport, and biochemistry are the three main process types) are written as functions of those same biological variables whose derivatives were specified. In this way, the inclusion of causal and feedback loops is automatic. Before the advent of the uncertainty principle, this view of the mathematics of causation was standard fare in college physics. Engineering and, later, biological modelers, largely immune to causeless quantum mechanics, rescued the paradigm and put it back to work.

Hence there is an underappreciated demarcation between statistical systems biology and mechanistic systems biology. Indeed, their worldviews are so distinct that these two branches of systems biology rarely meet or work together. This is unfortunate, because statistical surprises generate novel hypotheses that deserve incisive mechanistic tests.

Biological modeling is not monolithic. Some modelers feel that all the parameters should be measured experimentally before modeling begins. They are comfortable assuming that these in vitro measurements will apply in cells and recognize that a change in model structure may require new in vitro measurements. Others insist on a "minimal" model—one that has only as many parameters as can be resolved from the *current* data set and neither leverages nor is biased by previous work in the same field. Minimal models are small. They are tractable in the sense that we can "understand" them. But large models are inevitable, in my view, if biology aims to help the National Institutes of Health (NIH) achieve what the citizens expect.

Other groups aim at a "validated" model—one that has passed a second independent test. Still others see validation as inherently temporary. They view models as hypotheses that can sometimes be corroborated by experimental testing and are actually just as useful (perhaps more useful) when ruled out by such a test (Phair and Misteli, 2001; Anderson and Papachristodoulou, 2009).

A few paragraphs cannot do justice to the full family of modeling philosophies. But no matter which approach one chooses, experience suggests that the most effective strategy consists of *teams* of experimentalists and modelers working together closely (Phair, 2012). This is because we need both specialist depth and breadth of specialties to move successfully from reductionist to synthetic integrative work. Especially at the stage of model formulation, teams prevent key ideas (both physical and biological) from falling through the cracks.

It feels important for cell biology to encourage all modeling approaches. We want scientific progress to serve as the selection pressure. There is strength in diversity.

CONCERNS AND RESPONSES

Not everyone is convinced. Some biologists worry that it is too soon to model because we don't know all the parts yet. In 1865 Claude Bernard (Bernard, 1957) may have been the first great biologist to voice this concern, but modeling thrives on the unknown and does not require that we know all the parts. Modeling is quantitative hypothesis testing; it is classical scientific method combined with computation to help us to manage the enormous complexity of cell biology.

Another oft-heard concern is that we "don't want a lot of parameters whose values we don't know." Indeed, if your vision is something like the elegance of the ideal gas law, which has but one constant whose value is known to eight significant figures, you might well conclude that a pathway model with 100 or even just 20 parameters is an example of overfitting, a term that originated in statistics and describes a statistical model that is actually fitting the noise as well as the underlying relationship. The term has evolved to encompass any model that is thought to be too complex or to have too many parameters.

The criticism of overfitting usually reflects a modeling philosophy that derives from physics or from the natural human desire for simplicity. Simplicity and complexity are two distinct approaches to modeling; one is not better than the other. Nevertheless, physical law may be too high a standard for a field like cell biology, in which relative measurements (compared with control) are still far more common than absolute measurements with SI units, and absolute measurements of the same quantity can vary by a factor of five among laboratories. Modelers with backgrounds in physics see the goals of modeling differently from those with backgrounds in engineering. A physicist aims at something approaching a physical law; an engineer aims at a circuit diagram.

When a model is conceived as a working hypothesis (Popper, 1965; Phair, 1997; Phair and Misteli, 2001; Beard and Kushmerick, 2009) and is likely to be refuted by some future experiment, parameter identifiability (estimating numerical values with precision) is not a necessary goal. To me, a more practical modeling goal is discovering whether or not a proposed mechanism is actually capable of accounting for the data from many different experiments.

Another version of the "excessive parameters" objection was attributed to John von Neumann by Enrico Fermi: "With four parameters I can fit an elephant and with five I can make him wiggle his trunk." This dismissive epigram has entered the lore of biology and is often heard in the form, "With a complex model you can fit anything." Fermi and von Neumann were certifiably smart, but with five parameters one can describe at most two Krebs cycle reactions. A von Neumann elephant could neither oxidize pyruvate completely nor wiggle its trunk.

Modelers who work with the full complexity revealed by modern experimental cell biology have a very different experience. They do not find that a complex model can fit anything. The constraints imposed by hypothesizing a model with a specific mechanistic structure make it exceedingly difficult to discover even one set of parameter values that simultaneously accounts for any substantial number of different experiments (Alvarez-Vasquez *et al.*, 2004; Chen *et al.*, 2004; Wu *et al.*, 2007; Patterson *et al.*, 2008; Melnick *et al.*, 2009). For this reason, better computational tools for searching high-dimensional parameter spaces represent an active and essential research area (Oguz *et al.*, 2013).

CONCLUDING REMARKS: HOW MODELING CONNECTS CELL BIOLOGY TO HUMAN HEALTH

When I was a first-year assistant professor of physiology at Johns Hopkins, my lab was just one floor above Tom Pollard's. When I asked what distinguishes cell biology from other disciplines he said, "Cell biology is the study of the fundamental processes that every cell uses." Physiology, often conceived as dealing with organ systems, might equally well be seen as the study of what one cell does that others do not. To get from cell biology to physiology, we need increased focus on cell types. We need complete quantitative models of each cell type. Each of these models will leverage all of cell biology, but physiological models frequently entail multiple cell

types encoding what is unique about enterocytes, hepatocytes, and skeletal, cardiac, and smooth myocytes, adipocytes, neurons, renal tubular epithelial cells, endocrine and exocrine cells, stem cells, sensory cells, immune cells, endothelial cells, and osteoblasts or erythrocytes. There are more than 200 cell types in the human body. Perhaps cell type should be a required keyword for every poster at a meeting. Perhaps we can recruit the cell physiologists of the world to join us.

Modeling connects cell biology to the mission of the NIH by covering the essential biomedical spectrum from molecular cell biology to pathophysiology. This spectrum is essential in the sense that pharmaceuticals and biologics work at the level of molecules, while disease manifests at the level of physiology. This spectrum is the domain that systems biologists call multiscale modeling. It covers concentration scales from fM to mM, time scales from milliseconds to years, and length scales from nanometers to meters—factors of ~10^{10} each. We (Chasson and Phair, 2001) and many others (Slepchenko *et al.*, 2003; see also http://sbml.org/SBML_Software_Guide/SBML_Software_Matrix) have built software tools that aim to support modeling of physiology, building from molecular cell biology upward. We are currently focused on the cell biology of hepatocytes and the metabolic consequences of obesity for plasma lipoprotein metabolism, cholesterol trafficking, and mitochondrial energy metabolism. The challenges are enormous, but new ideas from engineering, physics, computer science, and mathematics appear seemingly every day. It is, as has been said about the whole of systems biology, "a good field for those seeking risk and adventure" (Kirschner, 2005, p. 504).

The year 2015 will be the sesquicentennial of Bernard's assertion that it is too soon to apply quantitative modeling to biology. But he also predicted the time would come. Successful combinations of modeling and experiment are now so common in the research literature that even Bernard would surely agree the time is now.

REFERENCES

Alvarez-Vasquez F, Sims KJ, Hannun YA, Voit EO (2004). Integration of kinetic information on yeast sphingolipid metabolism in dynamical pathway models. J Theor Biol 226, 265–291.

Anderson J, Papachristodoulou A (2009). On validation and invalidation of biological models. BMC Bioinformatics 10, 132.

Beard DA, Kushmerick MJ (2009). Strong inference for systems biology. PLoS Comput Biol 5, e1000459.

Bernard C (1957). An Introduction to the Study of Experimental Medicine, New York: Dover.

Chasson AK, Phair RD (2001). ProcessDB: a cellular process database supporting large-scale integrative kinetic modeling in cell biology. Proc 2nd Internat Conf Syst Biol Caltech, Pasadena, CA, Nov 4–7, 124. http://icsb-2001.org/Posters/114_chasson.pdf.

Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004). Integrative analysis of cell cycle control in budding yeast. Mol Biol Cell 15, 3841–3862.

Forrester JW (1971). World Dynamics, Cambridge, MA: Wright-Allen.

Hucka M, Bergmann FT, Hoops S, Keating SM, Sahle S, Schaff JC, Smith LP, Wilkinson DJ (2010). The Systems Biology Markup Language (SBML): language specification for level 3 version 1 core. Nat Precedings, doi:10.1038/npre.2010.4959.1 (accessed 24 October 2014).

Kirschner MW (2005). The meaning of systems biology. Cell 121, 503–504.

Kitano H, Funahashi A, Matsuoka Y, Oda K (2005). Using process diagrams for the graphical representation of biological networks. Nat Biotechnol 23, 961–966.

Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, *et al.* (2006). BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. Nucleic Acids Res 34, D689–D691.

Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, *et al.* (2009). The Systems Biology Graphical Notation. Nat Biotechnol 27, 735–741.

Lloyd CM, Halstead MDB, Nielsen PF (2004). CellML: its future, present and past. Prog Biophys Mol Biol 85, 433–450.

Lloyd CM, Lawson JR, Hunter PJ, Nielsen PF (2008). The CellML Model Repository. Bioinformatics 24, 2122–2123.

Melnick M, Phair RD, Lapidot SA, Jaskoll T (2009). Salivary gland branching morphogenesis: a quantitative systems analysis of the Eda/Edar/NFκB paradigm. BMC Dev Biol 9, 32.

Miller GA (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63, 81–97.

Oguz C, Laomettachit T, Chen KC, Watson LT, Baumann WT, Tyson JJ (2013). Optimization and model reduction in the high dimensional parameter space of a budding yeast cell cycle model. BMC Syst Biol 7, 53.

Patterson GH, Hirschberg K, Polishchuk RS, Gerlich D, Phair RD, Lippincott-Schwartz J (2008). Transport through the Golgi apparatus by rapid partitioning within a two-phase membrane system. Cell 133, 1055–1067.

Phair RD (1997). Development of kinetic models in the nonlinear world of molecular cell biology. Metabolism 46, 1489–1495.

Phair RD (2012). Why and how to expand the role of systems biology in pharmaceutical research and development. Adv Exp Med Biol 736, 533–542.

Phair RD, Misteli T (2001). Kinetic modelling approaches to in vivo imaging. Nat Rev Mol Cell Biol 2, 898–907.

Popper KR (1965). The Logic of Scientific Discovery, New York: Harper & Row.

Slepchenko BM, Schaff JC, Macara I, Loew LM (2003). Quantitative cell biology with the Virtual Cell. Trends Cell Biol 13, 570–576.

Wimalaratne SM, Halstead MDB, Lloyd CM, Cooling MT, Crampin EJ, Nielsen PF (2009). A method for visualizing CellML models. Bioinformatics 25, 3012–3019.

Wu F, Yang F, Vinnakota KC, Beard DA (2007). Computer modeling of mitochondrial tricarboxylic acid cycle, oxidative phosphorylation, metabolite transport, and electrophysiology. J Biol Chem 282, 24525–24537.

Yu T, Lloyd CM, Nickerson DP, Cooling MT, Miller AK, Garny A, Terkildsen JR, Lawson J, Britten RD, Hunter PJ, *et al.* (2011). The Physiome Model Repository 2. Bioinformatics 27, 743–744.