Multiscale Modeling: Physiome Project Standards, Tools, and Databases

Peter J. Hunter, Bioengineering Institute, University of Auckland
Wilfred W. Li, San Diego Supercomputer Center, University of California, San Diego
Andrew D. McCulloch, University of California, San Diego
Denis Noble, Oxford University

The Physiome project's markup languages and associated tools leverage the CellML and FieldML model databases published in peer-reviewed journals. As these tools mature, researchers can check models for conformance to underlying physics laws, using them to develop complex physiological models from separately validated components.

Biological complexity can only be understood quantitatively by using mathematics, the language of quantitative science.[1] Even the founder of experimental physiology, Claude Bernard, wrote as long ago as 1865 "the application of mathematics to natural phenomena is the aim of all science."[2] While championing the experimental method in physiology, Bernard clearly foresaw a time when it would become mathematical.

A characteristic of biological complexity, however, is the intimate connection that exists between different length scales—from the nanometer-length scale of molecules to the highly structured meter scale of the whole human body.[3] Subtle changes in molecular structure as a consequence of a single gene mutation can lead to catastrophic failure at the organ level, such as heart failure from reentrant arrhythmias that lead to ventricular fibrillation.[4] But information flows equally in the reverse direction: mechanoreceptors at the cell level sense the mechanical load on the musculoskeletal system and influence gene expression via signal transduction pathways.

Thus, interpreting the interactions that occur across the length scales from genes and proteins to cells, tissues, organs, and organ systems requires a multiscale mathematical modeling framework. The range of relevant time scales is even more spectacular—from the nanoseconds to microseconds of molecular events to the 10^{15}-second aging process that occurs during a 70-year human lifetime.

MATHEMATICAL MODELS

Mathematical models of a biological process, such as a signal transduction pathway or a set of transmembrane ion channels, provide a means of describing the behavior of that process and understanding how components such as proteins and small molecules relate to one another and to the behavior of the whole system quantitatively.

The cardiac action potential offers a good example of this. Cardiac myocytes contain about 30 membrane proteins that regulate the flow of Na+, K+, Ca^{2+} and Cl– ions across the cell membrane to generate the voltage change known as the "action potential" that propagates from one cell to its neighbor to synchronize contraction of the heart muscle. This rapid voltage change initiates Ca^{2+} entry into the cell, which stimulates further release of Ca^{2+} from internal stores to switch on the cell's contractile apparatus.

Researchers can study the properties of each membrane ion channel, ion exchangers, and pumps patch-clamping a small portion of membrane to measure the current flow of ions under controlled voltage changes.

Published by the IEEE Computer Society
0018-9162/06/$20.00 © 2006 IEEE

The channels are not passive conduits but rather have time-dependent ion-gating properties that are themselves voltage dependent. By measuring and mathematically characterizing the precise time- and voltage-dependence of each channel, researchers can build up composite models that contain all the membrane proteins involved in generating the action potential.

Indeed, our developing understanding of the cardiac action potential shows a close and fruitful interaction between experimental measurement and modeling. In some cases, the experiments showed the existence of new proteins; in other cases, the models predicted the need for new channels or properties of known channels that were then experimentally verified.

The power of models derives in part from their ability to introduce a priori physical constraints, such as the conservation of mass or charge, that must be satisfied. This power also derives partly from models' ability to describe biological processes that are otherwise too complex to understand and partly from their ability to provide quantitative predictions of behavior that researchers can subject to experimental testing.

As XML-based standards, CellML and FieldML models can be combined using many different techniques.

Researchers have developed three standards to describe biological processes: CellML, the Systems Biology Markup Language,[5] and FieldML.

CELLML

Typically, when a model is published in a scientific journal as a set of equations, any reader who wishes to use the information must write a simulation package and transcribe the equations from the printed text into computer code. This can be a difficult and time-consuming process because the published equations usually contain errors. Further, many biologists lack the ability or inclination to write the computer code needed to implement models, but they wish to use published models in the interpretation of their experimental data.

To remedy this situation, the Bioengineering Institute at the University of Auckland developed an XML markup language standard, CellML (www.cellml.org), to encapsulate mathematical models in an electronic form that a simulation package can read.[6] The mathematical content itself is marked up in Content MathML, the W3C standard for representing mathematics in a parsible form. Additional data about the model, such as the journal publication, authors, and other metadata, resides with the mathematics in the CellML file.

Further, many complex processes can only be understood by combining separately developed models of the process's different aspects. For example, a model of cardiac myocyte activity in electromechanical coupling requires, at the very least, the combination of ion channel models, a calcium transport model, and a myofilament force production model. Using this model to study myocardial ischemia requires building additional models for regulation of cellular pH and glycolysis and oxidative metabolism.

Further, many of the proteins in these models use phosphorylation regulation by protein kinases and dephosphorylation by phosphatases—processes themselves regulated by signal transduction pathways such as the adrenergic, calcium/calmodulin/CaM-kinase, and IP3/Ca/calcineurin/NFAT pathways. A different research group developed each of these models. Combining or importing such separately created models into a composite electromechanical model requires uniquely identifying the terms each model uses.

To do this, researchers use ontologies, which combine structured vocabularies with a small set of relationships between their components. Well-known examples of biological ontologies include the Gene Ontology project (www.geneontology.org) and BioPAX (www.biopax.org). The CellML metadata uses such ontologies to provide unique biological identification for each component in a CellML model.

A repository of CellML models (www.cellml.org/models) has been established that covers the following types: signal transduction and metabolic pathways, cardiac electrophysiology, calcium dynamics, immunology, cell cycle, simplified electrophysiology, other cell-type electrophysiology, and smooth and skeletal-muscle models.

Researchers have also written an application programming interface for the CellML 1.0 standard and provided a set of open source tools for authoring, validating, displaying, and running CellML-encoded models (www.cellml.org/tools). These tools make it straightforward to incorporate complex subcellular functions via CellML models into the solution of equations governing factors such as mechanical contraction, electrical activation, and blood flow at the tissue level.

FIELDML

As XML-based standards, CellML and FieldML models can be combined using many different techniques. For example, researchers can use XML namespaces (www.w3.org/TR/REC-xml-names/) to define models in a single serial file form. They also can use a linking standard such as Xlink[6] (www.w3.org/TR/xlink/) to store models in different files for easy reference.

The following example shows why storing models in separate files is most appropriate for multiscale physiome models. Biological processes use a 3D structure to help define function at all spatial scales. Protein struc-

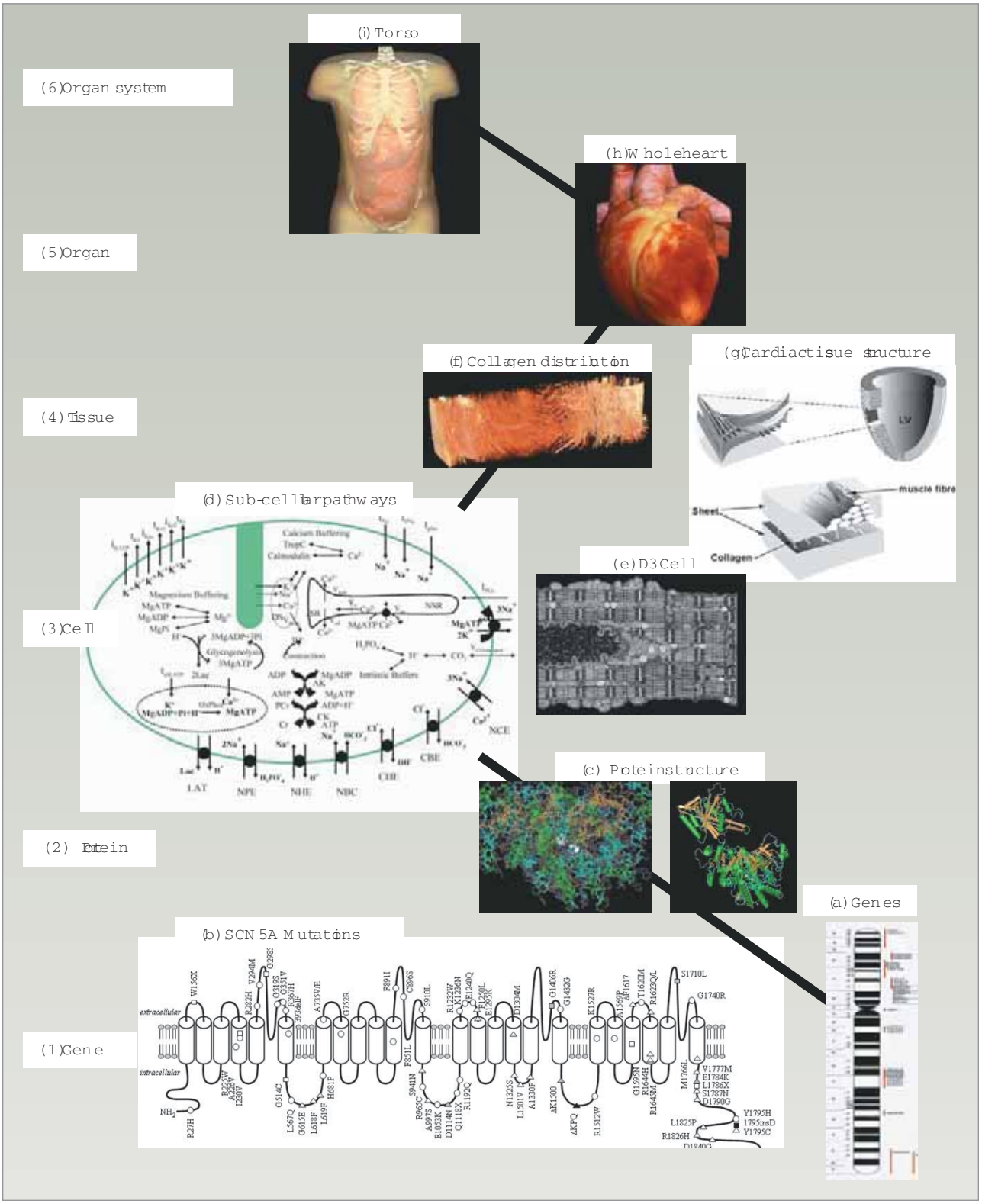

Figure 1. Hierarchy of spatial scales used in the Heart Physiome Project. Image (b) provided by Charles Antzelovitch; all other Figure 1 images courtesy of the Auckland Bioengineering Institute.

ture, for example, is critical to protein-protein and protein-ligand binding. Cell processes are usually highly localized spatially within a cell, and tissue structure is critical for function in connective, muscle, nerve, and epithelial tissues. Similarly, structure, or anatomy, closely links to function, or physiology, in all organs and organ

systems. This means that mathematical models of biological processes must capture the 3D structure at an appropriate spatial scale and that researchers need a markup language such as FieldML to standardize this mathematical parameterization of structure.

Defining an API for the FieldML standard allows creating software packages that solve partial differential equations over spatial domains to read and write the files, which facilitates the reuse of structural models. A software package for creating and visualizing any FieldML model is freely available under an open source license (www.cmiss.org/cmgui). Further, an API for FieldML is currently under development (www.physiome.org.nz/xml_languages/fieldml).

The CellML and FieldML files must be kept separate because the CellML models of ion channels, such as those based on Hodgkin-Huxley kinetics, Markov state models, lumped parameter models of metabolic pathways, or signal transduction pathways, can be reused many times in different cell and tissue types. These have either quite different anatomical descriptions or the same tissue type with spatially varying density of a given parameter, such as ion channels.

CellML and FieldML provide a mechanism for encapsulating the mathematics and metadata associated with a biological process model.

Nature reuses the same proteins or signaling systems in different cell and tissue types. Using FieldML parameterizations, together with the separately defined CellML parameterization of protein function, to capture the anatomy of tissue or a whole organ and the spatial distribution of protein density within it is fundamental to multiscale organ modeling.

A heart modeling example assigns one FieldML file to each anatomic field and solves the physics over these regions by calling on the CellML definitions of ion channel function at many separate points—typically 10 to 100 million—but all contained in one CellML file. For these problems, combining the two information types would be inefficient.

MODEL VALIDATION

Researchers have developed CellML and FieldML to provide a convenient and robust mechanism for encapsulating the mathematics and metadata associated with a biological process model.

The CellML model repository contains models from peer-reviewed journal publications. Making a CellML version of a published model available in the repository does not guarantee the model will be error free because the original publication might have had errors and, for issues of provenance, the corresponding CellML version needs to faithfully reproduce the mathematics in the paper, errors included. However, it is clearly desirable to then create version 2 of the CellML-encoded model in which various checks have been carried out that include determining that

- units are consistent,
- all parameters and initial conditions are defined, and
- running the model reproduces the results published in the paper.

As the CellML standard and the model repository gain journal acceptance, it should be possible to work with authors to achieve a level 2 version of the model at the time of publication. This has already happened for the publication of a metabolic model.[7] A further level of curation is anticipated that checks models for the extent to which they satisfy physical constraints such as conservation of mass, momentum, and charge.

MODELING THE HEART

Three research groups have developed a multiscale heart model, based on the CellML and FieldML standards: the Bioengineering Institute at the University of Auckland, the Laboratory of Physiology at Oxford University, and the cardiac mechanics group at the University of California in San Diego. This model, shown in Figure 1, links subcellular functions, such as ion channel currents, to structures and functions at the cell, tissue, and whole-organ levels.

At the cell level, the model includes the ion channels, pumps, and exchangers responsible for the cardiac action potential, the myofilament proteins responsible for calcium transport and active force generation, the membrane pumps and exchangers responsible for maintaining cellular pH, the glycolysis and oxidative metabolic pathways, and several of the signal transduction pathways that regulate these cellular subsystems and their responses to external stimuli, such as the adrenergic receptor signaling pathway.

At the tissue level, the heart model shown in Figure 2 currently includes the coupled equations governing cardiac electrophysiology,[8] large-deformation muscle mechanics,[9] metabolic processes and ventricular and coronary fluid mechanics, and the constitutive relations representing the anisotropic tissue-scale electrical and mechanical properties and the physicochemical interactions between the cells in three dimensions.

Researchers have used the model to understand the role of tissue microstructure in supporting the propagation of the activation wave that initiates contraction in the heart and the role of tissue microstructure in defibrillation of cardiac tissue. Developers are currently reengineering the CMISS software used for the heart model computations to leverage distributed-memory parallel processing in an open source project.

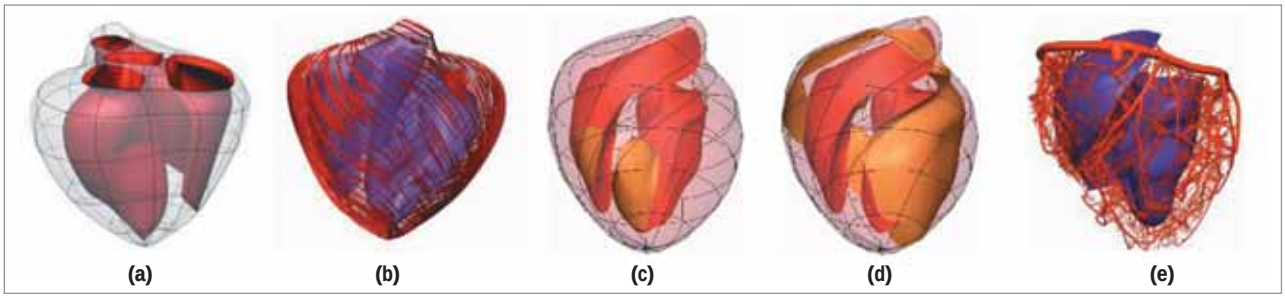

Figure 2. Organ-level heart model. (a) Geometry of the ventricular myocardium shown in a 3D finite-element model of the heart; the inner surfaces of the left and right ventricles appear in red through the heart model's translucent epicardium. (b) Fiber orientations on the epicardial surface of the heart. (c) Orange wave of electrical depolarization shown at an early stage of myocardial activation and (d) at a later stage. (e) Coronary arteries modeled from pig heart data. The physiome heart model couples these various physical processes and links the tissue-level properties to subcellular processes such as ion-channel current flow and mechanical force generation.

Several additional aspects of heart structure and function must be developed to provide a comprehensive model for understanding cardiac diseases. At the tissue and organ level, these include models of the heart valves and chordae tendinae, the atria, the pericardium, and the neural control system. At the cell level, much work remains on signal transduction pathways, especially the compartmentalization of signaling processes and the regulation of gene expression,[10] not to mention protein synthesis and posttranslational modification processes.

ENABLING TEAM SCIENCE

The Physiome Project,[3,11] which aims to provide the framework for multiscale anatomically- and biophysically-based human body modeling, is enormously complex and computationally demanding. In addition to driving the development of markup language standards and model repositories, it also drives innovations in scientific workflow management, information integration and query, cluster and grid computing, and environments for international collaboration.

The *computational grid* refers to the hardware and software infrastructure that assists researchers in solving complex problems using many computers, including the sharing of distributed resources across institutions or organizations.[12] The National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure[13] argued that making software applications easier to develop and share and reducing the amount of effort scientists must devote to developing information technologies could increase the efficiency, reliability, and effective sharing of new applications.

The grid's architecture has been described as an hourglass model, in that many varied applications and compute resources are connected by a core set of middleware components—such as connectivity protocols that support secure sharing of resources—and collective services that support the monitoring, scheduling, and accounting of different resources.[14]

The grid's emergence involves a developing story of standardization, reusability, service orientation, and interoperability. The Open Grid Services Architecture (OGSA) provides the community architectural standard for implementation of the grid as outlined by the Global Grid Forum (www.ggf.org). OGSA defines a grid service as a Web service with life cycle and state management in the grid environment.[14] A Web service is a platform- independent software system that performs operations described in the Web Service Definition Language (WSDL). Web services typically exchange data in an XML format using the HTTP/SOAP protocol. Web services currently provide the dominant technology for distributed computing as defined by the World Wide Web Consortium and developers can use them to build loosely coupled complex interactions and workflows (www.w3c.org).

Several international initiatives, such as the NSF Advanced Cyberinfrastructure Program (www.nsf.gov/news/special_reports/cyber/agrand.jsp) or the UK e-science program (www.rcuk.ac.uk/escience/), aim to advance the integration of the grid into science, engineering, and medicine.

The TeraGrid (www.teragrid.org), a major NSF initiative, involves eight institutions across the US, with 40 teraflops of computing power, 2 petabytes of storage, and 10 to 30 gigabits of network connection. The UK e-science program funds key projects such as e-science centers and the Open Middle Infrastructure Institute. The Pacific Rim Grid Application and Middleware Assembly (www.pragma-grid.net) is an NSF-funded project for international grid technology exchange and scientific collaboration. It offers a successful example of how grid computing helps build human networks with mutual trust and shared vision.[15]

To increase the usability and decrease the cost of entry to the grid, researchers are developing new programming models or application execution environments atop the grid software infrastructure. These are some-

times referred to as grid application-level tools[16] or upper middleware.[17]

For example, as part of its developmental effort to enable biomedical applications through grid computing, the National Biomedical Computation Resource (www.nbcr.net) has developed a generic Web service wrapper, *Opal*, for deploying legacy applications as Web services, accessible anywhere on the grid. Pursuant with its goal of developing a collaborative environment for biomedical research and clinical information management, the Biomedical Informatics Research Network (www.nbirn.net) sponsors the continued development of GridSphere, a grid portal environment for application deployment.

A key goal of this work is achieving interoperability by defining common interfaces and information representations like CellML and FieldML that support efficient integration of existing software tools and models. The UK MyGrid project has built grid tools for bioinformatics (www.mygrid.org.uk) that use SoapLab for wrapping bioinformatics applications as Web services and Taverna as a Web service workflow management system. NBCR's tool, *Vision*, is a Python-based visual programming environment with support for Web-service-based workflow execution. These upper middleware packages keep grid use as transparent as possible, while the adoption of Web service and XML-based messaging makes interoperability practical.

The standards that CellML and FieldML define for the Physiome Project help ensure software-component interoperability and knowledge integration at the semantic level. Although high-performance requirements can influence an application's internal design, the public interface—where interactions with other public services are desirable—could benefit from the service-oriented computing paradigm. The Physiome Project will combine them to expose a set of methods that supports common queries to the CellML database as a Web service that can use Grid Security Infrastructure authentication to verify that a user can access the data. Using the same credentials that the user supplies, additional operations can be carried out such as model computations using distributed high-performance computing resources.

In the Web services model, data must be self-describing and represented in XML, or metadata represented in XML, with associated operations or even software components provided otherwise. The built-in models and mathematical operations—software codes necessary in CellML models—will drive development of more advanced operations and complex interactions, all carried out transparently, either without the user's knowledge or, at most, with minimal input.

The Physiome Project has placed particular emphasis on developing standards for data exchange interoperability. The Integrative Biology project, an international consortium, is developing a suitable collection of middleware and customized solutions for combating fatal heart diseases and cancer (www.integrativebiology.ox.ac.uk/workprog.html).

NBCR, BIRN, and MyGrid use different scientific drivers to develop tools and middleware built upon grid computing resources. Each constructs an important component of the multiscale biological modeling space. As researchers develop a new, reusable, generalized technology and it becomes part of the Grid, more scientists and students will benefit from it. This feedback loop—technology enabling science and science driving technology—will transform how researchers do science and build an infrastructure.

At present, the Physiome Project consists of the markup languages and associated tools for authoring, validating, displaying, and executing models, together with the CellML and FieldML model databases that have been published in peer-reviewed journals. As the tools mature and the databases grow, developers can contemplate a further level of model curation. Models can be checked for conformance to underlying laws of physics such as mass and charge conservation.

Another important development will be building complex physiological models from separately validated model components[18] and linking CellML models to experimental data encoded in markup-language data standards.[19] The ability to link components will depend on the development of biological ontologies, which describe structured vocabularies and relationships between components.

The heart is just one of the human body's 12 organ systems, and similar progress has been made on several other systems, including the lungs and the musculoskeletal and digestive systems. All this is only a small beginning in using anatomically and biophysically based models to quantify the body's complex and highly integrated physiology. ■

References

1. D. Noble "The Rise of Computational Biology, *Nature Reviews Molecular Cell Biology*, vol. 3, 2002, pp. 460-463.
2. C. Bernard, *An Introduction to the Study of Experimental Medicine*, Claude Bernard, H.C. Greene (translator), Dover Publications, 1957.
3. P.J. Hunter and T.K. Borg, "Integration from Proteins to Organs: The Physiome Project," *Nature Reviews Molecular and Cell Biology*, vol. 4, 2003, pp. 237-243.
4. D. Noble, "Modeling the Heart: From Genes to Cells to the Whole Organ, *Science*, vol. 295, 2002, pp. 1678-1682.
5. M. Hucka et al., "The Systems Biology Markup Language (SBML): A Medium for Representation and Exchange of Biochemical Network Models," *Bioinformatics*, vol. 19, 2003, pp. 524-531.

6. AA. Cuellar et al., "An Overview of CellML 1.1, a Biological Model Description Language," *Trans. Society for Modeling and Simulation Int'l*, vol. 79, no. 12, 2003, pp. 740-747.
7. D.A. Beard "A Biophysical Model of the Mitochondrial Respiratory System and Oxidative Phosphorylation," *PLoS Computational Biology*, vol. 1, no. 4, 2005, **|Au: Inclusive page numbers?|**.
8. P.J. Hunter A.J. Pullan, and B.H. Smaill, "Modeling Total Heart Function," *Ann. Review Biomedical Eng.*, vol. 5, 2003, pp. 147-177.
9. D.P. Nickerson, N.P. Smith, and P.J. Hunter, "A Model of Cardiac Cellular Electromechanics," *Philosophical Trans. Royal Soc. of London*, A359, 2001, pp. 1159-1172.
10. R.L. Winslow and M.S. Boguski, "Genome Informatics—Current Status and Future Prospects, *Circulation Research*, vol. 92, 2003, pp. 953-961.
11. E.J. Crampin et al., "Computational Physiology and the Physiome Project," *J. Experimental Physiology*, vol. 89, no. 1, 2004, pp. 1-26.
12. S.L. Graham, M. Snir, and C.A. Patterson, *Getting Up to Speed: The Future of Supercomputing*, National Academies Press, 2004, p. 308.
13. D.E. Atkins et al., "Revolutionizing Science and Engineering through Cyberinfrastructure," *Nat'l Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, 2003; www.nsf.gov/od/oci/reports/APXA.pdf.
14. I. Foster and C. Kesselman, eds., *The Grid 2: Blueprint for a New Computing Infrastructure*. 2nd ed., Morgan Kaufmann, 2004.
15. C. Zheng et al., "The PRAGMA Testbed: Building a Multi-Application International Grid," *CCGrid*, **//publisher??//,** 2006, **|Au: Inclusive page numbers?|**.
16. H. Bal et al., "Application-Level Tools," *The Grid 2*, 2nd ed., I. Foster and C. Kesselman, eds., Morgan Kaufmann, 2004, **|Au: Inclusive page numbers?|**
17. D. Abramson et al., "Deploying Scientific Applications to the PRAGMA Grid Testbed: Lessons Learned," *Deploying Scientific Applications to the PRAGMA Grid Testbed: Strategies and Lessons*, **|Au: Publisher?|** 2006.
18. C.M. Lloyd, M.D.B. Halstead, and P.M.F. Nielsen, "CellML: Its Future, Present and Past," *Progress in Biophysics and Molecular Biology*, vol. 85, nos. 2-3, 2004, pp. 433-450.
19. FC Dewey et al., "ExperiBase—An Object Model Implementation for Biology?" *Proc. 5th Intelligent Systems in Medicine and Biology*, **|Au: Proceedings publisher?|**, 2005, **|Au: Inclusive page numbers?|**.

Peter J. Hunter is the director of the Bioengineering Institute, University of Auckland, New Zealand. His research interests include soft tissue biomechanics, electrophysiology, cell biophysics, and finite element techniques. Hunter received a PhD in physiology from Oxford University. Contact him at p.hunter@auckland.ac.nz.

Wilfred Li, is executive director of the National Biomedical Computation Resource at the San Diego Supercomputer Center, University of California, San Diego. His research interests include providing transparent access to the new and emerging grid infrastructure to deliver integrated compute, data, physical, experimental, and human resources to biomedical scientists investigating a wide range of medically important problems. Li received a PhD in biochemistry and molecular biology from the University of California, Los Angeles. Contact him at wilfred@sdsc.edu.

Andrew D. McCulloch, is a professor of bioengineering at the University of California, San Diego. His research focuses on building experimental and computational models to investigate the relationships between the cellular and extracellular structure of cardiac muscle and the electrical and mechanical function of the heart during ventricular remodeling and cardiac arrhythmia. McCulloch received a PhD in engineering science and physiology from the University of Auckland. Contact him at amcculloch@ucsd.edu.

Denis Noble is the codirector of computational biology at the University Laboratory of Physiology, Oxford. His research interests focus on using computer models of biological organs and systems to interpret function from the molecular through whole-body levels. Noble received a PhD in physiology from the University College, London. Contact him at denis.noble@physiol.ox.ac.uk.