Anthony Finkelstein
James Hetherington
Linzhong Li
Ofer Margoninski
Peter Saffrey
Rob Seymour
Anne Warner
University College London

Computational Challenges of Systems Biology

Progress in the study of biological systems such as the heart, brain, and liver will require computer scientists to work closely with life scientists and mathematicians. Computer science will play a key role in shaping the new discipline of systems biology and addressing the significant computational challenges it poses.

Bioinformatics is the computing response to the molecular revolution in biology. This revolution has reshaped the life sciences and given us a deep understanding of DNA sequences, RNA synthesis, and the generation of proteins. In the process of achieving this revolution in understanding, we have accumulated vast amounts of data.

The scale of this data, its structure, and the nature of the analytic task have merited serious attention from computer scientists and prompted work in intelligent systems, data mining, visualization, and more. It has also demanded serious efforts in large-scale data curation and developing a worldwide infrastructure to support this. Bioinformatics, the handmaiden of molecular biology, poses novel computational challenges, stretches the state of the art, and opens unanticipated uses of computing concepts. In tackling these, computer scientists have the additional satisfaction of contributing to a scientific Grand Challenge.

Bioinformatics is, however, only the first step in reshaping the life sciences. For further progress, we must return to the study of whole biological systems: the heart, cardiovascular system, brain, and liver—systems biology. To build an integrated physiology of whole systems, we must combine data from the many rich areas of biological information. Alongside the genome, which constitutes our knowledge about genes, we place the *proteome*, *metabolome*, and *physiome*, which embody knowledge about proteins, metabolic processes, and physiology.

Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention. Progressing in this discipline will involve computer scientists working in close partnership with life scientists and mathematicians. In contrast to the molecular biology revolution, computer science will proactively engage in shaping the endeavor rather than just clearing up afterwards!

The prize to be attained is immense. From in silico drug design and testing to individualized medicine that will take into account physiology and genetic profiles, systems biology has the potential to profoundly affect healthcare and medical science generally.

THE ROLE OF MODELING

Suppose we had a catalog of all the gene sequences, how they translate to make proteins, and which proteins interact with each other. Further, assume

Published by the IEEE Computer Society
0018-9162/04/$20.00 © 2004 IEEE

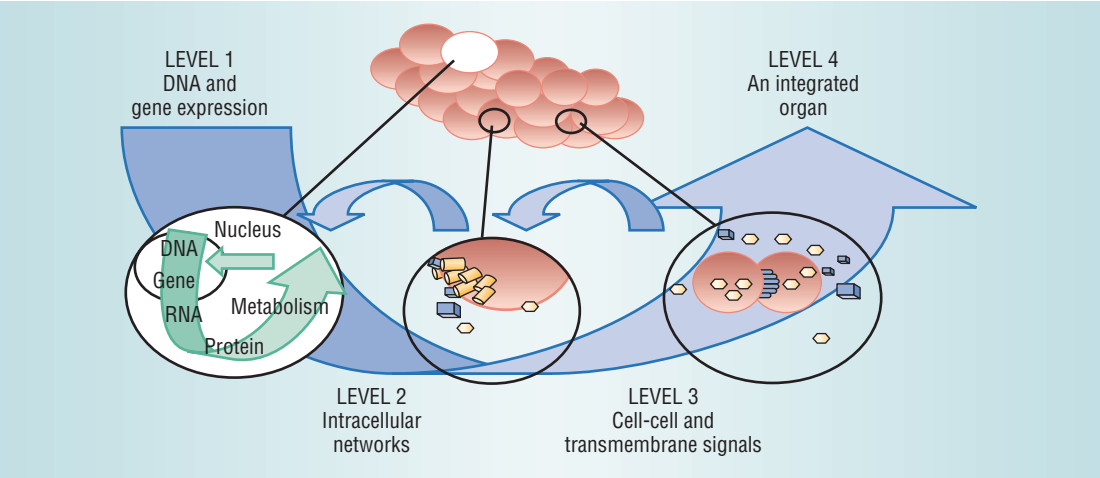

Figure 1. Building models in systems biology. The models should span from DNA and gene expression to intracellular networks to cell-to-cell and transmembrane signals and through to the organ level.

we know the way in which the protein backbones fold—whether into sheets, helices, or other shapes with differing properties. For several reasons, we would not be able to put them into a functionally meaningful framework simply from the data.

First, all proteins undergo post-translational modification that adds side chains like sugars to make, for example, *glycoproteins*—important constituents of cell membranes. These additions influence the shape and properties of proteins and hence their function and behavior. Further, just because two proteins can interact in principle does not mean that they do so in real cells. Also, metabolic processes synthesize many small, functionally important molecules. For example, many neurotransmitters are made by cells, not translated from RNAs. Biological systems are so enormously complicated that, however much we learn about them, it will be impossible to create a full simulation based on complete understanding.

Thus, a bottom-up, data-driven strategy will not work. We cannot build an understanding of biological systems from an understanding of the components alone. We must seek other approaches.

Modeling lies at the heart of systems biology. We can use experimental information to build models at different biological scales, integrating them to create an orchestrated assemblage ranging from gross models of physiological function through detailed models that build directly on molecular data. As Figure 1 shows, in principle these models should span from DNA and gene expression to intracellular networks, to cell-to-cell and transmembrane signals, and through to the organ level. Tenuously, we might eventually construct such models at the organism level.

We thus introduce two key concepts for systems biology, methodologies forced upon us by the peculiar complexity of biological systems. First, we acknowledge the importance of simplification because biological complexity requires us to model, not simulate. Second, we acknowledge the importance of both modularity and the integration of modules. Biological complexity requires us to break our systems into manageable components, but it also requires us to reassemble them because behaviors can emerge that we cannot understand from the components alone.

The resulting models can provide coarse-grained prediction, be used as a scaffold for our emerging understanding of the data, identify gaps in our biological knowledge, and, if the models are good, predict new behaviors that we can explore experimentally. Iteration between model and experiment provides the key to ensuring that models are realistic. Given that researchers may need a different technique to study each component, it is difficult if not impossible to undertake physiological studies of whole systems in which the individual components are monitored simultaneously.

This agenda poses some serious challenges to the construction, integration, and management of the models—challenges that computer scientists are well placed to meet.

MODELING STATE OF THE ART

Denis Noble and colleagues[1] developed the heart model that provides the paradigmatic example of systems biology. Their work provides a computational model of the heart's electrical and mechanical activity when healthy and when diseased. The model has been linked to sophisticated visualizations, particularly solid geometry models. It has also proven invaluable in developing an understanding of cardiac arrhythmia, with consequences for drug design and testing.[2]

The model itself has evolved from its relatively simple beginnings as an adaptation of the classic Hodgkin-Huxley squid axon model[3] to its current form, which involves hundreds of equations and adjunct models. Despite this sophistication and the large amount of effort it has consumed, the model only covers a small part of the heart's mechanical, electrophysiological, and chemical phenomena.

In addition to revealing what researchers can achieve, the heart model also suggests the scale of

Iteration between model and experiment provides the key to ensuring that models are realistic.

the challenge that systems biology presents. It has been the seed for the Physiome project,[4] which collects and catalogs biological models and supports access to these models. The Physiome project also provides Web-accessible databases of biological data that researchers can potentially link to models.

Other researchers have produced a plethora of stand-alone models to simulate various biological phenomena. Although most are relatively simple, some models demonstrate more sophistication. One example, the bacterial model that Dennis Bray and colleagues created,[5] simulates chemosensitivity and the motion of flagella, the thin projections from cells.

Many models are provisional, in that they embed contested hypotheses about biological function or structure or are otherwise only partially validated. Stand-alone biological modeling has attracted some attention from computer scientists. In particular, certain biological phenomena such as biochemical networks appear to lend themselves to representation in formal schemes such as process calculi, opening the possibility for formal analysis and reasoning—an avenue some researchers have already pursued.[6] Only a small proportion of stand-alone models are accessible to those outside their development groups or have been documented in a form other than the scientific papers in which they originally appeared.

MODEL INTEGRATION

Although vital to systems biology, model integration has only recently received the attention it deserves. In general, ad hoc, handcrafted, tightly coupled integration of stand-alone models is the state of the art. The Systems Biology Workbench project seeks to advance the practice of model integration. This project consists of two distinct components.

The Systems Biology Markup Language[7] is an XML language for representing biochemical network models. SBML has largely been driven by a pragmatic concern to facilitate the exchange of models across a range of popular modeling tools, and it has achieved some success in this regard. The Systems Biology Workbench[8] provides a software framework for interoperation among the heterogeneous tools and resources used in biological modeling.

The SBW standard is not tailored to biological modeling, but instead provides a generic middleware solution. Although neither SBML nor SBW focuses on model integration directly, SBML provides a common framework for documenting a small range of models, which is an important first step toward model integration.

Another approach, developed in parallel with the Physiome Project, has resulted in CellML.[9] This XML-based language seeks principally to store, exchange, and ultimately reuse biological models. CellML provides a high-level block-diagram representation scheme in which researchers can assemble and hierarchically compose networks of models. It uses the XML namespace mechanism to embed other languages such as Math ML. Some attention has been directed to descriptive metadata, but this remains a less-developed aspect of the project.

Unlike SBML, CellML explicitly attacks the model integration problem. Like SBML, however, CellML can only encompass a limited range of models that exclude, for example, discrete-event systems. CellML is less widely used than the more pragmatically driven SBML.

IDENTIFYING CHALLENGES

To map out the systems biology space more systematically, and to identify the computational challenges more precisely, we use the high-level information model shown in Figure 2. The metamodel is presented using a stripped-down entity-relationship modeling convention.

Model characteristics

Our information model has three overlapping regions, each representing a key concern in systems biology and consisting of several components:

- *construction*—the model, compound model, scheme, constraints, and view components;
- *analysis*—the model, context, engine, interpretation, and ground components; and
- *validation*—the model, aspect, observation, assumptions, and interpretation components.

Models represent *aspects,* a term that denotes a coherent set of properties or phenomena of biological interest. The aspect anchors the model in the real world. We establish a correspondence through an *ontology*, an explicit formal specification of how to represent the objects, concepts, and other entities assumed to exist in the biological domain being studied and the relationships that hold among them. The model and appropriate elements must then be linked to elements in the ontology.

Assumptions condition or determine the relationship between models and the aspects they represent. Assumptions underpin model construction,

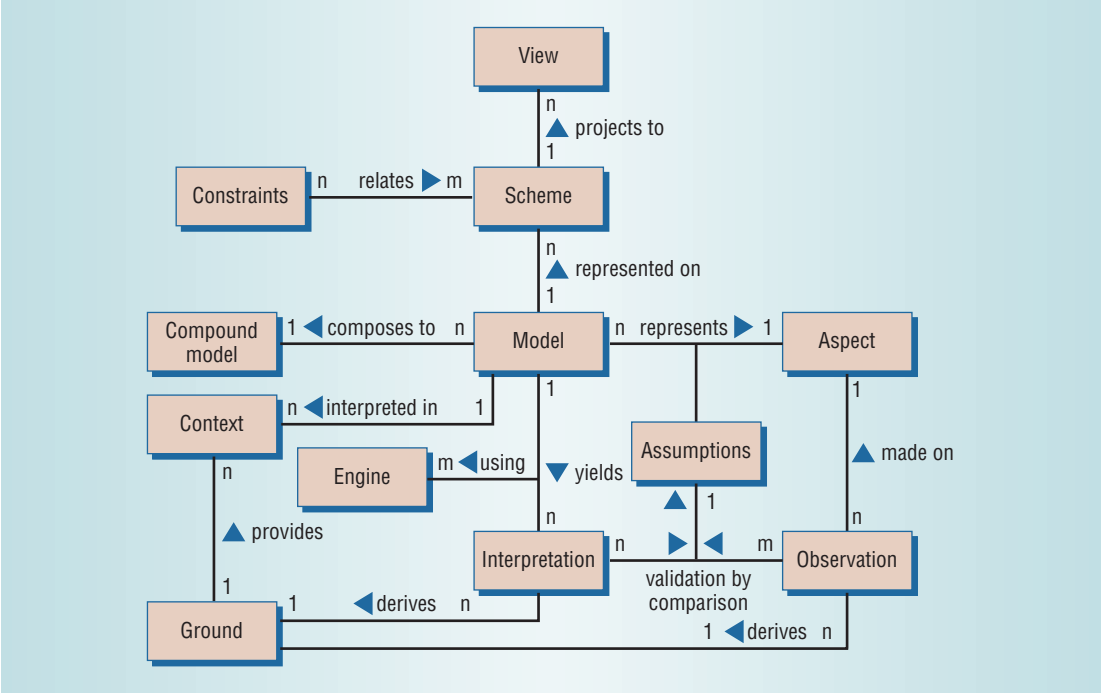

Figure 2. Systems biology metamodel, presented using entity-relationship modeling. This model identifies key concepts in systems biology and their relationships.

constitute the rationale for the model, and must be precisely documented and connected to the model for it to have meaning beyond the immediate use to which it has been put.

Experimental biologists make *observations* about phenomena of biological interest. Classically, these observations are used to validate interpretations derived from models. Commonly, however, models yield interpretations that prompt further observations or, when compared with observations, question the validity of the assumptions. Researchers document the observations in the scientific literature and in data resources associated with the experiments.

One of systems biology's central challenges involves the tie between descriptions of experiments, observations, experimental data, interpretations derived from models, and assumptions. In short, systems biology cannot be viewed independently of an information management framework that embraces a significant part of the experimental life sciences.

Acquiring scientific knowledge is an inductive process in which observations that agree with a model add to our confidence that it provides a good reflection of the system it describes. Thus, validation is a more troubled concept because it involves a matter of degree rather than certainty.

In principle, refutation is much simpler, but researchers must take care when deciding how best to modify the model to account for a disagreement between a previous version and observation. Many believe that in these circumstances modeling becomes most useful for developing scientific understanding. If we put our best scientific understanding into a model, and it does not fit the data, it suggests that our understanding is incomplete. This can be a powerful guide to new theories and experiments.

Models, once instantiated, yield *interpretations* through analysis. This can be a dynamic simulation process or a static mathematical reasoning process. The engine that both encompasses and executes a model determines the analytic process. Researchers can analyze the same model in many different ways using different procedures. The engine thus conditions an interpretation. We must precisely specify the engine to anchor the interpretation. In short, defining the model is insufficient—we must define how we use the model. Analysis can require significant computational resources.

Context is the data required to produce a model instance—it is the input to the model. Researchers could derive a context from observation, as in the straightforward case where experimental results provide a *ground* for data supplied to a model. In an alternative and somewhat more complex case, one model yields interpretations that constitute the context for another model. From an informational standpoint, we need to track the contexts supplied to the model and associate them with the interpretations to which they correspond. To maintain validation integrity, we must also track the context elements through their grounds.

Models are constructed in different languages, or *representation schemes*, each appropriate to the expression of and reasoning about different sets of properties. No universal language for systems biology can capture the many different phenomena we seek to explore.

We present these schemes through *views* defined as projections on the underlying scheme.

Computer scientists have developed techniques that can extend the schemes used in systems biology.

Modeling schemes relate to each other through *constraints* that define what it means for models in these schemes to be consistent with each other. Most schemes for modeling in the large provide a compositional mechanism that researchers can use to compose models and construct larger-scale *compound models*.

Modeling challenges

We are faced with three challenges:

- defining and managing the views, languages, and constraints;
- providing the means for checking the constraints and devising modeling schemes with sound compositional mechanisms; and
- managing models that may not be consistent with each other, either across schemes or across scales.

There is ample scope to extend the range of modeling schemes used in systems biology. Computer scientists have developed an extensive arsenal of formal modeling techniques that can be usefully employed here.

This complex picture excludes two key dimensions, however. Models may be produced in different versions over time and by different teams. Disagreements can arise and observations can be contested. Different researchers may generate models in different versions and configurations.

These unpredictable factors mean that systems biology is unlikely to produce a set of canonical models. Rather, a complex ecology of models embedded within a framework that enables debate and collaboration among contributors will arise. Ultimately, our objective might include individualized models that account for variations in physiology, rather than generic models of biological phenomena.

MODELING THE LIVER

As a first step in crafting a meaningful research agenda, we need further convincing exemplars of systems biology of the general type of the heart model. Such examples will necessarily be restricted in scope and scale. Ideally, however, they will be more explicitly engineered, with some systematic modularity and separation of concerns among component models. These models can then act as test beds for the broader conception of systems biology and for the information management frameworks that must accompany it.

The UK Department of Trade & Industry is supporting high-adventure science Beacon projects that offer the possibility of advances with significant industrial potential. One such project at University College London focuses on producing a physiological model of the human liver that is integrated across scales.[10] The project brings together physiologists and experimental life scientists, engineers with expertise in systems modeling, applied mathematicians with an interest in integrating models across differing temporal and spatial scales, and computer scientists who can build and deploy the information management and computational infrastructure.

The liver has been selected as an exemplar of systems biology because it is medically important and has a relatively homogeneous structure. Primarily a chemical system, the liver offers a more challenging subject than electromechanical organs such as the heart. Electromechanical systems have a long history of quantitative description and modeling, and research in this area is comparatively advanced. Several ongoing efforts also seek to build in vitro livers, artificial organs that patients recovering from liver damage can use. Researchers could use models to understand and overcome some of the problems experienced by those who build such livers.

The human liver has three principal functions:

- storing materials for release into the blood stream when needed;
- synthesizing proteins and peptides from amino acids; and
- detoxifying the system by breaking down harmful materials such as alcohol, which are then excreted.

Examining an example of the first function—glucose release from the hepatocytes, liver cells, in response to circulating adrenaline or glucagon—helps illustrate current work on systems biology.

Adrenaline triggers the classic fight-or-flight response to stress. Glucagon contributes to the homeostatic control of blood glucose. Both these systems are compromised in diabetes when the cellular uptake of glucose, driven by insulin, is defective.

Both adrenaline and glucagon activate the same intracellular mechanisms: These hormones, circulating in the bloodstream, bind to specific receptors on the hepatocyte's membrane. As a result, ion channels—specialized protein molecules that let specific ions enter or leave cells—open in the membrane. Calcium enters the cell through these channels, raising the concentration of calcium in the

cytoplasm—that is, the cellular material located within the cell membrane but outside the nucleus.

The binding of adrenaline or glucagon to receptors simultaneously activates linked G-proteins and initiates a chain reaction within the cell, releasing calcium and causing an increase in cytoplasmic calcium. At different concentrations, calcium both stimulates and inhibits calcium release from stores, causing cytoplasmic calcium levels to oscillate. The increase in calcium also mobilizes glucose release from glycogen, the stored form of glucose, which leaves the cell on glucose transporters.

This abbreviated description shows the complexity of the dynamic relationships involved in a relatively straightforward physiological process. Researchers can construct models of each of these subprocesses, such as G-protein activation or cytoplasmic calcium oscillation, in isolation. Typically, researchers model these subprocesses as ordinary differential equations, although certain processes appear to lend themselves to discrete event modeling.

The processes have, in this case, been well studied experimentally. Thus, researchers can relate the parameters that constitute the context systematically to values in the literature. Ideally, this should be done using a mediating ontology. Several significant projects are constructing such ontologies for human physiology, including, for example, the Digital Anatomist Foundational Model.[11] The richer ontologies developed for genetic and bioinformatic work, such as the Gene Ontology,[12] can also be useful for cell physiological work.

We must, however, look carefully at the reliability of the experimental data when selecting the parameters to use with the model. Assuming homogeneous models of the subprocesses, we can connect them to build a detailed model of the entire network. Representational heterogeneity naturally makes this more difficult. The resulting model can be investigated numerically by varying its context.

Alongside this model, we can build a simplified model. To make the system piecewise linear, we assume that ion channel opening, protein activation, and so on behave as perfect switches. The simplified system is biologically unrealistic, and many features, such as the shape or period of oscillations, are lost. Some features are retained, however, and we can use algebraic analysis to develop an understanding of the system. For example, we can learn how certain elements of the context control specific features of the system's behavior. Even in the absence of analytical results, a model simple enough to hold in the human mind provides a useful tool for understanding and as a comparator to the fuller, more unwieldy model.

Both the detailed and simplified models are constructed and analyzed using standard tools for scientific modeling, which must be wrapped to support model integration. They also must be connected to standard scientific visualizations, such as graphs or more sophisticated animated views.

We intend to take modeling of this system much further. An immediate extension will incorporate the homeostatic activation of glucose release through glucagon receptors. We could, for example, build models of gap junctions, which are constructed from *connexins,* membrane-inserted proteins that bridge the space between cells and provide direct channels through which the cytoplasm of one cell communicates with that of adjacent cells.

We could use this model to link more than one cell and scale up to multicellular models. Another approach to the scaling issue would consider the effects that signaling molecules have on gene expression by acting as transcription factors—proteins that bind to regulatory regions—thus moving down to the molecular machinery.

Judicious simplification will drive the art of systems biology.

MODELING STRATEGIES

Representing all aspects of a biological system in the smallest conceivable detail is infeasible, even when the data is available. We cannot and need not recreate the world as an isomorphic in silico image of itself. Therefore, judicious simplification will drive the art of systems biology. This is particularly true when trying to link different processes at different spatial or temporal scales, such as gene and protein networks.

Simplification

Selecting the appropriate simplification will depend on the topic being researched. For example, to represent biochemical networks that involve many different proteins, we could model the interactions between proteins as simple stimulus-response functions. Alternatively, we could choose to focus on a few proteins and model the extremely complex transformational processes between them in great detail.

Model simplification has at least three facets:

- *Choosing a modeling scheme.* The scheme must provide sufficient descriptive fidelity, flexibility when linking to other models, contextualization in terms of known or obtainable data, and reasonable ease of interpretation.

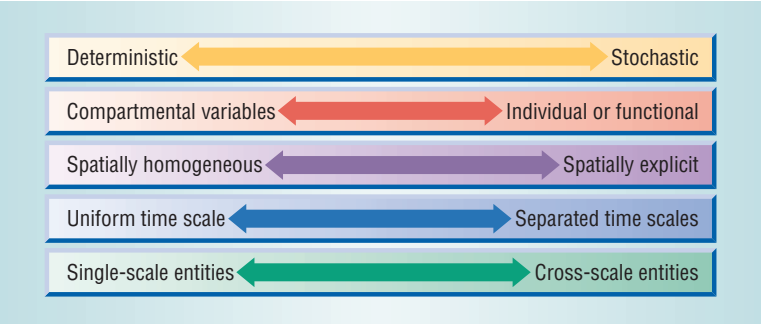

Figure 3. Taxonomic framework for modeling schemes. The framework contrasts modeling schemes based on different principles, such as spatial and temporal parameters.

- *Choosing a level of detail.* The choice of level of detail within a given representation determines how many links in a signaling pathway to represent explicitly, if and how to model space, and the dominant timescale.
- *Determining sensitivity.* A useful simplification scheme must have robust context and interpretation. The model must include the backbone elements that give robustness to the real biological system.

Some of these issues can be clarified by thinking about the interpretation obtained from the simplified model of calcium oscillations with square waves. One value of this model derives from its position at the extreme end of a continuum of models with Hill-function response functions, all of which behave with qualitative similarity. Thus, the models represent a continuum, only one of which provides a true representation of the real world.

That all these models behave qualitatively in the same way tells us that, in some sense, the detail of the real-world response may be incidental: There may be recognizable and potentially real worlds in which calcium oscillations differ. Thus, some perfectly feasible creature may have square calcium waves.

We can therefore attribute the human wave shape to some kind of fine tuning. We must determine how strenuously we should chase this kind of fine-scale effect, rather than being content with more robust, qualitative phenomena. Deciding which behaviors a model must reproduce can be difficult. This is, in our experience, an area where researchers from different backgrounds often disagree strongly.

Construction

Function can be an important guide for model construction and interpretation. That is, we know roughly what a liver is for. With other study subjects, however, this may not be the case. At the fine grain in biological systems, we can observe phenomena whose function we do not understand. In a deep sense, these phenomena may not be "for" anything—there is no logic to evolution. If we don't know which phenomena are central and which incidental, assessing the model's validity becomes extremely difficult.

In many cases, we must model both the physiological process and the experimental protocol. For example, we conducted an experiment that, when taken at face value, seemed to refute an assumption underpinning a model of protein production by cells. This, however, turned out to be true only if we interpreted the model in the most naive way. It is not always clear just what an experiment does and does not tell us about a model. More sophisticated interpretations, involving the explicit representation of stochastic effects, offer a means for analyzing the laboratory experiment and its predicted result while remaining compatible with the original hypothesis.

Integration

Our framework represents the relationships between models in different schemes in terms of constraints that define what it means if those models are consistent when we place them in conjunction with each other. Expressing these constraints, or understanding how the models relate, poses many difficulties when we are integrating different kinds of models. Figure 3 shows a simple taxonomic framework that contrasts modeling schemes based on different principles.

Problems arise when working with stochastic models or models formulated so that some act as discrete-time systems and others as continuous-time systems. Our strategy assumes that designing coherent collections of models is preferable to struggling to integrate fundamentally incompatible schemes. What these should be, and how they should be structured, remain open questions.

Although we can identify some important staging posts, systems biology has, in contrast to projects that map genomes, no clear end point. Models that provide thin vertical slices across scales offer one possibility. Our models of glucose release in the hepatocyte already approach cross-scale integration from gene expression through multicellular responses. Another example we are working on is fluid transport, a key part of liver physiology.

In the past five years, life scientists have identified the genes for aquaporins, the membrane water channels that control the movement of water into and out of cells. Moving from the gene through aquaporin models to bile flow would be a significant achievement.

An important staging post could be achieved by developing drug testing models that would satisfy

the strict requirements of validity, reliability, transparency, and traceability. Establishing global *collaboratories* in which researchers can exchange, review, and analyze models would also be significant. Finally, when we can use our models to dependably diagnose health issues and identify novel treatments, systems biology will have come of age. ■

References

1. D. Noble, "Modeling the Heart: From Genes to Cells to the Whole Organ," *Science*, vol. 295, 2002, pp. 1678-1682.
2. D. Noble and T. Colatsky, "A Return to Rational Drug Discovery: Computer-Based Models of Cells, Organs and Systems in Drug Target Identification," *Emerging Therapeutic Targets*, vol. 4, 2000, pp. 39-49.
3. A. Hodgkin and A. Huxley, "A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve," *J. Physiology*, vol. 117, 1952, pp. 500-544.
4. The Physiome Project, 2003; www.physiome.org.
5. M. Levin et al., "Origins of Individual Swimming Behavior in Bacteria," *Biophysical J.*, vol. 74, 1998, pp. 175-181.
6. C. Priami, ed., "*Computational Methods in Systems Biology, Proc. 1st Int'l Workshop Computational Methods in Systems Biology*, Springer, 2003.
7. SBML: Systems Biology Markup Language, 2003; http://sbml.org/index.psp.
8. The Systems Biology Workbench, 2004; http://sbw.sourceforge.net.
9. CellML.org, 2001; www.cellml.org/public/about/what_is_cellml.html.
10. The UCL Beacon Project, 2003; http://grid.ucl.ac.uk/biobeacon/php/index.php.
11. C. Rosse and J. Mejino, "A Reference Ontology for Bioinformatics: The Foundational Model of Anatomy," *J. Biomedical Informatics*, vol. 36, 2003, pp. 478-500.
12. Gene Ontology Consortium, *Gene Ontology*, 2004; www.geneontology.org.

Anthony Finkelstein *is a professor of software systems engineering in the Department of Computer Science and the Centre for Mathematics and Physics in the Life Sciences and Experimental Biology (CoMPLEX), University College London. His research interests include systems modeling. He received a PhD in engineering design from the Royal College of Art. Contact him at a.finkelstein@ucl.ac.uk.*

James Hetherington *is a research fellow for CoMPLEX, University College London. His research interests include mathematical modeling of biological systems. Hetherington received a PhD in physics from Cambridge. Contact him at j.hetherington@ucl.ac.uk.*

Linzhong Li *is a research fellow for CoMPLEX, University College London. His research interests include the mathematical modeling of biological systems. Li received a PhD in applied mathematics from University College London. Contact him at l.li@ucl.ac.uk.*

Ofer Margoninski *is a research fellow for CoMPLEX, University College London. His research interests include systems modeling. Margoninski received an MSc in computer science from Hebrew University. Contact him at o.margoninski@ucl.ac.uk.*

Peter Saffrey *is a research fellow for CoMPLEX, University College London. His research interests include systems modeling. Saffrey received a PhD in computer science from the University of Glasgow. Contact him at p.saffrey@ucl.ac.uk.*

Rob Seymour *is a professor of mathematics, Department of Mathematics, at CoMPLEX, University College London. His research interests include biomathematics. Seymour received a PhD in mathematics from Warwick University. Contact him at r.seymour@ucl.ac.uk.*

Anne Warner *is a professor of developmental biology, Department of Anatomy and Developmental Biology, and director of CoMPLEX, University College London. Warner's research interests include integrative biology. She is a Fellow of the Royal Society. Contact her at a.warner@ucl.ac.uk.*