

'In Silico' *Simulation of Biological Processes: Novartis Foundation Symposium, Volume 247*
Edited by Gregory Bock and Jamie A. Goode
Copyright © Novartis Foundation 2002.
ISBN: 0-470-84480-9

'IN SILICO'
SIMULATION OF
BIOLOGICAL
PROCESSES

The Novartis Foundation is an international scientific and educational charity (UK Registered Charity No. 313574). Known until September 1997 as the Ciba Foundation, it was established in 1947 by the CIBA company of Basle, which merged with Sandoz in 1996, to form Novartis. The Foundation operates independently in London under English trust law. It was formally opened on 22 June 1949.

The Foundation promotes the study and general knowledge of science and in particular encourages international co-operation in scientific research. To this end, it organizes internationally acclaimed meetings (typically eight symposia and allied open meetings and 15–20 discussion meetings each year) and publishes eight books per year featuring the presented papers and discussions from the symposia. Although primarily an operational rather than a grant-making foundation, it awards bursaries to young scientists to attend the symposia and afterwards work with one of the other participants.

The Foundation's headquarters at 41 Portland Place, London W1B 1BN, provide library facilities, open to graduates in science and allied disciplines. Media relations are fostered by regular press conferences and by articles prepared by the Foundation's Science Writer in Residence. The Foundation offers accommodation and meeting facilities to visiting scientists and their societies.

Information on all Foundation activities can be found at
<http://www.novartisfound.org.uk>

Novartis Foundation Symposium 247

'IN SILICO'
SIMULATION OF
BIOLOGICAL
PROCESSES

2002



JOHN WILEY & SONS, LTD

Copyright © Novartis Foundation 2002
Published in 2002 by John Wiley & Sons Ltd,
The Atrium, Southern Gate,
Chichester, West Sussex PO19 8SQ, UK

National 01243 779777
International (+44) 1243 779777
e-mail (for orders and customer service enquiries): cs-books@wiley.co.uk
Visit our Home Page on <http://www.wileyurope.com>
or <http://www.wiley.com>

All Rights Reserved. No part of this book may be reproduced, stored in a retrieval system or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning or otherwise, except under the terms of the Copyright, Designs and Patents Act 1988 or under the terms of a licence issued by the Copyright Licensing Agency Ltd, 90 Tottenham Court Road, London W1T 4LP, UK, without the permission in writing of the Publisher. Requests to the Publisher should be addressed to the Permissions Department, John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England, or emailed to permreq@wiley.co.uk, or faxed to (+44) 1243 770620.

This publication is designed to provide accurate and authoritative information in regard to the subject matter covered. It is sold on the understanding that the Publisher is not engaged in rendering professional services. If professional advice or other expert assistance is required, the services of a competent professional should be sought.

Other Wiley Editorial Offices

John Wiley & Sons Inc., 111 River Street, Hoboken, NJ 07030, USA

Jossey-Bass, 989 Market Street, San Francisco, CA 94103-1741, USA

Wiley-VCH Verlag GmbH, Boschstr. 12, D-69469 Weinheim, Germany

John Wiley & Sons Australia Ltd, 33 Park Road, Milton, Queensland 4064, Australia

John Wiley & Sons (Asia) Pte Ltd, 2 Clementi Loop #02-01, Jin Xing Distripark, Singapore 129809

John Wiley & Sons Canada Ltd, 22 Worcester Road, Etobicoke, Ontario, Canada M9W 1L1

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Novartis Foundation Symposium 247
viii+262 pages, 39 figures, 5 tables

Library of Congress Cataloging-in-Publication Data

'In silico' simulation of biological processes / [editors, Gregory Bock and Jamie A. Goode.
p. cm. — (Novartis Foundation symposium ; 247)

“Symposium on ‘In silico’ simulation of biological processes, held at the Novartis Foundation, London, 27–29 November 2001” — Contents p.

Includes bibliographical references and index.

ISBN 0-470-84480-9 (alk. paper)

1. Biology—Computer simulation—Congresses. 2. Bioinformatics—Congresses. I.

Bock, Gregory. II. Goode, Jamie. III. Symposium on ‘In Silico’ Simulation of Biological Processes (2001 : London, England) IV. Series.

QH324.2 .I5 2003

570.1'13—dc21

2002035730

British Library Cataloguing in Publication Data

A catalogue record for this book is available from the British Library

ISBN 0 470 84480 9

Typeset in 10¹/₂ on 12¹/₂ pt Garamond by Dobbie Typesetting Limited, Tavistock, Devon.

Printed and bound in Great Britain by Biddles Ltd, Guildford and King's Lynn.

This book is printed on acid-free paper responsibly manufactured from sustainable forestry, in which at least two trees are planted for each one used for paper production.

Contents

Symposium on 'In silico' simulation of biological processes, held at the Novartis Foundation, London, 27–29 November 2001

Editors: Gregory Bock (Organizer) and Jamie A. Goode

This symposium is based on a proposal made by Dr Paul Herrling

- Denis Noble** Chair's introduction 1
- Andrew D. McCulloch and Gary Huber** Integrative biological modelling
in silico 4
Discussion 20
- Mike Giles** Advances in computing, and their impact on scientific computing 26
Discussion 34
- David Krakauer** From physics to phenomenology. Levels of description and levels of selection 42
- Philip K. Maini** Making sense of complex phenomena in biology 53
Discussion 60
- Michael Ashburner and Suzanna Lewis** On ontologies for biologists: the Gene Ontology—untangling the web 66
Discussion 80
- General discussion I** Model validation 84
- Minoru Kanehisa** The KEGG database 91
Discussion 101
- Shankar Subramaniam** and the Bioinformatics Core Laboratory Bioinformatics of cellular signalling 104
Discussion 116

General discussion II	Standards of communication	119
	Semantics and intercommunicability	121
Raimond L. Winslow, Patrick Helm, William Baumgartner Jr., Srinivas Peddi, Tilak Ratnanather, Elliot McVeigh and Michael I. Miller	Imaging-based integrative models of the heart: closing the loop between experiment and simulation	129
	<i>Discussion</i>	141
General discussion III	Modelling Ca^{2+} signalling	144
Leslie M. Loew	The Virtual Cell project	151
	<i>Discussion</i>	160
Thomas Simon Shimizu and Dennis Bray	Modelling the bacterial chemotaxis receptor complex	162
	<i>Discussion</i>	194
Denis Noble	The heart cell <i>in silico</i> : successes, failures and prospects	182
	<i>Discussion</i>	194
General discussion IV		198
P. J. Hunter, P. M. F. Nielsen and D. Bullivant	The IUPS Physiome Project	207
	<i>Discussion</i>	217
Jeremy M. Levin, R. Christian Penland, Andrew T. Stamps and Carolyn R. Cho	Using <i>in silico</i> biology to facilitate drug development	222
	<i>Discussion</i>	238
Final discussion	Is there a theoretical biology?	244
	Index of contributors	253
	Subject index	255

Participants

Michael Ashburner EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD and Department of Genetics, University of Cambridge, Cambridge CB2 3EH, UK

Michael Berridge The Babraham Institute, Laboratory of Molecular Signalling, Babraham Hall, Babraham, Cambridge CB2 4AT, UK

Jean-Pierre Boissel Service de Pharmacologie Clinique, Faculté RTH Laennec, rue Guillaume Paradin, BP 8071, F-69376 Lyon Cedex 08, France

Marvin Cassman NIGMS, NIH, 45 Center Drive, Bethesda, MD 20892, USA

Edmund Crampin University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, UK

Mike Giles Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK

Jutta Heim Novartis Pharma AG, CH-4002 Basel, Switzerland

Rob Hinch OCIAM, Mathematical Institute, 24–29 St Giles', Oxford OX1 3LB, UK

Peter Hunter Department of Engineering Science, University of Auckland, Private Bag 92019, Auckland, New Zealand

Minoru Kanehisa Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan

Jeremy Levin Physiome Sciences, Inc., 150 College Road West, Princeton, NJ 08540-6604, USA

Leslie M. Loew Center for Biomedical Imaging Technology, Department of Physiology, University of Connecticut Health Center, Farmington, CT 06030-3505, USA

Philip Maini Centre for Mathematical Biology, Mathematical Institute, 24–29 St Giles', Oxford OX1 3LB, UK

Andrew D. McCulloch Department of Bioengineering, Whitaker Institute of Biomedical Engineering and San Diego Supercomputer Center, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412, USA

David Nickerson (*Novartis Foundation Bursar*) Bioengineering Research Group, Level 6–70 Symonds Street, Department of Engineering Science, University of Auckland, Auckland, New Zealand

Denis Noble (*Chair*) University Laboratory of Physiology, University of Oxford, Parks Road, Oxford OX1 3PT, UK

Thomas Paterson Entelos, Inc., 4040 Campbell Ave, Suite #200, Menlo Park, CA 94025, USA

Mischa Reinhardt Novartis Pharma AG, Lichtstrasse 35, WSJ-88.10.10, CH-4002, Basel, Switzerland

Tom Shimizu Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

Shankar Subramaniam Departments of Chemistry & Biochemistry and Bioengineering, San Diego Supercomputing Center, Dept. 0505, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92037, USA

Raimond Winslow The Whitaker Biomedical Engineering Institute, The Johns Hopkins University, Center for Computational Medicine & Biology, Rm 201B Clark Hall, 3400 N. Charles Street, Baltimore, MD 21218, USA

Chair's introduction

Denis Noble

University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, UK

This meeting establishes a major landmark since it is the first fully published meeting on the growing field of computer (*in silico*) representation of biological processes. The first International Conference on Computational Biology was held earlier in 2001 (Carson et al 2001) but was not published. Various funding bodies (INSERM, MRC and NIH) have held strategy meetings, also unpublished. And there is a lot of interest in the industrial world of pharmaceutical, biotechnology and medical device companies. Now is the ripe time to explore the issues in depth. That is the purpose of this meeting.

The Novartis Foundation has already played a seminal role in the thinking that forms the background to our discussions. Two previous meetings were fertile breeding grounds for the present one. The first was on *The limits of reductionism in Biology* (Novartis Foundation 1998), proposed and chaired by Lewis Wolpert. That meeting set the scene for one of the debates that will feature again in this meeting, which is the issue of reduction versus integration. There cannot be any doubt that most of the major successes in biological research in the last few decades have come from the reductionist agenda—attempting to understand biological processes entirely in terms of the smallest entities, i.e. genes, proteins and other macromolecules, etc. We have, successfully, broken Humpty Dumpty down into his smallest bits. Do we now have to worry about how to put him back together again? That is the agenda of integration, and most of the people I have spoken to believe that this absolutely requires simulation in order to succeed. I also suggest that there needs to be a constructive tension between reduction and integration. Neither alone gives the complete story.

The reason is that in order to unravel the complexity of biological processes we need to model in an integrative way at all levels: gene, protein, pathways, sub-cellular, cellular, tissue, organ, system. This was the issue debated in the symposium on *Complexity in biological information processing* (Novartis Foundation 2001), chaired by Terry Sejnowski. An important discussion in that meeting focused on the question of whether modelling should be tackled from the bottom-up (starting with genes and biomolecules) or top-down (starting with physiological and pathological states and functions). A conclusion of that

discussion, first proposed by Sydney Brenner, was that modelling had to be ‘middle-out’, meaning that we must begin at whatever level at which we have most information and understanding, and then reach up and down towards the other levels.

These issues will feature again, sometimes in new guise, in the present meeting. But there will also be some new issues to discuss. What, for example, is computational biology? How does it differ from and relate to mathematical biology? Could we view the difference as that between being descriptive and being analytical?

Then, what are the criteria for good modelling? I would suggest that biological models need to span at least three levels. Level 1 would be primarily descriptive. It will be the level at which we insert as much data as possible. At this data-rich level, we don’t worry about how many parameters are needed to describe an elephant! The elephant is a given, and the more details and data the better. Far from making it possible to build anything given enough parameters, at this level data will be restrictive. It will set the boundaries of what is possible. Biological molecules are as much the prisoners of the system as they are its determinants.

Level 2 will be integrative—how do all these elements interact? This is the level at which we need to do the heaviest calculations, literally to ‘integrate’ the data into a working model.

Level 3 is the level (or better still, multiple levels) at which we can be explanatory and predictive; to gain physiological insight.

Another issue we will tackle concerns the role of biological models. Models do not serve a single purpose. Here is a preliminary list that I propose:

- (1) To systematize information and interactions
- (2) For use in computational experiments
- (3) For analysis of emergent properties
- (4) To generate counter-intuitive results
- (5) To inspire mathematical analysis
- (6) . . . but ultimately to fail

The last is important and is poorly understood in biological work. All models must fail at some point since they are always only partial representations. It is *how* models fail that advances our understanding. I will illustrate this principle in my own paper at this meeting (Noble 2002a, this volume).

So, the questions to be debated at this meeting will include:

- What does *in silico* refer to and include?
- What are the roles of modelling in biology?
- What is the role of mathematics in modelling?

- What is the relation of modelling to bioinformatics?
- What about model validation?
- What are the hardware and software constraints and opportunities?
- What are the applications to health and disease?
- What are the industrial applications?
- Could we eventually be so successful that we can move towards a virtual organism/human?
- Even more ambitiously, can we envisage the development of a theoretical biology?

My own tentative answer to the last question is that if there is to be a theoretical biology, it will have to emerge from the integration of many pieces of the reconstruction of living systems (see Noble 2002b). We will, appropriately, keep this big issue for the concluding discussion.

I look forward to a lively debate, touching on everything from the immensely practical to the audaciously theoretical.

References

- Carson JH, Cowan A, Loew LM 2001 Computational cell biologists snowed in at Cranwell. *Trends Cell Biol* 11:236–238
- Noble D 2002a The heart *in silico*: successes, failures and prospects. In: '*In silico*' simulation of biological processes. Wiley, Chichester (Novartis Found Symp 247) p 182–197
- Noble D 2002b Biological Computation. In: Encyclopedia of life sciences, <http://www.els.net>. Nature Publishing Group, London
- Novartis Foundation 1998 The limits of reductionism in biology. Wiley, Chichester (Novartis Found Symp 213)
- Novartis Foundation 2001 Complexity in biological information processing. Wiley, Chichester (Novartis Found Symp 239)

Integrative biological modelling *in silico*

Andrew D. McCulloch and Gary Huber

Department of Bioengineering, The Whitaker Institute of Biomedical Engineering, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093-0412, USA

Abstract. *In silico* models of biological systems provide a powerful tool for integrative analysis of physiological function. Using the computational models of the heart as examples, we discuss three types of integration: structural integration implies integration across physical scales of biological organization from protein molecule to whole organ; functional integration of interacting physiological processes such as signalling, metabolism, excitation and contraction; and the synthesis of experimental observation with physicochemical and mathematical principles.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 4–25

During the past two decades, reductionist biological science has generated new empirical data on the molecular foundations of biological structure and function at an accelerating rate. The list of organisms whose complete genomes have been sequenced is growing by the week. Annotations of these sequences are becoming more comprehensive, and databases of protein structure are growing at impressive, indeed formerly unimaginable rates. Molecular mechanisms for fundamental processes such as ligand–receptor interactions and signal transduction are being elucidated in exquisite structural detail.

But as attention turns from gene sequencing to the next phases such as cataloguing protein structures (proteomics), it is clear to biologists that the challenge is much greater than assigning functions to individual genes. The great majority of cell functions require the coordinated interaction of numerous gene products. Metabolic or signalling pathways, for example, can be considered the expression of a 'genetic circuit', a network diagram for cellular function (Palsson 1997). But the layers of complexity do not end at the plasma membrane. Tissue and organ functions require the interactions of large ensembles of cells in functional units and networks (Boyd & Noble 1993). No amount of biochemical or single-cellular detail is sufficient to describe fully memory and learning or cardiac rhythm and pumping.

To identify the comprehensive approach that will be needed to reintegrate molecular and genetic data into a quantitative understanding of physiology and

pathophysiology in the whole organism, Bassingthwaight coined the term *physiome* (Bassingthwaight 1995; see <http://www.physiome.org/>). Other terms conveying the same general concept such as *functional genomics* and *systems biology* have entered the scientific lexicon. While achieving these goals will require the convergence of many new and emerging technologies, biology is increasingly becoming an information science, and there is no doubt that there will be a central role for information technology and mathematics, in general, and computational modelling, in particular.

Projects such as the Human Genome Project and its spin-offs have generated thousands of databases of molecular sequence and structure information such as GenBank (<http://www.ncbi.nlm.nih.gov/Genbank/>) and the Protein Data Bank (<http://www.rcsb.org/pdb/>). These databases in turn have generated demand for on-line tools for data mining, homology searching, sequence alignment and numerous other analyses. One of the best entry points for those interested in the burgeoning field of *bioinformatics* is the National Center for Biotechnology Information web site (<http://www.ncbi.nlm.nih.gov/>). Others include the Biology Workbench (<http://workbench.sdsc.edu/>) and the Integrative Biosciences portal at the San Diego Supercomputer Center (<http://biology.sdsc.edu/>). In contrast to this progress, a major obstacle to the progress in the computational modelling of integrative biological function is the lack of databases of the morphology and physiological function of cells, tissues and organs.

While there are, for example, some excellent databases of metabolic pathways such as the Metabolic Pathways Database (<http://wit.mcs.anl.gov/MPW/>) and KEGG, the Kyoto Encyclopedia of Genes and Genomes (<http://www.genome.ad.jp/kegg/>), there are not yet comprehensive public databases of myocyte ion channel kinetics or coronary vascular structure. This is one reason that investigators have focused on developing integrated theoretical and computational models. Models, even incomplete ones, can provide a formal framework for classifying and organizing data derived from experimental biology, particularly those data that serve as model parameters. Using numerical models to simulate interacting processes, one can reveal emergent properties of the system, test prediction against experimental observation, and define the specific needs for new experimental studies. The integrated models have the potential to support and inform decisions about drug design, gene targeting, biomedical engineering, and clinical diagnosis and management.

Integrative biological modelling: structural, functional and empirical–theoretical

Computational modelling of biological systems can achieve integration along several intersecting axes (Fig. 1): *structural integration* implies integration across

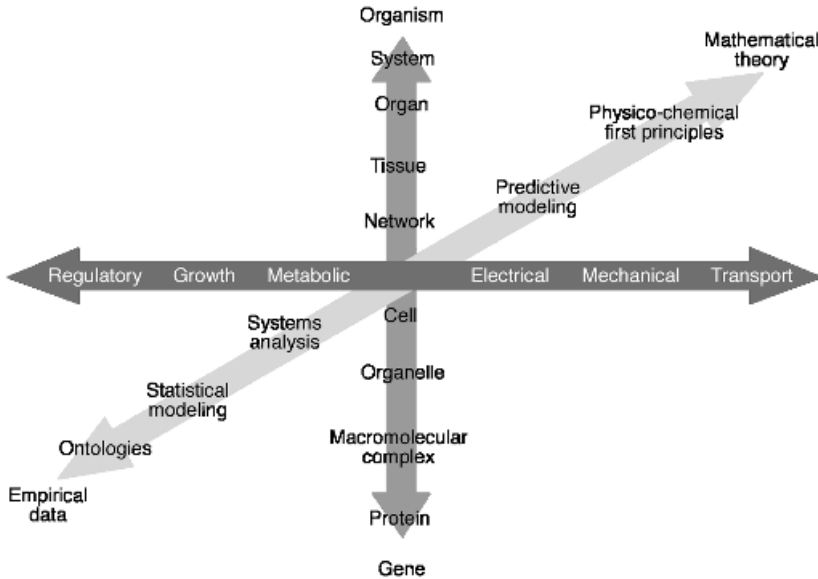


FIG. 1. Three intersecting axes of integration in computational biology: functional (darkest gray) left–right; structural (mid-gray), bottom to top; and (light gray) between data and theory.

physical scales of biological organization from protein to cell, tissue, organ, and whole organism; by *functional integration*, we mean the logical integration of coupled physiological subsystems such as those responsible for gene expression, protein synthesis, signal transduction, metabolism, ionic fluxes, cell motility and many other functions; last, but not least, as is well known from the traditions of physics and engineering, computational models serve as a powerful tool to integrate theoretical principles with empirical observations. We call this *data integration* for short.

The challenges of structurally integrated and functionally integrated computational modelling tend to be different. Functionally integrated biological modelling is a central goal of what is now being called *systems biology* (Ideker et al 2001). It is strongly data driven and therefore data intensive. Structurally integrated computational biology (such as molecular dynamics and other strategies that predict protein function from structure) is driven by physicochemical first principles and thus tends to be more computationally intensive.

Both approaches are highly complementary. Systems science is needed to bridge the large space and time scales of structural organization that span from molecule to organism, without leaving the problem computationally intractable. Structural models based on physicochemical first principles allow us to make best use of the growing databases of structural data and yet constrain the space of possible

solutions to the systems models by imposing physicochemical constraints, e.g. the protein folding problem, or the application of mass balances to metabolic flux analyses.

Therefore, most integrative biological modelling employs a combination of analysis based on physicochemical first principles and systems engineering approaches by which information can be communicated between different subsystems and across hierarchies of the integrated system. Systems models also provide a means to include within the integrated system, necessary sub-systems that are not yet characterized in sufficient detail to be modelled from first principles. This effort in turn demands new software tools for data integration, model implementation, software interoperability and model validation. It will also require a large and dedicated multidisciplinary community of scientists to accept the chore of defining ontologies and standards for structural and functional biological data representation and modelling.

Examples of the intersections between structurally and functionally integrated computational biology are becoming easier to find, not least due to the efforts of the contributors to this book:

- The linkage of biochemical networks and spatially coupled processes such as calcium diffusion in structurally based models of cell biophysics (see Loew & Schaff 2001, Loew 2002 this volume).
- The use of physicochemical constraints to optimize genomic systems models of cell metabolism (Palsson 1997, Schilling et al 2000).
- The integration of genomic or cellular systems models into multicellular network models of memory and learning (Durstewitz et al 2000, Tiesinga et al 2002), developmental pattern formation (Davidson et al 2002) or action potential propagation (Shaw & Rudy 1997).
- The integration of structure-based predictions of protein function into systems models of molecular networks.
- The development of kinetic models of cell signalling coupling them to physiological targets such as energy metabolism, ionic currents or cell motility (see Levin et al 2002, this volume).
- The use of empirical constraints to optimize protein folding predictions (Salwinski & Eisenberg 2001).
- The integration of systems models of cell dynamics into continuum models of tissue and organ physiology (Winslow et al 2000, Smith et al 2002).

Functionally integrated computational modelling of the heart

There are many reasons why a structurally and functionally integrated model of the heart is an important goal:

- Common heart diseases are multifactorial and multigenic; they are frequently linked to other systemic disorders such as diabetes, hypertension or thyroid disease.
- Cardiac structure and function are heterogeneous and most pathologies such as myocardial infarction or heart failure, are regional and non-homogeneous.
- Basic cellular functions such as pacemaker activity involve the coordinated interaction of many gene products.
- Many functional subsystems interact in fundamental physiological processes, e.g. substrate and oxygen delivery \leftrightarrow energy metabolism \leftrightarrow cross-bridge mechanoenergetics \leftrightarrow ventricular wall stress \leftrightarrow coronary flow \leftrightarrow substrate and oxygen delivery.
- Many cardiac pathologies with known or putative molecular aetiologies also depend critically on anatomic substrates for their expression *in vivo*, e.g. atrial and ventricular re-entrant arrhythmias.

Some of the aims of integrative cardiac modelling have been to integrate data and theories on the anatomy and structure, haemodynamics and metabolism, mechanics and electrophysiology, regulation and control of the normal and diseased heart. The challenges of integrating models of many aspects of such an organ system, including its structure and anatomy, biochemistry, control systems, haemodynamics, mechanics and electrophysiology has been the theme of several workshops over the past decade or so (Hunter et al 2001, McCulloch et al 1998, Noble 1995, Glass et al 1991).

Some of the major components of an integrative cardiac model that have been developed include ventricular anatomy and fibre structure (Vetter & McCulloch 1998), coronary network topology and haemodynamics (Kassab et al 1997, Kroll et al 1996), oxygen transport and substrate delivery (Li et al 1997), myocyte metabolism (Gustafson & Kroll 1998), ionic currents (Luo & Rudy 1994, Noble 1995) and impulse propagation (Winslow et al 1995), excitation–contraction coupling (Jafri et al 1998), neural control of heart rate and blood pressure (Rose & Schwaber 1996), cross-bridge cycling (Zahalak et al 1999), tissue mechanics (Costa et al 1996a,b), cardiac fluid dynamics and valve mechanics (Peskin & McQueen 1992), ventricular growth and remodelling (Lin & Taber 1995).

Of particular interest to the physician are whole organ lumped-parameter models describing transport and exchange of substrates, and accounting for the spatial distribution of the coronary arteries, regional myocardial blood flows, the uptake and metabolism of glucose, fatty acids and oxygen used for the energy to form ATP, which is in turn used to fuel the work of contraction and ion pumping. Data from nuclear medicine have been essential in this area both for estimating the kinetic parameters of mass transport in the heart, but also for providing independent measurements with which to validate such models. A unique

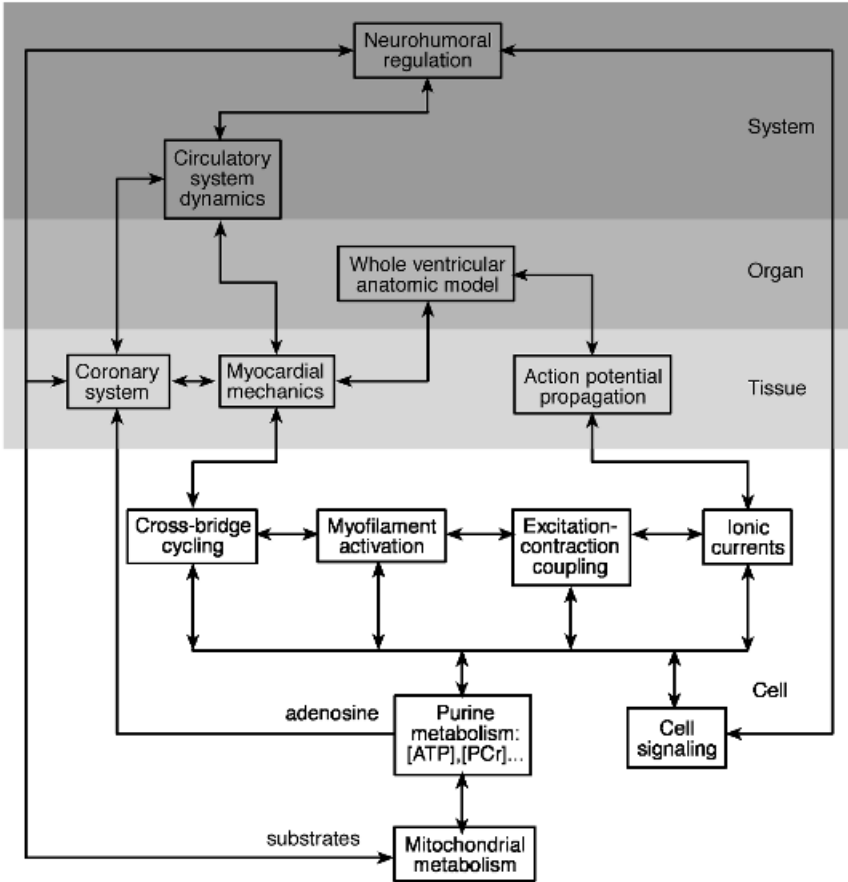


FIG. 2. Some major functional sub-systems of an integrated heart model and their hierarchical relationships from cell to tissue to organ and cardiovascular system.

resource for numerical models and simulation for circulatory mass transport and exchange is the National Simulation Resource (<http://nsr.bioeng.washington.edu>).

To explore, how these models can be extended and integrated with others, workers in the field have defined several major functional modules for initial attention, as shown in Fig. 2, which has been adapted and expanded from the scheme proposed by Bassingthwaighte (Bassingthwaighte 1997). They include:

- Coronary artery anatomy and *regional myocardial flows* for substrate and oxygen delivery.

- Metabolism of the substrate for *energy metabolism*, fatty acid and glucose, the tricarboxylic acid (TCA) cycle, and *oxidative phosphorylation*.
- *Purine nucleoside and purine nucleotide metabolism*, describing the formation of ATP and the regulation of its degradation to adenosine in endothelial cells and myocytes, and its effects on coronary vascular resistance.
- The *transmembrane ionic currents* and their *propagation* across the myocardium
- *Excitation-contraction coupling*: calcium release and reuptake, and the relationships between these and the strength and extent of sarcomere shortening.
- *Sarcomere dynamics* of myofilament activation and cross-bridge cycling, and the *three-dimensional mechanics* of the ventricular myocardium during the cardiac cycle.
- *Cell signalling* and the *autonomic control* of cardiac excitation and contraction.

Naturally, the scheme in Fig. 2 contains numerous omissions such as the coronary venous system and its interactions with myocardial stresses, regulation of intracellular enzymes by secondary processes, vascular and tissue remodelling, protein metabolism, systemic influences on total body vascular resistance, changes in cardiac pool sizes of glycogen and di- and triphosphoglycerides, neurohumoral regulation of contractility and coronary flow, and many other features. Nevertheless, it provides a framework to incorporate these features later. More importantly, despite these limitations, a model like this should provide an opportunity to answer important questions in integrative cardiac physiology that have eluded intuitive understanding. One excellent example is the physical and biological basis of flow and contractile heterogeneity in the myocardium. Another is the role of intracellular inorganic phosphate accumulation on contractile dysfunction during acute myocardial ischaemia.

While Fig. 2 does show different scales in the structural hierarchy, it emphasizes functional integration, and thus it is not surprising that the majority of functional interactions take place at the scale of the single cell. In this view, a systems model of functionally interacting networks in the cell can be viewed as a foundation for structurally coupled models that extend to multicellular networks, tissue, organ and organ system. But it can also be viewed as a focal point into which feed structurally based models of protein function and subcellular anatomy and physiology. We explore this view further in the following section.

Structurally integrated models of the heart

A fundamental challenge of biological science is the integration of information across scales of length and time that span many orders of magnitude from molecular structures and events to whole-organ anatomy and physiology. As more and more detailed data accumulate on the molecular structure and diversity

of living systems, there is an increasing need to develop computational analyses that can be used to integrate functions across the hierarchy of biological organization, from atoms to macromolecules, cells, tissues, organs, organ systems and ultimately the whole organism.

Predictive computational models of various processes at almost every individual level of the hierarchy have been based on physicochemical first principles. Although important insight has been gained from empirical models of living systems, models become more predictive if the number of adjustable parameters is reduced by making use of detailed structural data and the laws of physics to constrain the solution. These models, such as molecular dynamics simulations, spatially coupled cell biophysical simulations, tissue micromechanical models and anatomically based continuum models are usually computationally intensive in their own right.

But to be most valuable in post-genomic biological science, they must also be integrated with each other across scales of biological organization. This will require a computational infrastructure that will allow us to integrate physically based biological models that span the hierarchy from the dynamics of individual protein molecules up to the regional physiological function of the beating heart. This software will have to make use of computational resources that are distributed and heterogeneous, and be developed in a modular manner that will facilitate integration of new models and levels.

Two examples from cardiac physiology illustrate the potential significance of structurally integrated modelling: In the clinical arrhythmogenic disorder long-QT syndrome, a mutation in a gene coding for a cardiomyocyte sodium or potassium selective ion channel alters its gating kinetics. This small change at the molecular level affects the dynamics and fluxes of ions across the cell membrane and thus affects the morphology of the recorded electrocardiogram (prolonging the QT interval) and increasing the vulnerability to life-threatening cardiac arrhythmia. Such an understanding could not be derived by considering only the single gene, channel or cell; it is an integrated response across scales of organization. A hierarchical integrative simulation could be used to analyse the mechanism by which this genetic defect can lead to sudden cardiac death by, for example, exploring the effects of altered repolarization on the inducibility and stability of re-entrant activation patterns in the whole heart. A recent model study by Clancy & Rudy (1999) made excellent progress at spanning some of these scales by incorporating a Markov model of altered channel gating—based on the structural consequences of the genetic defect in the cardiac sodium channel—into a whole cell kinetic model of the cardiac action potential that included all the major ionic currents.

As a second example, it is becoming clearer that mutations in specific proteins of the cardiac muscle contractile filament system lead to structural and developmental

abnormalities of muscle cells, impairment of tissue contractile function and the eventual pathological growth (hypertrophy) of the whole heart as a compensatory response (Chien 1999). In this case, the precise physical mechanisms at each level remain speculative, though much detail has been elucidated recently, so an integrative model will be useful for testing various hypotheses regarding the mechanisms. The modelling approach could be based on the same integrative paradigm commonly used by experimental biologists, wherein the integrated effect of a specific molecular defect or structure can be analysed using techniques such as *in vivo* gene targeting.

Investigators have developed large-scale numerical methods for *ab initio* simulation of biophysical processes at the following levels of organization: molecular dynamics simulations based on the atomic structure of biomolecules; hierarchical models of the collective motions of large assemblages of monomers in macromolecular structures (Huber 2002); biophysical models of the dynamics of cross-bridge interactions at the level of the cardiac contractile filaments (Landesberg et al 2000); whole-cell biophysical models of the regulation of muscle contraction (Bluhm et al 1998); microstructural constitutive models of the mechanics of multicellular tissue units (MacKenna et al 1997); continuum models of myocardial tissue mechanics (Costa et al 2001) and electrical impulse propagation (Rogers & McCulloch 1994); and anatomically detailed whole organ models (Vetter & McCulloch 2000).

They have also investigated methods to bridge some of the boundaries between the different levels of organization. We and others have developed finite-element models of the whole heart, incorporating microstructural constitutive laws and the cellular biophysics of thin filament activation (Mazhari et al 2000). Recently, these mechanics models have been coupled with a non-linear reaction–diffusion equation model of electrical propagation incorporating an ionic cellular model of the cardiac action potential and its regulation by stretch (Vetter & McCulloch 2001). At the other end of the hierarchy, Huber (2002) has recently developed a method, the Hierarchical Collective Motions method, for integrating molecular dynamics simulation results from small sections of a large molecule into a quasi-continuum model of the entire molecule.

The different levels of description are illustrated in Fig. 3. In order to prevent the models from being overwhelmed by an explosion of detail, only a representative subset of structures from the finer level can be used directly; the behaviour of the remainder must be inferred by spatial interpolation. This approach has been used in software packages such as our program *Continuity* or the CONNFESSIT models of polymer rheology (Laso & Ottinger 1993) to span two or three levels of organization. The modelling infrastructure must therefore support not only software modules required to solve the structures at each level of the hierarchy, it must also support adapter functions between modules. In some cases the

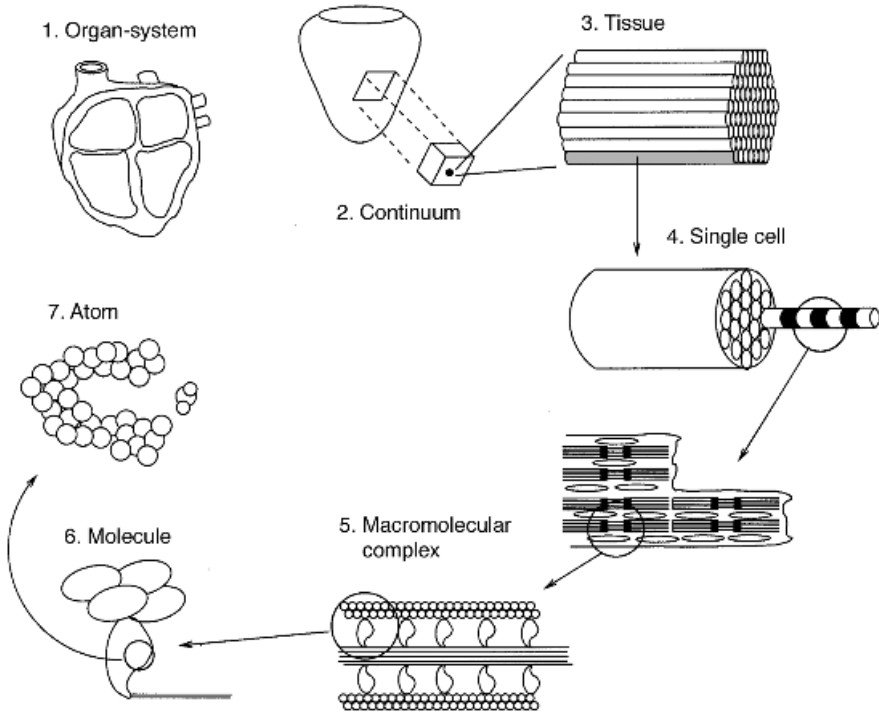


FIG. 3. Scales of a structurally integrated heart model from atomic resolution to organ system.

communication between levels is direct; the output of one level, such as transmembrane potential or myofibril stress is a more or less direct input to the level above. In others, the results of computations on the finer structure need to be parameterized to meet the requirements of the coarser level. The amount of detail and bidirectional communication required between levels is not only a function of the structures being modelled but the question being investigated. Experimenting with different degrees of coupling between levels of the hierarchy will likely be an important new path to scientific discovery.

The disparity of time scales is as significant as that of spatial scale. For example, the period of the cardiac cycle is about 1 s, the time steps of the cellular model of the cardiac action potential are shorter than a millisecond for the fastest kinetics, while the time steps of an atomic-level simulation are on the order of femtoseconds. Running atomic-level simulations for the entire length of a physiological simulation time step would not be feasible. However, in many situations it is not necessary to run the simulation for the full duration of the time step of the level immediately above, because the response of the lower level will converge relatively

TABLE 1 Models at each physical scale and the bridges between them

<i>Scale</i>	<i>Class of Model</i>	<i>Mechanics Example</i>	<i>Electrophysiology example</i>
Organ system	Lumped parameter model	Arterial circuit equivalent	Equivalent dipole EKG
	External boundary conditions	Haemodynamic loads	No flux condition
Whole organ	Continuum PDE model	Galerkin FE stress analysis	Collocation FE model
	Constitutive model	Constitutive law for stress	Anisotropic diffusion
Tissue	Multicellular network model	Tissue micromechanics model	Resistively coupled network
Multicellular	Cell–cell/cell–matrix coupling	Matrix micromechanics model	Gap junction model
Single cell	Whole cell systems model	Myocyte 3D stiffness and contractile mechanics	Myocyte ionic current and flux model
Subcellular	Subcellular compartment model	Sarcomere dynamics model	Intracellular calcium fluxes
	Stochastic state-transition model	Cross-bridge model of actin–myosin interaction	Single channel Markov model
Macromolecular	Weighted ensemble Brownian dynamics	Single cross-bridge cycle	Ion transport through single channel
Molecular	Hierarchical collective motions	Actin, myosin, tropomyosin	Na ⁺ , K ⁺ and Ca ⁺ channels
Atomic	Molecular dynamics simulation	PDB coordinates	PDB coordinates

EKG, electrocardiogram; FE, finite element; PDB, Protein Data Bank; PDE, partial differential equation.

quickly. Such a response will be characterized by either equilibrium or quasi-steady-state behaviour. On levels close to the atomic end of the hierarchy, the response is characterized by the infrequent crossing of free energy barriers, driven by thermal fluctuations. In such cases, we have developed special algorithms, such as *weighted-ensemble Brownian dynamics* (Huber & Kim 1996), to circumvent the disparity between the frequency of barrier crossing and the simulation time step size.

We identify eight levels of biological organization from atomic scale to whole organ system as depicted in Fig. 3. Separate classes of model represent each scale with intervening models that bridge between across scales. For example, a weighted ensemble Brownian dynamics simulation of ion transport through a

single channel can be used to compute channel gating properties from the results of a hierarchical collective motions simulation of the channel complex. Homogenization theory can be used to derive a constitutive model that re-parameterizes the results of a micromechanical analysis into a form suitable for continuum scale stress analysis. Table 1 shows these scales, the classes of models that apply at each scale and that bridge between each scale, and examples from possible simulations of cardiac electrical and mechanical function. At each level, investigators have already implemented models (some sophisticated and some more simple) that model this level or that bridge between them.

Organ system model

The top level can be represented by a lumped parameter systems model of arterial impedance used to generate the dynamic pressure boundary conditions acting on the cardiac chambers. In the case of electrophysiology, we have the transfer function for integrating the electrical dipole and whole body electrocardiogram from the current sources generated by the sequence of cardiac electrical activation and repolarization.

Whole heart continuum model

Finite element methods have been used to solve the continuum equations for myocardial mechanics (Costa et al 1996) or action potential propagation (Rogers & McCulloch 1994). In the case of cardiac mechanics, boundary conditions such as ventricular cavity pressures are computed from the lumped parameter model in the top level. Detailed parametric models of three-dimensional cardiac geometry and muscle fibre orientations have been used to represent the detailed structure of the whole organ with sub-millimetre resolution (Vetter & McCulloch 1998).

Tissue model

Constitutive laws for the continuum models are evaluated at each point in the continuum scale model and obtained by homogenizing the results of multicellular network models. In the case of tissue mechanics, these represent ensembles of cell and matrix micromechanics models and, in some cases, the microvascular blood vessels too (May-Newman & McCulloch 1998). These models represent basic functional units of the tissue, such as the laminar myocardial sheets. Workers have used a variety of approaches for these models including stochastic models based on measured statistical distributions of myofibre orientations (Usyk et al 2001). In cardiac electrophysiology, this level is typically modelled as resistively coupled networks of discrete cellular models interconnected in three dimensions (Leon & Roberge 1991).

Single cell model

This level models representative myocytes from different myocardial regions, such as epicardial cells, mid-ventricular M-cells and endocardial cells. For mechanics models, individual myofibrils and cytoskeletal structures are modelled by lattices and networks of rods, springs and dashpots in one, two or three dimensions. Single cell electrophysiological models are well established as described elsewhere in this book (Noble 2002, this volume). Single cell models bridge to stochastic state-transition models of macromolecular function through subcellular compartment models of representative structures such as the sarcomere. Another example is diffusive or Monte-Carlo models of intracellular calcium transfer between restricted micro-domains and the bulk myoplasm.

Macromolecular complex model

This is the level of representative populations of cross-bridges or ion channels. They are described by Markov models of stochastic transitions between discrete states of, for example, channel gating, actin-myosin binding or nucleotide bound to myosin.

Molecular model

The penultimate level is composed of reduced-variable, or normal-mode-type models of the single cross-bridges and ion channels as computed by the hierarchical collective motions (HCM) model. The cross-bridges will move according to Brownian dynamics, and it will be necessary to use weighted-ensemble dynamics to allow the simulation to clear the energy barriers. The flexibility of the cross bridges themselves will be derived from the HCM method, and the interactions with other molecules will be computed using continuum solvent approximations.

Atomic model

The final level involves descriptions at the atomic scale based on crystallography structures of these molecules in public repositories such as the Protein Data Bank. The dynamics of representative myosin heads, actin monomers, ion channel or troponin subunits, are simulated at atomic resolution using molecular dynamics, in order to build the HCM model. Certain key parts, such as binding sites, channel gating sites, or voltage sensor, must be kept at atomic detail during coupling with the level above.

Summary

Although the main emphasis of this paper is on the mechanics and electrophysiology of the heart, other aspects of cardiac physiology could be modelled using a similar framework. The approach should also be adaptable to other tissues and organs especially those with physical functions, such as lung and cartilage. Such integrative models are composed of a hierarchy of simulation levels, each implemented by a set of communicating program modules. Substantial experimental data and theoretical modelling has been done at each level from the biomechanics of the myocardium and myocytes to the biophysics of the sarcomere and the structural biology of the cross-bridge and contractile filament lattice. Many other questions remain unanswered: for example, how the geometry of the myofibril lattice leads to transverse as well as longitudinal stresses remains unclear (Lin & Yin 1998).

In order to carry out numerical experiments to complement *in vitro* and *in vivo* experiments, a flexible and composable simulation infrastructure will be required. It is not realistic to expect that any single integrative analysis will include atomic or even molecular resolution detail of more than a small subset of proteins involved in the physiological response. Instead, the path to discovery will follow the one used in experimental biology. Models will be used to compare the effects of a specific molecular structure or mutation on the integrated response.

Acknowledgements

Both authors are supported by grants from the National Science Foundation. ADM is also supported by the Procter and Gamble International Program for Animal Alternatives and grants from the National Institutes of Health, including the National Biomedical Computation Resource (<http://nbcrc.sdsc.edu/>) through a National Center for Research Resources program grant (P 41 RR08605).

References

- Bassingthwaighte JB 1995 Toward modeling the human physiome. *Adv Exp Med Biol* 382:331–339
- Bassingthwaighte JB 1997 Design and strategy for the Cardionome Project. *Adv Exp Med Biol* 430:325–339
- Bluhm WF, Sung D, Lew WY, Garfinkel A, McCulloch AD 1998 Cellular mechanisms for the slow phase of the Frank-Starling response. *J Electrocardiol* 31:S13–S22
- Boyd CAR, Noble D (eds) 1993 *The logic of life: the challenge of integrative physiology*. Oxford University Press, New York
- Chien KR 1999 Stress pathways and heart failure. *Cell* 98:555–558

- Clancy CE, Rudy Y 1999 Linking a genetic defect to its cellular phenotype in a cardiac arrhythmia. *Nature* 400:566–569
- Costa KD, Hunter PJ, Rogers JM, Guccione JM, Waldman LK, McCulloch AD 1996a A three-dimensional finite element method for large elastic deformations of ventricular myocardium: I—Cylindrical and spherical polar coordinates. *J Biomech Eng* 118:452–463
- Costa KD, Hunter PJ, Wayne JS, Waldman LK, Guccione JM, McCulloch AD 1996b A three-dimensional finite element method for large elastic deformations of ventricular myocardium: II—Prolate spheroidal coordinates. *J Biomech Eng* 118:464–472
- Costa KD, Holmes JW, McCulloch AD 2001 Modeling cardiac mechanical properties in three dimensions. *Phil Trans R Soc Lond A Math Phys Sci* 359:1233–1250
- Davidson EH, Rast JP, Oliveri P et al 2002 A genomic regulatory network for development. *Science* 295:1669–1678
- Durstewitz D, Seamans JK, Sejnowski TJ 2000 Neurocomputational models of working memory. *Nat Neurosci* 3:S1184–S1191
- Glass L, Hunter P, McCulloch AD (eds) 1991 *Theory of heart: biomechanics, biophysics and nonlinear dynamics of cardiac function*. Institute for Nonlinear Science. Springer-Verlag, New York
- Gustafson LA, Kroll K 1998 Downregulation of 5′-nucleotidase in rabbit heart during coronary underperfusion. *Am J Physiol* 274:H529–H538
- Huber G 2002 The Hierarchical Collective Motions method for computing large-scale motions of biomolecules. *J Comp Chem*, in press
- Huber GA, Kim S 1996 Weighted-ensemble Brownian dynamics simulations for protein association reactions. *Biophys J* 70:97–110
- Hunter PJ, Kohl P, Noble D 2001 Integrative models of the heart: achievements and limitations. *Phil Trans R Soc Lond A Math Phys Sci* 359:1049–1054
- Ideker T, Galitski T, Hood L 2001 A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2:343–372
- Jafri MS, Rice JJ, Winslow RL 1998 Cardiac Ca²⁺ dynamics: the roles of ryanodine receptor adaptation and sarcoplasmic reticulum load [published erratum appears in 1998 *Biophys J* 74:3313]. *Biophys J* 74:1149–1168
- Kassab GS, Berkley J, Fung YC 1997 Analysis of pig's coronary arterial blood flow with detailed anatomical data. *Ann Biomed Eng* 25:204–217
- Kroll K, Wilke N, Jerosch-Herold M et al 1996 Modeling regional myocardial flows from residue functions of an intravascular indicator. *Am J Physiol* 271:H1643–1655
- Landesberg A, Livshitz L, Ter Keurs HE 2000 The effect of sarcomere shortening velocity on force generation, analysis, and verification of models for crossbridge dynamics. *Ann Biomed Eng* 28:968–978
- Laso M, Ottinger HC 1993 Calculation of viscoelastic flow using molecular models: the CONFESSIT approach. *Non-Newtonian Fluid Mech* 47:1–20
- Leon LJ, Roberge FA 1991 Directional characteristics of action potential propagation in cardiac muscle. A model study. *Circ Res* 69:378–395
- Levin JM, Penland RC, Stamps AT, Cho CR 2002 In: *'In silico' simulation of biological processes*. Wiley, Chichester (Novartis Found Symp 247) p 227–243
- Li Z, Yipintsoi T, Bassingthwaighte JB 1997 Nonlinear model for capillary-tissue oxygen transport and metabolism. *Ann Biomed Eng* 25:604–619
- Lin IE, Taber LA 1995 A model for stress-induced growth in the developing heart. *J Biomech Eng* 117:343–349
- Lin DHS, Yin FCP 1998 A multi-axial constitutive law for mammalian left ventricular myocardium in steady-state barium contracture or tetanus. *J Biomech Eng* 120:504–517
- Loew LM 2002 In: *'In silico' simulation of biological processes*. Wiley, Chichester (Novartis Found Symp 247) p 151–161

- Loew LM, Schaff JC 2001 The virtual cell: a software environment for computational cell biology. *Trends Biotechnol* 19:401–406
- Luo C-H, Rudy Y 1994 A dynamic model of the cardiac ventricular action potential. I. Simulation of ionic currents and concentration changes. *Circ Res* 74:1071–1096
- MacKenna DA, Vaplon SM, McCulloch AD 1997 Microstructural model of perimysial collagen fibers for resting myocardial mechanics during ventricular filling. *Am J Physiol* 273: H1576–H1586
- May-Newman K, McCulloch AD 1998 Homogenization modelling for the mechanics of perfused myocardium. *Prog Biophys Mol Biol* 69:463–482
- Mazhari R, Omens JH, Covell JW, McCulloch AD 2000 Structural basis of regional dysfunction in acutely ischemic myocardium. *Cardiovasc Res* 47:284–293
- McCulloch A, Bassingthwaite J, Hunter P, Noble D 1998 Computational biology of the heart: from structure to function [editorial]. *Prog Biophys Mol Biol* 69:153–155
- Noble D 1995 The development of mathematical models of the heart. *Chaos Soliton Fract* 5: 321–333
- Noble D 2002 The heart *in silico*: successes, failures and prospects. In: ‘*In silico*’ simulation of biological processes. Wiley, Chichester (Novartis Found Symp 247) p 182–197
- Palsson BO 1997 What lies beyond bioinformatics? *Nat Biotechnol* 15:3–4
- Peskin CS, McQueen DM 1992 Cardiac fluid dynamics. *Crit Rev Biomed Eng* 29:451–459
- Rogers JM, McCulloch AD 1994 Nonuniform muscle fiber orientation causes spiral wave drift in a finite element model of cardiac action potential propagation. *J Cardiovasc Electrophysiol* 5:496–509
- Rose WC, Schwaber JS 1996 Analysis of heart rate-based control of arterial blood pressure. *Am J Physiol* 271:H812–H822
- Salwinski L, Eisenberg D 2001 Motif-based fold assignment. *Protein Sci* 10:2460–2469
- Schilling CH, Edwards JS, Letscher D, Palsson BO 2000 Combining pathway analysis with flux balance analysis for the comprehensive study of metabolic systems. *Biotechnol Bioeng* 71: 286–306
- Shaw RM, Rudy Y 1997 Electrophysiologic effects of acute myocardial ischemia: a mechanistic investigation of action potential conduction and conduction failure. *Circ Res* 80:124–138
- Smith NP, Mulquiney PJ, Nash MP, Bradley CP, Nickerson DP, Hunter PJ 2002 Mathematical modelling of the heart: cell to organ. *Chaos Soliton Fract* 13:1613–1621
- Tiesinga PH, Fellous JM, Jose JV, Sejnowski TJ 2002 Information transfer in entrained cortical neurons. *Network* 13:41–66
- Usyk TP, Omens JH, McCulloch AD 2001 Regional septal dysfunction in a three-dimensional computational model of focal myofiber disarray. *Am J Physiol* 281:H506–H514
- Vetter FJ, McCulloch AD 1998 Three-dimensional analysis of regional cardiac function: a model of rabbit ventricular anatomy. *Prog Biophys Mol Biol* 69:157–183
- Vetter FJ, McCulloch AD 2000 Three-dimensional stress and strain in passive rabbit left ventricle: a model study. *Ann Biomed Eng* 28:781–792
- Vetter FJ, McCulloch AD 2001 Mechanoelectric feedback in a model of the passively inflated left ventricle. *Ann Biomed Eng* 29:414–426
- Winslow R, Cai D, Varghese A, Lai Y-C 1995 Generation and propagation of normal and abnormal pacemaker activity in network models of cardiac sinus node and atrium. *Chaos Soliton Fract* 5:491–512
- Winslow RL, Scollan DF, Holmes A, Yung CK, Zhang J, Jafri MS 2000 Electrophysiological modeling of cardiac ventricular function: From cell to organ. *Ann Rev Biomed Eng* 2:119–155
- Zahalak GI, de Laborderie V, Guccione JM 1999 The effects of cross-fiber deformation on axial fiber stress in myocardium. *J Biomech Eng* 121:376–385

DISCUSSION

Noble: You have introduced a number of important issues, including the use of modelling to lead the way in problem resolution. You gave some good examples of this. You also gave a good example of progressive piecing together: building on what is already there. One important issue you raised that I'd be keen for us to discuss is that of modelling across scales. You referred to something called HCM: would you explain what this means?

McCulloch: The principle of HCM is an algorithm by which Gary Huber breaks down a large protein molecule — the example he has been working on is an actin filament — and models a small part of it. He then extracts modes that are of interest from this molecular dynamics simulation over a short time (e.g. principle modes of vibration of that domain of the protein). He takes this and applies it to the other units, and repeats the process at a larger scale. It is a bit like a molecular multigrid approach, whereby at successive scales of resolution he attempts to leave behind the very high-frequency small-displacement perturbations that aren't of interest, and accumulate the larger displacements and slower motions that are of interest. The result is that in early prototypes he is able to model a portion of an actin filament with, say, 50 G-actin monomers wiggling around and accumulates the larger Brownian motion scale that would normally be unthinkable from a molecular dynamics simulation.

Subramaniam: That is a fairly accurate description. HCM involves coarse-graining in time scale and length scale. He is successfully coarse graining where the parameterization for the next level comes from the lower level of coarse graining. Of course, what Gary would eventually like to resolve, going from one set of simulations to the next hierarchy of simulations, is starting from molecular dynamics to go into Brownian dynamics or stochastic dynamics, from which he can go into continuum dynamics and so forth. HCM is likely to be very successful in large-scale motions of molecular assemblies, where we cannot model detailed atomic-level molecular dynamics.

Noble: Is this effectively the same as extracting from the lower level of modelling just those parameters in which changes are occurring over the time-scale relevant to the higher-level modelling?

Subramaniam: Yes, with one small caveat. Sometimes very small-scale motions may contribute significantly to the next hierarchy of modelling. This would not be taken into account in a straightforward parameterization approach. Since the scales are not truly hierarchically coupled, there may be a small-scale motion that can cause a large-scale gradient in the next level of hierarchy. Gary's method would take this into account.

Noble: Is the method that this can automatically be taken into account, or will it require a human to eyeball the data and say that this needs to be included?

McCulloch: He actually does it himself; it is not automatic yet. But the process that he uses is not particularly refined. It could certainly be automated.

Cassman: You are extracting a certain set of information out of a fairly complex number of parameters. You made a decision that these long time-scales are what you are going to use. But of course, if you really want to know something about the motion of the protein in its native environment, it is necessary to include all of the motions. How do you decide what you put in and what you leave out, and how do you correct for this afterwards? I still don't quite see how this was arrived at.

McCulloch: The answer is that it probably depends on what the purpose of the analysis is. In the case of the actin filament, Gary was looking for the motion of a large filament. A motion that wouldn't affect the motion of neighbouring monomers was not of interest. In this case it was fairly simple, but when it comes to biological functions it is an oversimplification just to look at whether it moves or not.

Noble: When you say that it all depends on what the functionality is that you want to model, this automatically means that there will be many different ways of going from the lower level to the upper level. This was incidentally one of the reasons why in the discussion that took place at the Novartis Foundation symposium on *Complexity in biological information processing* (Novartis Foundation 2001), the conclusion that taking the bottom-up route was not possible emerged. In part, it was not just the technical difficulty of being able to do it—even if you have the computing power—but also because you need to take different functionalities from the lower-level models in order to go to the higher-level ones, depending on what it is you are trying to do.

Hunter: There is a similar example of this process that might illustrate another aspect of it. For many years we have been developing a model of muscle mechanics, which involves looking at the mechanics of muscle trabeculae and then from this extracting a model that captures the essential mechanical features at the macro level. Recently, Nic Smith has been looking at micromechanical models of cross-bridge motion and has attempted to relate the two. In this, he is going from the scale of what a cross-bridge is doing to what is happening at the continuum level of a whole muscle trabecula. The way we have found it possible to relate these two scales is to look at the motion at the cross-bridge level and extract the eigenvectors that represent the dominant modes of action of that detailed structural model. From these eigenvectors we then get the information that we can relate to the higher-level continuum models. This does seem to be an effective way of linking across scales.

Subramaniam: Andrew McCulloch, in your paper you illustrated nicely the fact that you need to integrate across these different time-scales. You took a phenomenon at the higher level, and then used biophysical equations to model it. When you think of pharmacological intervention, this happens at a molecular

level. For example, take cardiomyopathy: intervention occurs by means of a single molecule acting at the receptor level. Here, you have used parameters that have really abstracted this molecular level.

McCulloch: In the vast majority of our situations, where we do parameterize the biophysical model in terms of quantities that can be related to drug action, the source of the data is experimental. It is possible to do experiments on single cells and isolated muscles, such as adding agonists and then measuring the alteration in channel conductance or the development of force. We don't need to use *ab initio* simulations to predict how a change in myofilament Ca^{2+} sensitivity during ischaemia gives rise to alterations in regional mechanics. We can take the careful measurements that have been done *in vitro*, parameterize them in terms of quantities that we know matter, and use these.

Subramaniam: So your parameters essentially contain all the information at the lower level.

McCulloch: They don't contain it all, but they contain the information that we consider to be important.

Noble: You gave some nice examples of the use of modelling to lead the way in trying to resolve the problem of the Anrep effect. I would suggest that it is not just a contingent fact that in analysing this Anrep effect your student came up with internal Na^+ being a key. The reason for this is that I think that one of the functions of modelling complex systems is to try to find out what the drivers are in a particular situation. What are the processes that, once they have been identified, can be regarded as the root of many other processes? Once this is understood, we are then in the position where we have understood part of the logic of the situation. The reason I say that it is no coincidence that Na^+ turned out to be important is that is a sort of driver. There is a lot of Na^+ present, so this will change relatively slowly. Once you have identified the group of processes that contribute to controlling that, you will in turn be able to go on to understand a huge number of other processes. The Anrep effect comes out. So also will change in the frequency of stimulation. I could go on with a whole range of things as examples. It seems that one of the functions of complex modelling is to try to identify the drivers. Do you agree?

McCulloch: Yes, I think that is a good point. I think an experienced electrophysiologist would perhaps have deduced this finding intuitively. But in many ways the person who was addressing the problem was not really an experienced electrophysiologist, so the model became an 'expert system' as much as a fundamental simulation for learning about the cell and rediscovering phenomena. This was a situation where we were able to be experimentally useful by seeking a driver.

Winslow: I think this is a good example of a biological mechanism that is a kind of nexus point. Many factors affect Na^+ and Ca^{2+} in the myocyte, which in turn affect

many other processes in the myocyte. These mechanisms are likely to be at play across a wide range of behaviours in the myocyte. Identifying these nexus points with high fan in and high fan out in biological systems is going to be key.

Noble: Andrew McCulloch, when you said that you thought a good electrophysiologist could work it out, this depends on there being no surprises or counterintuitive effects. I think we will find during this meeting that modelling has shown there to be quite a lot of such traps for the unwary. I will do a *mea culpa* in my paper on some of the big traps that nature has set for us, and the way in which modelling has enabled us to get out of these.

Cassman: You are saying that one of the functions of modelling is to determine what the drivers are for a process. But what you get out depends on what you put in. You are putting into the model only those things that you know. What you will get out of the model will be the driver based on the information that you have. It could almost be seen as a circular process. When do you get something new out of it, that is predictive rather than simply descriptive of the information that you have already built into the model?

McCulloch: The only answer I can give is when you go back and do more experiments. It is no accident that three-quarters of the work in my laboratory is experimental. This is because at the level we are modelling, the models in and of themselves don't live in isolation. They need to go hand in hand with experiments. In a way, the same caveat can be attached to experimental biology. Experimental biology is always done within the domain of what is known. There are many assumptions that are implicit in experiments. Your point is well taken: we were never going to discover a role for Na^+/H^+ exchange in the Anrep effect with a model that did not have that exchanger in it.

Noble: No, but what you did do was identify that given that Na^+ was the driver, it was necessary to take all the other Na^+ transporters into account. In choosing what then to include in your piecemeal progressive building of humpty dumpty, you were led by that.

Paterson: Going back to the lab, the experiments were preceded by having a hypothesis. Where things get really interesting is when there is a new phenomenon that you hadn't anticipated, and when you account for your current understanding of the system, that knowledge cannot explain the phenomenon that you just observed. Therefore, you know that you are missing something. You might be able to articulate several hypotheses, and you go back to the lab to find out which one is correct. What I find interesting is how you prioritize what experiment to run to explore which hypothesis, given that you have limited time and resources. While the iterative nature of modelling and data collection is fundamental, applied research, as in pharmaceutical research and development, must focus these iterations on improving their decision-making under tremendous time and cost pressures.

Boissel: I have two points. First, I think that this discussion illustrates that we are using modelling simply as another way of looking at what we already know. It is not something that is very different from the literary modelling that researchers have been doing for centuries. We are integrating part of what we know in such a way that we can investigate better what we know, nothing more. Second, all the choices that we have to make in setting up a model are dependent on the purpose of the model. There are many different ways of modelling the same knowledge, depending on the use of the model.

McCulloch: I agree with your second point. But I don't agree with your first point — that models are just a collection of knowledge. These models have three levels or components. One is the set of data, or knowledge. The second is a system of components and their interactions. The third is physicochemical first principles: the conservation of mass, momentum, energy and charge. Where these types of models have a particular capacity to integrate and inform is through imposing constraints on the way the system could behave. In reality, biological processes exist within a physical environment and they are forced to obey physical principles. By imposing physicochemical constraints on the system we can do more than simply assemble knowledge. We can exclude possibilities that logic may not exclude but the physics does.

Boissel: I agree, but for me, the physicochemical constraints you put in the model are also a part of our knowledge.

Loew: It seems to me that the distinction between traditional modelling that biologists have been doing for the last century, and the kind of modelling that we are concerned with here, is the application of computational approaches. The traditional modelling done by biologists has all been modelling that can be accomplished by our own brain power or pencil and paper. In order to deal with even a moderate level of complexity, say of a dozen or so reactions, we need computation. One of the issues for us in this meeting is that someone like Andrew McCulloch, who does experiments and modelling at the same time, is relatively rare in the biological sciences. Yet we need to use computational approaches and mathematical modelling approach to understand even moderately complicated systems in modern biology. How do we get biologists to start using these approaches?

Boissel: I used to say that formal modelling is quite different from traditional modelling, just because it can integrate quantitative relations between the various pieces of the model.

Levin: A brief comment: I thought that what has been highlighted so well by Andrew McCulloch, and illustrates the distinction of what modelling was 20 years ago and what modelling is today, is the intimate relationship between experimentation and the hypotheses that are generated by modelling.

Reference

Novartis Foundation 2001 Complexity in biological information processing. Wiley, Chichester
(Novartis Found Symp 239)

Advances in computing, and their impact on scientific computing

Mike Giles

Oxford University Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, UK

Abstract. This paper begins by discussing the developments and trends in computer hardware, starting with the basic components (microprocessors, memory, disks, system interconnect, networking and visualization) before looking at complete systems (death of vector supercomputing, slow demise of large shared-memory systems, rapid growth in very large clusters of PCs). It then considers the software side, the relative maturity of shared-memory (OpenMP) and distributed-memory (MPI) programming environments, and new developments in 'grid computing'. Finally, it touches on the increasing importance of software packages in scientific computing, and the increased importance and difficulty of introducing good software engineering practices into very large academic software development projects.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 26-41

Hardware developments

In discussing hardware developments, it seems natural to start with the fundamental building blocks, such as microprocessors, before proceeding to talk about whole systems. However, before doing so it is necessary to make the observation that the nature of scientific supercomputers has changed completely in the last 10 years.

Ten years ago, the fastest supercomputers were highly specialized vector supercomputers sold in very limited numbers and used almost exclusively for scientific computations. Today's fastest supercomputers are machines with very large numbers of commodity processors, in many cases the same processors used for word processing, spreadsheet calculations and database management. This change is a simple matter of economics. Scientific computing is a negligibly small fraction of the world of computing today, so there is insufficient turnover, and even less profit, to justify much development of custom hardware for scientific applications. Instead, computer manufacturers build high-end systems out of the

building blocks designed for everyday computing. Therefore, to predict the future of scientific computing, one has to look at the trends in everyday computing.

Building blocks

Processors. The overall trend in processor performance continues to be well represented by Moore's law, which predicts the doubling of processor speed every 18 months. Despite repeated predictions of the coming demise of Moore's law because of physical limits, usually associated with the speed and wavelength of light, the vast economic forces lead to continued technological developments which sustain the growth in performance, and this seems likely to continue for another decade, driven by new demands for speech recognition, vision processing and multimedia applications.

In detail, this improvement in processor performance has been accomplished in a number of ways. The feature size on central processing unit (CPU) chips continues to shrink, allowing the latest chips to operate at 2 GHz. At the same time, improvements in manufacturing have allowed bigger and bigger chips to be fabricated, with many more gates. These have been used to provide modern CPUs with multiple pipelines, enabling parallel computation within each chip. Going further in this direction, the instruction scheduler becomes the bottleneck, so the newest development, in IBM's Power4 chip, is to put two completely separate processors onto the same chip. This may well be the direction for future chip developments.

One very noteworthy change over the last 10 years has been the consolidation in the industry. With Compaq announcing the end of Alpha development, there are now just four main companies developing CPUs: Intel, AMD, IBM and Sun Microsystems. Intel is clearly the dominant force with the lion's share of the market. It must be tough for the others to sustain the very high R&D costs necessary for future chip development, so further reduction in this list seems a distinct possibility.

Another change which may become important for scientific computing is the growth in the market for mobile computing (laptops and personal data assistants [PDAs]) and embedded computing (e.g. control systems in cars) both of which have driven the development of low-cost low-power microprocessors, which now are not very much slower than the regular CPUs.

Memory. As CPU speed has increased, applications and the data they use have grown in size too. The price of memory has varied erratically, but main memory sizes have probably doubled every 18 months in line with processor speed. However, the speed of main memory has not kept pace with processor speeds, so that data throughput from main memory to processor has become probably the

most significant bottleneck in system design. Consequently, we now have systems with a very elaborate hierarchy of caches. All modern chips have at least two levels of cache, one on the CPU chip, and the other on a separate chip, while the new IBM Power4 has three levels. This introduces a lot of additional complexity into the system design, but the user is shielded from this.

Hard disks. Disk technology has also progressed rapidly, in both size and reliability. One of the most significant advances has been the RAID (redundant array of inexpensive disks) approach to providing very large and reliable file systems. By ‘striping’ data across multiple disks and reading/writing in parallel across these disks it has also been possible to greatly increase aggregate disk read/write speeds. Unfortunately, backup tape speeds have not improved in line with the rapid increase in disk sizes, and this is now a significant problem.

System interconnect. Connecting the different components within a computer is now one of the central challenges in computer design. The general trend here is a change from system buses to crossbar switches to provide sufficient data bandwidth between the different elements. The chips for the crossbar switching are themselves now becoming commodity components.

Networking. In the last 10 years, networking performance, for example for file servers, has improved by a factor of 100, from Ethernet (10 Mb/s) to Gigabit Ethernet (1 Gb/s), and 10 Gb/s Ethernet is now under development. This has been driven by the development of the Internet, the World Wide Web and multimedia applications. It seems likely that this development will continue, driven by the same forces, perhaps with increasing emphasis on tight integration with the CPU to maximize throughput and minimise delays. These developments would greatly aid distributed-memory parallel computing for scientific purposes.

Very high performance networking for personal computer (PC) clusters and other forms of distributed-memory machine remains the one area of custom hardware development for scientific computing. The emphasis here of companies such as Myricom and Dolphin Interconnect is on very low latency hardware, minimizing the delays in sending packets of data between machines. These companies currently manufacture proprietary devices, but the trend is towards adoption of the new Infiniband standard which will lead to the development of low-cost very high performance networking for such clusters, driven in part by the requirements of the ASPs (application service providers), to be described later.

Visualization. 10 years ago, scientific visualization required very specialized visualization workstations. Today, there is still a small niche market for specialized capabilities such as ‘immersive technologies’, but in the more

conventional areas of scientific visualization the situation has changed enormously with the development of very low cost but incredibly powerful 3D graphics cards for the computer games marketplace.

Systems

Vector computers. The days of vector computing are over. The huge development costs could not be recouped from the very small scientific supercomputing marketplace. No new codes should be written with the aim of executing them on such systems.

Shared-memory multiprocessors. Shared-memory systems have a single very large memory to which is connected a number of processors. There is a single operating system, and each application task is usually a single Unix 'process'. The parallelism comes from the use of multiple execution 'threads' within that process. All threads have access to all of the data associated with the process. All that the programmer has to worry about to achieve correct parallel execution is that no two threads try to work with, and in particular update, the same data at the same time.

This simplicity for the programmer is achieved at a high cost. The problem is that each processor has its own cache, and in many cases the cache will have a more up-to-date value for the data than the main memory. If another processor wants to use that data, then it needs to be told that the cache has the true value, not the main memory. In small shared-memory systems, this problem of cache coherency is dealt with through something called a 'snoopy bus', in which each processor 'snoops' on requests by others for data from the main memory, and responds if its cache has a later value. In larger shared-memory systems, the same problem is dealt with through specialized distributed cache management hardware.

This adds significantly to the cost of the system interconnect and memory subsystems. Typically, such systems cost three-to-five times as much as distributed memory systems of comparable computing power. Furthermore, the benefits of shared-memory programming can be illusional. To get really good performance on a very large shared-memory system requires the programmer to ensure that most data is used by only one processor, so that it stays within the cache of that processor as much as possible. This ends up pushing the programmer towards the style of programming necessary for distributed-memory systems.

Shared-memory multiprocessors from SGI and Sun Microsystems account for approximately 30% of the machines in the TOP500 list of the leading 500 supercomputers in the world which are prepared to provide details of their systems. The SGI machines tend to be used for scientific computing, and the Sun systems for financial and database applications, reflecting the different marketing emphasis of the two companies.

An interesting development is that the major database companies, such as Oracle, now have distributed-memory versions of their software. As a consequence of this, and the cost of large shared-memory systems, my prediction is that the market demand for very large shared-memory systems will decline. On the other hand, I expect that there will continue to a very large demand for shared-memory machines with up to 16 processors for commercial computing and applications such as web servers, file servers, etc.

Distributed-memory systems. Distributed-memory systems are essentially a number of separate computers coupled together by a very high speed interconnect. Each individual computer, or 'node', has its own memory and operating system. User's applications have to decide how to split the data between the different nodes. Each node then works on its own data, and they communicate with each other as necessary when the data belonging to one is needed by another. In the simplest case, each individual node is a single processor computer, but in more complex cases, each node may itself be a shared-memory multiprocessor.

IBM is the manufacturer of approximately 40% of the systems on the TOP500 list, and almost all of these are distributed-memory systems. Many are based on its SP architecture which uses a cross-bar interconnect. This includes the system known as ASCI White which is officially the world's fastest computer at present, at least of those which are publicly disclosed.

Another very important class of distributed-memory systems are Linux PC clusters, which are sometimes also known as Beowulf clusters. Each node of these is usually a PC with one or two Intel processors running the Linux operating system. The interconnect is usually Myricom's high-speed low-latency Myrinet 2000 network, whose cost is approximately half that of the PC itself. These systems provide the best price/performance ratio for high-end scientific applications, which demand tightly-coupled distributed-memory systems. The growth in these systems has been very dramatic in the past two years, and there are now many such systems with at least 128 processors, and a number with as many as 1024 processors. This includes the ASCI Red computer with 9632 Pentium II processors, which was the world's fastest computer when it was installed in 1999, and is still the world's third fastest.

Looking to the future, I think this class of machines will become the dominant force in scientific computing, with Infiniband networking and with each node being itself a shared-memory multiprocessor, possibly with the multiple processors all on the same physical chip.

Workstation/PC farms. Workstation and PC farms are similar to distributed-memory systems but connected by a standard low-cost Fast Ethernet network. They are ideally suited for 'trivially parallel' applications which involve very large

numbers of independent tasks, each of which can be performed on a single computer. As with PC clusters, there has been very rapid development in this area. The big driving force now is to maximize the ‘density’ of such systems, building systems with as much computing power as possible within a given volume of rack space. It is this desire to minimize the space requirements that is leading to the increasing use of low-power mobile processors. These consume very little power and so generate very little heat to be dissipated and can therefore be packaged together very tightly. A single computer rack with 128 processors seems likely in the very near future, so larger systems with 1024 processors could become common in a few years.

Software developments

Operating systems

Unix remains the dominant choice for scientific computing, although Windows dominance in everyday computing means it cannot be discounted.

Within the Unix camp, the emergence and acceptance of Linux is the big story of the last 10 years, with many proprietary flavours of Unix disappearing.

The big issue for the next 10 years will be the management of very large numbers of PCs or workstations, including very large PC clusters. The cost of support staff is becoming a very significant component of overall computing costs, so there are enormous benefits to be obtained from system management tools that enable support staff to look after, and upgrade, large numbers of machines.

Another key technology is DRM (Distributed Resource Management) software such as Sun Microsystems’ Grid Engine software, or Platform Computing’s LSF software. These provide distributed queuing systems which manage very large numbers of machines, transparently assigning tasks to be executed on idle systems, as appropriate to the requirements of the job and the details of the system resources.

Programming languages

Computer languages evolve much more slowly than computer hardware. Many people still use Fortran 77/90, but increasingly C and C++ are the dominant choice for scientific computing, although higher-level, more application-specific languages such as MATLAB are used heavily in certain areas.

OpenMP

For shared-memory computing, OpenMP is the well-established standard with support for both Fortran and C. The development of this standard five years ago

has made it possible for code developers to write a single code which can run on any major shared-memory system, without the extensive code porting effort that was previously required.

MPI

For distributed-memory computing, the standard is MPI (message passing interface) which has superseded the earlier PVM (parallel virtual machine). Again this standard includes library support for both Fortran and C, and it has been adopted by all major system manufacturers, enabling software developers to write fully portable code.

It remains the case unfortunately that the writing of a message-passing parallel code can be a tedious task. It is usually clear enough how one should parallelize a given algorithm, but the task of actually writing the code is still much harder than writing an OpenMP shared-memory code. I wish I could be hopeful about improvements in this area over the next 10 years, but I am not optimistic; there is only limited research and development in this area within academia or by commercial software vendors.

Grid computing

‘Grid computing’ is a relatively new development which began in the USA and is now spreading to Europe; within the UK it is known as ‘E-Science’. The central idea is collaborative working between groups at multiple sites, using distributed computing and/or distributed data.

One of the driving examples is in particle physics, in which new experiments at CERN and elsewhere are generating vast quantities of data to be worked on by researchers in universities around the world.

An entirely different example application is in engineering design, in which a number of different companies working jointly on the development of a single complex engineering product, such as an aircraft, need to combine their separate analysis capabilities with links into a joint design database.

In the simulation of biological processes, there is also probably a strong need for collaboration between leading research groups around the world. Each may have expert knowledge in one or more aspects, but it is by combining their knowledge that the greatest progress can be achieved.

Another aspect of grid computing is remote access to, and control of, very expensive experimental facilities. One example is astronomical telescopes; another is transmission electron microscopes. This may have relevance to the use of robotic equipment for drug discovery.

Other trends

ASPs and remote facility management

It was mentioned earlier that the cost of computing support staff is a significant component of overall computing costs. As a consequence, there is a strong trend to 'outsource' this. Within companies, this can mean an organization such as EDS, CSC or IBM Global Services managing the company's computing systems. Within universities, as well as companies, this may in the future lead to specialist companies remotely managing special facilities such as very large PC clusters. This is made feasible by the advances in networking. The economic benefits come from the economies of scale from having a team of people with specialist knowledge supporting many such systems at different sites.

Another variation on the same theme is ASPs (application service providers) which offer a remote computing service to customers, managing the systems at their own site. This requires much higher bandwidth between the customer and the ASP, so it does not seem so well suited for scientific computing, but it is a rapidly developing area for business computing.

Development of large software packages

My final comments concern the process of developing scientific software. The codes involved in simulation software are becoming larger and larger. In engineering, they range from 50 000 lines to perhaps 2 000 000 lines of code, with development teams of 5–50 people. I suspect the same is true for many other areas of science, including biological simulations.

Managing such extensive software development requires very able programmers with good software engineering skills. However, academic researchers are more focused on the scientific goals of the research, and academic salaries are not attractive to talented information technology (IT) staff. In the long term, I think the trend must be for much of the software development to be done in private companies, but for mechanisms to exist whereby university groups can contribute to the scientific content of these packages.

I do not underestimate the difficulty in this. Joint software development by multiple teams increases the complexity significantly, and developing software so that one group can work on one part without extensive knowledge of the whole code is not as easy as it may appear. Equally, the non-technical challenges in agreeing intellectual property rights provisions, properly crediting people for their academic contributions, etc., are not insignificant. However, I think it is unavoidable that things must move in this direction. Otherwise, I do not see how university groups will be able to take part in the development of extremely large and complex simulation systems.

Reference webpages

http://www.intel.com/home/pentium4/tech_info.htm
<http://www.intel.com/home/pentiumiii-m/index.htm>
<http://www.ibm.com/servers/eserver/pseries/hardware/whitepapers/power4.html>
<http://www.sun.com/sparc/UltraSPARC-III/>
<http://www.myricom.com>
<http://www.dolphinics.com>
<http://www.infinibandta.org/home.php3>
<http://www.intel.com/technology/infiniband/>
<http://www.top500.org>
<http://clusters.top500.org>
<http://www.ibm.com/servers/eserver/pseries/hardware/largescale/supercomputers/asciwhite/>
<http://www.ibm.com/servers/eserver/pseries/solutions/stc/brief.html>
<http://www.openmp.org>
<http://www.unix.mcs.anl.gov/mpi/index.html>
<http://www.globus.org>
<http://www.gridforum.org>
<http://www.ipg.nasa.gov>
<http://www.research-councils.ac.uk/escience/>

DISCUSSION

Asburner: You were fairly optimistic that Moores's law (a doubling of CPU power every 18 months) would continue to hold, at least over the next 18 months. The trouble in my field is that the amount data, even its most simple form, is quadrupling or so every 12 months. If one includes data such as those from microarray experiments, this is probably an underestimate of the rate of growth. We are therefore outstripping Moore's law by a very significant factor.

Paterson: I think the problem is even worse than this. As we start building this class of large models, there are never enough data to give us definitive answers as to how these systems are working. We are always in the process of formulating different hypotheses of what might be going on in these systems. Fundamentally, in systems biology/integrative physiology we have to deal with combinatorics. We may be seeing a geometric growth in computing power, but I would argue that the permutations of component hypotheses within an integrated physiological system also grow geometrically with the size of the biological system being studied. The two tend to cancel each other out, leaving a linear trend in time for the size of models that can be appropriately analysed.

Asburner: If I wanted today to do an all-against-all comparison of two human genome sequences, I don't know whether I'd see this through before I retire. This is desperately needed. We only have one human sequence (in fact, we have two, but one is secret), but five years down the line we will have about 50.

Reinhardt: In genomics, we commonly face the problem of simultaneously having to analyse hundreds of microarray experiments, for example for the

prediction of protein interactions from expression data. The rate-limiting step is getting the data out of the database. The calculation time is only a percentage of the whole run-time of the program. For algorithms we can add processors and we can parallelize procedures. What we don't have is a solution for how to accelerate data structures. This is needed for faster retrieval of data from a data management system. This would help us a lot.

Subramaniam: I agree with you that bang for the buck is very good with the distributed computing processors. But your statement that databases such as Oracle deal efficiently with data distributed computing is not true. We deal with this on a day-to-day basis. If you are talking about particle physics where you have 4–10 tables this may be true, but in cell biology we are dealing typically with 120 tables. We don't have the tools at the moment to do data grid computing and feeding back to a database.

Biology computing is qualitatively distinct from physics equation space computing. First, a lot of data go into the computing process. Second, we don't have idealized spheres and cylinders: there are very complex geometries, and the boundary conditions are very complicated. Third, biologists think visually most of the time. They need visualization tools, which rules out Fortran, because it is not possible to write a sphere program in Fortran very easily. This is one of the reasons why Mike Pique wrote his first sphere program in C++. I am just trying to point out that graphical user interfaces (GUIs) are an integral component of biology. GUIs warrant programming in Java and Perl and so on.

Giles: It is possible to combine different languages, although all the visualization software we use is written in C. Yes, visualization is crucial. But visualization software exists for handling distributed memory data. I have no idea how efficient Oracle's distributed databases are, but what is important is that this is the way they are heading. This is the platform that they see as becoming the dominant one. If they haven't got it right with their first release, by the time they get to their tenth release they will surely have got it right.

Noble: Mike Giles, since you have started to use your crystal ball, I'd like you to go a little further with it. I will show a computation in my paper which I did in 1960, on a machine that occupied a fairly large room. It was an old valve machine and the computation took about two hours. When I show it, it will flash across my rather small laptop so fast that I'll have to slow it down by a factor of about 50 in order to let you see what happens. Jump 40 years in the future: where is the limit in processing power? You were looking ahead no further than 5–10 years. Were you constraining yourself because you can see some obvious physical limits, or was this because you thought that speculating any further ahead is not possible?

Giles: It gets really tough to look beyond about 10 years. People have been talking about reaching physical limits of computing for a while, but manufacturing technology keeps advancing. Currently, these chips are generated

with photolithography, laying down patterns of light that etch in and define the pathways. Now we are at the level where the feature of individual gates is actually less than the wavelength of UV light. This is achieved by the use of interference patterns. The next stage will involve going to X-rays, using synchrotrons. There's enough money in the marketplace to make this feasible. In the labs they are already running 4 GHz chips. Power is definitely becoming a concern. Very large systems consume a great deal of electricity, and dissipating this heat is a problem. Then answer is probably to move in the direction of low-voltage chips.

Noble: So at the moment you feel we can just keep moving on?

Giles: Yes. For planning purposes Moore's law is as good as anything.

Asburner: In genomics we are outpacing this by a significant factor, as I have already said.

Giles: Well, I can't see any hope for beating Moore's law. My other comment is that any effort spent on improving algorithms that will enable us to do something twice as fast, is a gain for all time. The funding agencies are trying to get biologists and computer scientists to talk to each other more on algorithm issues.

Noble: This brings us round to the software issue, including languages.

Loew: You mentioned e-computing, or grid computing, and how this might relate to modelling. The results of modelling really require a different publication method than the traditional 'flat' publications that we are used to. Even the current electronic publications are still quite flat. Collaborative computing seems to be the ideal sort of technology to bring to bear on this issue of how to publish models. It is an exciting area. Is there any effort being made to deal with this in the computer science community?

Noble: There is in the biological science community: there is very serious discussion going on with a number of journals on this question.

Loew: I've been involved a little with the *Biophysical Journal* on the issue, but we are still trying to get the journal to move beyond including movies. It's a hard sell.

Levin: There are publishing companies at this stage who are looking quite proactively at establishing web-based publishing of interactive models. One of the issues bedevilling all of them is standardization. Even in the scientific community we haven't yet adopted compatible standards for developing models. Once we have reached consensus in the community as to what are the right standardization forms, it will be much easier for the publishers to adopt one or the other. Wiley has made steps towards doing this in the lab notebook area. But these aren't sophisticated enough to accommodate complex models. What we are talking about here is the ability to put onto the web a model, and then for an individual investigator to place their own separate data into that model and run the model. This is currently feasible and can be made practical. It is a question of deciding which of the standards to adopt. It is likely to be based on being able to use a form of XML (extensible mark-up language) as a standard.

I have one other point that concerns the educational issue. Modelling has been the preserve of just a few ‘kings’ over the years: in order for it to devolve down to the pawns and penetrate across the entire spectrum of biology, I think it will take a number of proactive efforts, including publication of interactive models on the web; the development of simple tools for modelling; and the use of these tools not only in companies but also in places of education, to answer both applied and research biological questions.

Noble: The publication issue has become a very serious one in the UK. I remember when the *Journal of Physiology* switched to double-column publication, instead of the old-fashioned across-the-page style. The big issue was whether it would ever be possible again to publish a paper like the Hodgkin–Huxley paper on the nerve impulse! Much more seriously, journals that were taking a very good view of the extensive article covering some 30–40 pages are no longer doing so. *The Proceedings of the Royal Society* has gone over to short paper publication. *Philosophical Transactions of the Royal Society*, which was the journal that no one buys but everyone relies on, no longer takes original papers, though it is noteworthy that the Royal Society journals do a good job on publishing extensive focused issues. *Progress in Biophysics and Molecular Biology* does this also. These were the places where people were gravitating towards in order to publish huge papers.

Hunter: This is not the case in some areas of engineering and mathematics. SIAM (Society for Industrial and Applied Mathematics) publishes very long, detailed mathematical papers.

Loew: There’s another issue related to this that I still think has to do with collaborative computing and databasing. Once you start including this kind of interactive modelling environment in electronic publications, how is it archived so that people can look for pieces of models, in order to get the much richer kind of information, as opposed to the rather flat information that we now get through PubMed or Medline? I think there are a great number of possibilities for really enriching our ability to use published material in ways that just haven’t been possible before.

Subramaniam: With regard to databases, one of the missing elements here is ontologies. Establishing well-defined ontologies will be needed before we will have databases that can be distributed widely.

Asburner: These are complex issues, many of which are not scientific but social. Philip Campbell, the editor of *Nature*, recently wrote in answer to a correspondent stating that supplementary information attached to papers published by *Nature* is archived by the journal in perpetuity. If I take that statement literally, then Philip Campbell must know something I don’t! There is clearly an inherent danger here, because we all know that any commercial company can be taken over and the new owners may not have the same commitment: no guarantees on earth will lead me to

believe that the new owners will necessarily respect the promises of the original owner. This is not a scientific problem but a social one.

Cassman: There are answers to this that have nothing to do with publication or journals. There are nationally supported databases, such as the protein database. This is the natural place for these models to be located. The difficulty is, as Shankar Subramaniam has pointed out, that for databases of interacting systems we lack ontologies that people will agree on. Ontologies exist, but there is no common ontology. Attempts that we made to get people to agree on this issue a couple of years ago simply failed. I don't know what the answer is. It's a critical issue: if these models are to be more than descriptive, then they have to be easily accessible in ways that are common (or at least interconvertible) for all of the models. This hasn't happened yet, but it needs to happen reasonably quickly. Molecular genetics, when it started, was a very specific discipline used by a small number of people. Now everyone uses it: it is a standard tool for biology. If we want modelling to be a standard tool also (as it should be) then we need all these things to happen, and some group is going to have to catalyse it.

Berridge: When it comes to archiving array data, for example, do we have to draw a distinction between databases that store sequence information, and those in which we store experimental data? If you are running an array experiment comparing cells A and B, and the result of that experiment is that there are 20 new genes being expressed in cell B, do we have to keep the original data? Does someone accessing this paper have to be able to interrogate the original data? And is there a cut-off where there is so much information, that we just need to keep the information that was extracted from the experiment rather than archiving all the array data? I suspect this is a balance that we will have to strike.

Asburner: Access to these data is essential for the interpretation of these sorts of experiment: they must be publicly available.

Berridge: There must be a balance somewhere, because it simply won't be physically possible to store every bit of data.

Asburner: The particular issue of microarray data is whether or not the primary images should be stored. I believe they should, but the problem is that they are very large. Although memory is relatively cheap, it is still a major problem. Moreover, even if the images are stored, to attempt to transmit them across the internet would require massive bandwidth and is probably not currently feasible.

McCulloch: There is an emerging standard for microarray data representation, called MAGEML. This includes the raw image as part of the XML document, in the form of a URL (uniform resource locator) that points to the image file. At least in principle these databases are readily federated and distributed. But then, the likelihood of being able to retrieve and efficiently query the image is not great, especially after a long period. The consensus though is that the raw experimental

data should be available. At least in part, this is driven by the significant differences in the way people interpret them.

Asburner: And the software will improve, too. We may want to go back and get more out of the original data in the future.

Levin: I would like to address the issue of storage of data, referring in particular to modelling. There is a need to institutionalize memory. Without institutionalizing memory, whether it be in an academic or commercial organization, what happens is that frequently we are forced to recreate the errors of the past by redoing the experiments again. The cost is unsupportable, particularly as the number of hypotheses that are being generated by data such as microarray data rises. With the primary data we need to come to a consensus as to how we store them and what we store. There is an essential requirement to be able to store models that contain within them a hypothesis and the data upon which the hypothesis was based. So, when a researcher leaves a laboratory, the laboratory retains that particular body of work in the form of a database of models. This enables other researchers to go back and query, without having to recreate the data and the models.

Boissel: The raw data are already stored somewhere, and the real problem is just one of accessing these data. Of course, the raw data alone, even if they are accessible, are difficult to use without proper annotation. I don't think we need a huge database containing everything. We just need to have proper access to existing data. This access will be aided if there is proper ontology, such that each researcher can store their own raw data in the proper way.

I have a feeling that we are discussing two separate issues here: the storage of data and the access to data, and the storage of models and the access to models.

McCulloch: The discussion started out about hardware and software, and has quickly gravitated towards data, which is not surprising in a biological setting. It is the large body of data, and how to get at this and query it, that is the central driving force of modern computational biology. But let's confine the discussion for a minute to that set of information that comprises the model, and that people have discussed encapsulating in formats such as CellML or SBML (systems biology mark-up language). It will be helpful to the 'kings' (the modellers), but it will not in itself make the models available to other biologists without appropriate software tools. Mike Giles, I'd like to comment on the issue you raised about software. First, you said you looked into C++ about 10 years ago and found that it wasn't stable. There are now excellent C++ compilers, so stability of this language is no longer a problem. But there is, at the moment, a perceived problem with object-oriented languages such as C++ and Java for scientific programming, and that is performance. We found, somewhat to our surprise, that C++ has features that can more than compensate for the performance trade-offs of modularity and flexibility. For example, using templated meta-programming facilities of C++ we

achieved speed-ups of over 10-fold compared with legacy FORTRAN code. These generic programming techniques allow the programmer to optimize the executable code at compile time by identifying data that won't change during execution. The idea that modern object-oriented languages must sacrifice performance needs to be revised because sometimes they can actually improve it.

Paterson: There is one point that hasn't been addressed yet that I think is relevant here. In terms of getting a common language and being able to get a model published, this is moving the bottleneck from an issue of portability to an issue of scientifically sound usage now that the model is in the hands of a much larger group of people. I would be interested to understand to what extent ontologies have been able to solve the problem that I think is going to arise. Models are an exercise in abstracting reality. They aren't reality. The amount of documentation it takes to make a model stand alone—explaining how to use, interpret and modify it—is going to be an issue. My concern is that the bottleneck is going to come back to the researcher. Now that everyone has the model and is able to start doing things with it, this is likely to create a huge support/documentation burden on the person publishing the model. Either they will have to manage the flood of 'support' questions, or worse, anticipate limits and caveats to using the model in unanticipated applications and document the model accordingly.

Noble: This is one of the reasons why in the end we had to commercialize our models. The reason my group launched Oxsoft is that we published a paper (Di Francesco & Noble 1985) and couldn't cope with the response. It wasn't just the 500–1000 reprint requests, but also about 100 requests for the software. There simply wasn't any other way of coping with that demand. Those were the days when a disk cost £100.

Levin: You have stimulated a thought: effectively all biologists are modellers in one fashion or another, we just don't interpret the way we conduct science in this way. A person who has drawn a pathway on a piece of paper showing potential proteins and how they interact has modelled in one dimension what the relationships are. I think the challenge is less being concerned with researchers and their use of models, or their ability to refer back to the original formation and documentation of the model (although these are important). Rather, the obligation resides on those who are building the software and the underlying mathematics (including the ontologies and standardization) to ensure that the end-user finds the modelling tools sufficiently intuitive to utilize it in the same way that other standardized biological tools, such as PCR, gained acceptance once the basic technology (in the case of PCR, the thermal cycler) was simple enough to be used universally. The onus is on those responsible for building intuitive, practical and functional capabilities into the technologies and making them available for modelling.

Asbburner: Denis Noble, I'm sure you are correct that at the time commercialization was the only way to cope. But now there exist robust systems by means of which you can deposit your software and make it accessible to others (for example, on the Sourceforge website; <http://sourceforge.net>). I agree that it has to be well documented, but it is then freely available for anyone to download. The lesson of Linux has to be taken seriously by the biological community. We are working entirely through open-source sites and with open-source software. There is no distribution problem.

Subramaniam: In addition to ontology, in order to make models universally accessible we need to create problem-solving environments such as the Biology Workbench.

Reference

Di Francesco D, Noble D 1985 A model of cardiac electrical activity incorporating ionic pumps and concentration changes. *Philos Trans R Soc Lond B Biol Sci* 307:353–398

From physics to phenomenology. Levels of description and levels of selection

David Krakauer

Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

Abstract. Formal models in biology are traditionally of two types: simulation models in which individual components are described in detail with extensive empirical support for parameters, and phenomenological models, in which collective behaviour is described in the hope of identifying critical variables and parameters. The advantage of simulation is greater realism but at a cost of limited tractability, whereas the advantage of phenomenological models, is greater tractability and insight but at a cost of reduced predictive power. Simulation models and phenomenological models lie on a continuum, with phenomenological models being a limiting case of simulation models. I survey these two levels of model description in genetics, molecular biology, immunology and ecology. I suggest that evolutionary considerations of the levels of selection provides an important justification for many phenomenological models. In effect, evolution reduces the dimension of biological systems by promoting common paths towards increased fitness.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 42-52

... In that Empire, the art of cartography attained such perfection that the map of a single province occupied the entirety of a city, and the map of the Empire, the entirety of a province. In time those unconscionable maps no longer satisfied, and the Cartographers Guilds struck a map of the Empire whose size was that of the Empire and which coincided point for point with it.

Jorge Luis Borges *On Exactitude in Science*

Levels of description

The natural sciences are all concerned with many-body problems. These are problems in which an aggregate system is made up from large numbers of a few

basic types or particles, and where these types interact according to some well-defined rules. The state of a system at any one time can be captured by a microscopic description of the individual particles, a macroscopic description of system level properties such as entropy and temperature, or by statistical descriptions of the whole system. Both microscopic and macroscopic descriptions relate to some constituent parts of a system, whereas the statistical description deals exclusively with aggregate properties. In the language of biology the microscopic and macroscopic descriptions would be referred to as mechanistic whereas the statistical property as a functional character. Formal mathematical or computational approaches in the natural sciences reflect these different scales in the choice of either individual-based simulations, population-based phenomenological models, or statistical models. The natural scale and model choice do not map one to one. The most abstract description—the statistical model—can be applied only to the whole system, whereas the most specific description—the microscopic model—can be used to generate results at all three scales. Consider the combinatorial game of ‘Go’. A microscopic model would describe the colour and position of every piece on the 19×19 Go board and deduce the future state of the board through the application of a strategy operating within certain basic constraints or game-rules. A macroscopic model could describe configurations of pieces forming triangles, chains, ladders and eyes, and calculate likely transitions between these configurations through the application of pattern-based strategies such as cutting, connecting and Ko.

A statistical model could describe the mean numbers of each colour, the mean territory size, and the temporal variance in score as a means of estimating probable outcomes. The statistical model is of little use to a player of the game, whereas the microscopic model will produce both macroscopic patterns and improve parameter estimates for the statistical model.

An important limitation of the simulation-based approach is that the possible states of the model are of the same order as the possible states of the game. The simulation model is not significantly simpler than the system it describes. This exposes an apparent paradox of simulation models, namely, is the natural system the best simulation of itself (see Borges’ epigraph)? Both the phenomenological model and the statistical model are considerably simpler than the natural system.

The natural system can exist in many more configurations, and produce many more behaviours, than the phenomenological model or the statistical model could describe. The key to worthwhile phenomenological modelling is in the choice of restrictions we apply in developing our macroscopic descriptions, whereas the key to simulation-based modelling is speed. It is because we require that our simulations reproduce possible system configurations over a shorter time than real time, that the structure of the simulation model and the natural system must be different. The way in which this increase in speed is achieved is to

make simulation models partially phenomenological through simplifying approximations. We see therefore that the distinction between modelling approaches becomes somewhat arbitrary, as all models are phenomenological models. The differences are not qualitative but quantitative, and relate to the number of variables and parameters we are happy to plug into our brains or into the circuitry of a computer. A smaller number of variables and parameters is always preferable, but our willingness to move toward the phenomenological, depends on how reliable is the derivation of the macroscopic equations from the microscopic interactions. A formal approach to rescaling many-body problems—a method for reducing the number of variables—is to use renormalization group theory (Wilson 1979).

Here I am going to present an evolutionary perspective on this complex topic. Rather than discuss Monte Carlo methods, agent based models, interacting particle systems, and stochastic and deterministic models, and their uses at each scale. I restrict myself to a biological justification for phenomenological modelling. The argument is as follows. Natural selection works through the differential replication of individuals. Individuals are complex aggregates and yet the fitness of individuals is a scalar quantity, not a vector of component fitness contributions. This implies that the design of each component of an aggregate must be realized through the differential replication of the aggregate as a whole. We are entitled therefore to characterize the aggregate with a single variable, fitness, rather than enumerate variables for all of its components. This amounts to stating that identifying levels of selection can be an effective procedure for reducing the dimensionality of our state space.

Levels of selection

Here I shall briefly summarize current thinking on the topic of units and levels of selection (for useful reviews see Keller 1999, Williams 1995). The levels of selection are those units of information (whether genes, genetic networks, genomes, individuals, families, populations, societies) able to be propagated with reasonable fidelity across multiple generations, and in which these units, possess level-specific fitness enhancing or fitness reducing properties. All of the listed levels are in principle capable of meeting this requirement (that is the total genetic information contained within these levels), and hence all can be levels of selection. Selection operates at multiple levels at once. However, selection is more efficient in large populations, and drift dominates selection in small populations. As we move towards increasingly inclusive organizations, we also move towards smaller population sizes. This implies that selection is likely to be more effective at the genetic level than, say, the family level. Furthermore, larger organizations are more likely to undergo fission, thereby reducing the fidelity of replication. These

two factors have led evolutionary biologists to stress the gene as a unit of selection. This is a quantitative approximation. In reality there are numerous higher-order fitness terms derived from selection at more inclusive scales of organization.

From the foregoing explanation it should be apparent that the ease with which a component can be an independent replicator, helps determine the efficiency of selection. In asexual haploid organisms individual genes are locked into permanent linkage groups. Thus individual genes do not replicate, rather whole genomes or organisms. The fact of having many more genes than genomes is not an important consideration for selection. This is an extreme example highlighting the important principle of linkage disequilibrium. Linkage disequilibrium describes a higher than random association among alleles in a population. In other words, picking an AB genome from an asexual population is more likely than first picking an A allele and subsequently a B allele. Whenever A and B are both required for some function we expect them to be found together, regardless of whether the organism is sexual, asexual, or even if the alleles are in different individuals! (Consider obligate symbiotic relationships.) This implies that the AB aggregate can now itself become a unit of selection. This process can be extended to include potentially any number of alleles, spanning all levels of organization. The important property of AB versus A and B independently is that we can now describe the system with one variable whereas before we had to use two. The challenge for evolutionary theory is to identify selective linkage groups, thereby exposing units of function, and allowing for a reduction in the dimension of the state space. These units of function can be genetic networks, signal transduction modules, major histocompatibility complexes, and even species. In the remainder of this paper I shall describe individual level models and their phenomenological approximations, motivated by the assumption of higher levels of selection.

Levels of description in genetics

Population genetics is the study of the genetic composition of populations. The emphasis of population genetics has been placed on the changes in allele frequencies through time, and the forces preserving or eliminating genetic variability. Very approximately, mutation tends to diversify populations, whereas selection tends to homogenize populations. Population genetics is a canonical many-body discipline. It would appear that we are required to track the abundance of every allele at each locus of all members in a randomly mating population. This would seem to be required assuming genes are the units of selection, and all replicate increasing their individual representation in the gene pool. However, even a cursory examination of the population genetics literature reveals this expectation to be unjustified. The standard assumption of population genetics modelling is that whole genotypes can be assigned individual fitness

values. Consider a diploid population with two alleles, A_1 and A_2 and corresponding fitness values $W_{11}=1$, $W_{12}=W_{21}=1-bs$ and $W_{22}=1-s$. The value s is the selection coefficient and b the degree of dominance. Population genetics aims to capture microscopic interactions among gene products by varying the value of b . When $b=1$ then A_1 is dominant. When $0 < b < \frac{1}{2}$ then A_1 is incompletely dominant. When $b=0$, A_1 is recessive. Denoting as p the frequency of A_1 and $1-p$ the frequency of A_2 , the mean population fitness is given by

$$\overline{W} = 1 - s + 2s(1 - b)p - s(1 - 2b)p^2$$

and the equilibrium abundance of A_1 ,

$$\hat{p} = \frac{1 - b}{1 - 2b}$$

These are very general expressions conveying information about the fitness and composition of a genetic population at equilibrium. The system is reduced from two dimensions to one dimension by assuming that dominance relations among autosomal alleles can be captured through a single parameter (b). More significantly, the models assume that autosomal alleles are incapable of independent replication. The only way in which an allele can increase its fitness is through some form of cooperation (expressed through the dominance relation) with another allele.

The situation is somewhat more complex in two-allele two-locus models (A_1, A_2, B_1, B_2). In this case we have 16 possible genotypes. The state space can be reduced by assuming that there is no effect of position, such that the fitness of $A_1B_1A_2B_2$ is equal to that of $A_1B_2A_2B_1$. We therefore have 9 possible genotypes. We can keep the number of parameters in such a model below 9 while preventing our system from becoming underdetermined, by assuming that genotype fitness is the result of the additive or multiplicative fitness contributions of individual alleles. This leaves us with us 6 free parameters. The assumption of additive allelic fitness means that individual alleles can be knocked out without mortality of the genotype. With multiplicative fitness knockout of any one allele in a genome is lethal. These two phenomenological assumptions relate to very different molecular or microscopic processes. Once again this modelling approach assumes that individual alleles cannot increase their fitness by going solo; alleles increase in frequency only as members of the complete genome and they cooperate to increase mean fitness.

When alleles or larger units of DNA (microsatellites, chromosomes) no longer cooperate, that is when they behave selfishly, then the standard population genetics approximations for the genetic composition of populations breaks down (Buss

1987). This requires that individual genetic elements rather than whole genotypes are assigned fitness values. The consequence is a large increase in the state space of the models.

Levels of description in ecology

Population genetics was described as the study of the genetic structure of populations. In a like fashion, ecology might be described as the study of the species composition of populations. More broadly, ecology seeks to study the interactions between organisms and their environments. This might lead one to expect that theory in ecology is largely microscopic, involving extensive simulation of large populations of different individuals. Once again this is not the case. The most common variable in ecological models is the species. In order to understand the species composition of populations, theoretical ecologists ascribe replication rates and birth rates to whole species, and focus on species level relations. We can see this by looking at typical competition equations in ecology. Assume that we have two species X and Y with densities x and y . We assume that these species proliferate at rates ax and dy . In isolation each species experiences density limited growth at rates bx^2 and fy^2 . Finally, each species is able to interfere with the other such that y reduces the growth of x at a rate cyx and x reduces the growth of y at a rate exy . With these assumption we can write down a pair of coupled differential equations describing the dynamics of species change,

$$\begin{aligned}\dot{x} &= x(a - bc - cy) \\ \dot{y} &= y(d - ex - fy)\end{aligned}$$

This system produces one of two solutions, stable coexistence or bistability. When the parameter values satisfy the inequalities,

$$\frac{b}{e} > \frac{a}{d} > \frac{c}{f}$$

The system converges to an equilibrium in which both species coexist. When the parameter values satisfy the inequalities,

$$\frac{c}{f} > \frac{a}{d} > \frac{b}{e}$$

then depending on the initial abundances of the two species one or the other species is eliminated producing bistability. These equations describe infinitely large populations of identical individuals constituting two species. The justification for this approximation is derived from the perfectly reasonable assumption that

evolution at the organismal level is far slower than competition among species. This separation of time scales is captured by Hutchinson's epigram, 'The ecological theatre and the evolutionary play'. In effect these models have made the species the vehicle for selection.

An explicit application of the separation of time scales to facilitate dimension reduction lies at the heart of adaptive dynamics (Diekman & Law 1996). Here the assumption is made to allow individual species composition to be neglected in order to track changes in trait values. The canonical equation for adaptive dynamics is,

$$\dot{s}_i = k_i(s) \cdot \frac{\partial}{\partial s'_i} W_i(s'_i, s)|_{s'_i=s_i}.$$

The s_i with $i=1, \dots, N$ denote the values of an adaptive trait in a population of N species. The $W(s'_i, s)$ are the fitness values of individual species with trait values given by s_2 when confronting the resident trait values s . The $k_i(s)$ values are the species-specific growth rates. The derivative $(\partial/\partial s'_i)W_i(s'_i, s)|_{s'_i=s_i}$ points in the direction of the maximal increase in mutant fitness. The dynamics describes the outcome of mutation which introduces new trait values (s'_i) and selection that determines their fate — fixation or extinction. It is assumed that the rapid time scale of ecological interactions, combined with the principle of mutual exclusion, leads to a quasi-monomorphic resident population. In other words, populations for which the periods of trait coexistence are negligible in relation to the time scale of evolutionary fixation. These assumptions allow for a decoupling of population dynamics (changes in species composition) from adaptive dynamics (changes in trait composition).

While these levels of selection approximations have proved very useful, there are numerous phenomena for which we should like some feeling for the individual behaviours. This requires that we do not assume away individual contributions in order to build models, but model them explicitly, and derive aggregate approximations from the behaviour of the models. This can prove to be very important as the formal representation of individuals, can have a significant impact on the statistical properties of the population. Durrett & Levin (1994) demonstrate this dependence by applying four different modelling strategies to a single problem: mean field approaches (macroscopic), patch models (macroscopic), reaction diffusion equations (macroscopic) and interacting particle systems (microscopic). Thus the models move between deterministic mean field models, to deterministic spatial models, to discrete spatial models. Durrett and Levin conclude that there can be significant differences at the population level as a consequence of the choice of microscopic or macroscopic model. For example spatial and non-spatial models disagree when two species

compete for a single resource. The importance of this study is to act as cautionary remark against the application of levels of selection thinking to justify approximate macroscopic descriptions.

Levels of description in immunology

The fundamental subject of experimental immunology is the study of those mechanisms evolved for the purpose of fighting infection. Theoretical immunology concerns itself with the change in composition of immune cells and parasite populations. Once again we might assume that this involves tracking the densities of all parasite strains and all proliferating antigen receptors. But consideration of the levels of selection can free us from the curse of dimensionality. The key to thinking about the immune system is to recognize that selection is now defined somatically rather than through the germ line. The ability of the immune system to generate variation through mutations, promote heredity through memory cells, and undergo selection through differential amplification, allows us to define an evolutionary process over an ontogenetic time scale. During somatic evolution, we assume that receptor diversity and parasite diversity are sufficiently small to treat the immune response as a 1 dimensional variable. Such an assumption underlies the basic model of virus dynamics (Nowak & May 2000). Denote uninfected cell densities as x , infected cells y , free virus as v and the total cytotoxic T lymphocyte (CTL) density as z . Assuming mass action we can write down the macroscopic differential equations,

$$\dot{x} = \lambda - dx - \beta xv \quad (1)$$

$$\dot{y} = \beta xv - ay - pyz \quad (2)$$

$$\dot{v} = ky - uv \quad (3)$$

$$\dot{z} = cyz - bz \quad (4)$$

The rate of CTL proliferation is assumed to be cyz and the rate of decay of CTLs bz . Uninfected cells are produced at a rate λ , die at a rate λx , and are infected at a rate βxv . Free virus is produced from infected cells at a rate ky and dies at a rate uv . The immune system eliminates infected cells proportional to the density of infected cells and available CTLs pyz . Assuming that the inequality $cy > b$ then CTLs increase to attack infected cells. The point about this model is that individuals are not considered: the population of receptor types, cell types and virus types are all assumed to be monomorphic. As with the ecological theatre and evolutionary play,

we assume rapid proliferation and selection of variants, but much slower production. When these assumptions are unjustified, such as with rapidly evolving RNA viruses, then we require a more microscopic description of our state space. We can write down a full quasi-species model of infection,

$$\dot{x} = \lambda - dx - x \sum_i \beta_i v_i \quad (5)$$

$$\dot{y}_i = x \sum_j \beta_j Q_{ij} v_j - a_i y_i - p y z \quad (6)$$

$$\dot{v}_i = k_i y_i - u_i v_i \quad (7)$$

$$\dot{z} = \sum_j c_j y_j z - b z \quad (8)$$

Here the subscript i denotes individual virus strains and Q_{ij} the probability that replication of virus j results in the production of a virus i . In such a model receptor diversity is ignored, assuming that the immune response is equally effective at killing all virus strains. In other words, receptors are neutral (or selectively equivalent) with respect to antigen. In this way we build increasingly microscopic models of the immune response, increasing biological realism but at a cost of limited analytical tractability.

Levels of description in molecular biology

Unlike population genetics, ecology and immunology, molecular biology does not explicitly concern itself with evolving populations. However, molecular biology describes the composition of the cell, a structure that is the outcome of mutation and selection at the individual level. There are numerous structures within the cell, from proteins, to metabolic pathways through to organelles, which remain highly conserved across distantly related species. In other words, structures that have the appearance of functional modules (Hartwell et al 1999). Rather than modify individual components of these modules to achieve adaptive benefits at the cellular level, one observes that these modules are combined in different ways in different pathways. In other words, selection has opted to combine basic building blocks rather than to modify individual genes. (Noble has stated this as genes becoming physiological prisoners of the larger systems in which they reside.) This gives us some justification for describing the dynamics of populations of modules rather than the much larger population of proteins comprising these modules.

A nice experimental and theoretical example of functional modularity comes from Huang & Ferrell's (1996) study of ultrasensitivity in the mitogen-activated protein kinase (MAPK) cascades. The MAPK cascade involves the phosphorylation of two conserved sites of MAPK. MAPKKK activates MAPKK by phosphorylation, and MAPKK activates MAPK. In this way a wave of activation triggered by ligand binding is propagated from the cell surface towards the nucleus. Writing down the kinetics of this reaction (using the simplifying assumptions of mass action, and mass conservation), Huang and Ferrell observed that the density of activated MAPK varied ultrasensitively with an increase in the concentration of the enzyme (E) responsible for phosphorylating MAPKKK. Formally, the dose-response curve of MAPKKK against E can be described phenomenologically using a Hill equation with a Hill coefficient of between 4 and 5. The function is of the form,

$$MAPKKK^* = \frac{E^m}{E^m + a^m}$$

where $4 < m < 5$. The density of activated MAPKs at each tier of the cascade can be described with a different value of m . With $m=1$ for MAPK, $m=1.7$ for MAPKK and $m=4.9$ for MAPKKK. The function of the pathway for the cell is thought to be the transformation of a graded input at the cell surface into a switch-like behaviour at the nucleus. With this information, added to the conserved nature of these pathways across species, we can approximate pathways with Hill functions rather than large systems of coupled differential equations.

Not all of molecular biology is free from the consideration of evolution over the developmental time scale. As with the immune system, mitochondrial function and replication remains partially autonomous from the expression of nuclear genes and the replication of whole chromosomes. A better way of expressing this is to observe that mitochondrial genes are closer to linkage equilibrium than nuclear genes. This fact allows for individual mitochondria to undergo mutation and selection at a faster rate than genes within the nucleus. Mitochondrial genes can experience selection directly, rather than exclusively through marginal fitness expressed at the organismal level. The molecular biology of cells must contend with a possible rogue element. This requires that we increase the number of dimensions in our models when there is variation in mitochondrial replication rates.

Conclusions

Models of many-body problems vary in the number of bodies they describe. Predictive models often require very many variables and parameters. For these

simulation models, speedy algorithms are at a premium. Phenomenological models provide greater insight, but tend to do less well at prediction. These models have the advantage of being more amenable to analysis. Even predictive, simulation models are not of the same order as the system they describe, and hence they too contain phenomenological approximations. The standard justifications for phenomenological approaches are: (1) limiting case approximations, (2) neutrality of individual variation, (3) the reduction of the state space, (4) ease of analysis, and (5) economy of computational resources. A further justification can be furnished through evolutionary considerations: (6) levels of selection. Understanding the levels of selection helps us to determine when natural selection begins treating a composite system as a single particle. Thus rather than describe the set of all genes, we can describe a single genome. Rather than describe the set of all cellular protein interactions, we can describe the set of all pathways. Rather than describe the set of all individuals in a population, we can describe the set of all competing species. The identification of a level of selection remains however non-trivial. Clues to assist us in this objective include: (1) observing mechanisms that restrict replication opportunities, (2) identifying tightly coupled dependencies in chemical reactions, (3) observing low genetic variation across species within linkage groups, and (4) identifying group level benefits.

References

- Buss LW 1987 *The evolution of individuality*. Princeton University Press. Princeton, NJ
- Dieckmann U, Law R 1996 The dynamical theory of coevolution: a derivation from stochastic ecological processes. *J Math Biol* 34:579–612
- Durrett R, Levin S 1994 The importance of being discrete (and spatial). *Theor Popul Biol* 46:363–394
- Hartwell LH, Hopfield JJ, Leibler S, Murray AW 1999 From molecular to modular cell biology. *Nature* 402:C47–C52
- Huang C-Y, Ferrell JE 1996 Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci USA* 93:10078–10083
- Keller L 1999 *Levels of selection in evolution*. Princeton University Press. Princeton, NJ
- Nowak MA, May RM 2000 *Virus dynamics: mathematical principles of immunology and virology*. Oxford University Press, New York
- Williams GC 1995 *Natural selection: domains, levels and challenges*. Oxford University Press, Oxford
- Wilson KG 1979 Problems in physics with many scales of length. *Sci Am* 241:158–179

Making sense of complex phenomena in biology

Philip K. Maini

Centre for Mathematical Biology, Mathematical Institute, 24–29 St Giles, Oxford OX1 3LB

Abstract. The remarkable advances in biotechnology over the past two decades have resulted in the generation of a huge amount of experimental data. It is now recognized that, in many cases, to extract information from this data requires the development of computational models. Models can help gain insight on various mechanisms and can be used to process outcomes of complex biological interactions. To do the latter, models must become increasingly complex and, in many cases, they also become mathematically intractable. With the vast increase in computing power these models can now be numerically solved and can be made more and more sophisticated. A number of models can now successfully reproduce detailed observed biological phenomena and make important testable predictions. This naturally raises the question of what we mean by understanding a phenomenon by modelling it computationally. This paper briefly considers some selected examples of how simple mathematical models have provided deep insights into complicated chemical and biological phenomena and addresses the issue of what role, if any, mathematics has to play in computational biology.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 53–65

The enormous advances in molecular and cellular biology over the last two decades have led to an explosion of experimental data in the biomedical sciences. We now have the complete (or almost complete) mapping of the genome of a number of organisms and we can determine when in development certain genes are switched on; we can investigate at the molecular level complex interactions leading to cell differentiation and we can accurately follow the fate of single cells. However, we have to be careful not to fall into the practices of the 19th century, when biology was steeped in the mode of classification and there was a tremendous amount of list-making activity. This was recognized by D'Arcy Thompson, in his classic work *On growth and form*, first published in 1917 (see Thompson 1992 for the abridged version). He had the vision to realize that, although simply cataloguing different forms was an essential data-collecting exercise, it was also vitally important to develop theories as to how certain forms arose. Only then could one really comprehend the phenomenon under study.

Of course, the identification of a gene that causes a certain deformity, or affects an ion channel making an individual susceptible to certain diseases, has huge benefits for medicine. At the same time, one must recognize that collecting data is, in some sense, only the beginning. Knowing the spatiotemporal dynamics of the expression of a certain gene leads to the inevitable question of why that gene was switched on at that particular time and place. Genes contain the information to synthesize proteins. It is the physicochemical interactions of proteins and cells that lead to, for example, the development of structure and form in the early embryo. Cell fate can be determined by environmental factors as cells respond to signalling cues. Therefore, a study at the molecular level alone will not help us to understand how cells interact. Such interactions are highly non-linear, may be non-local, certainly involve multiple feedback loops and may even incorporate delays. Therefore they must be couched in a language that is able to compute the results of complex interactions. Presently, the best language we have for carrying out such calculations is mathematics. Mathematics has been extremely successful in helping us to understand physics. It is now becoming clear that mathematics and computation have a similar role to play in the life sciences.

Mathematics can play a number of important roles in making sense of complex phenomena. For example, in a phenomenon in which the microscopic elements are known in detail, the integration of interactions at this level to yield the observed macroscopic behaviour can be understood by capturing the essence of the whole process through focusing on the key elements, which form a small subset of the full microscopic system. Two examples of this are given in the next section. Mathematical analysis can show that several microscopic representations can give rise to the same macroscopic behaviour (see the third section), and that the behaviour at the macroscopic level may be greater than the sum of the individual microscopic parts (see the Turing model section).

Belousov–Zhabotinskii reaction

The phenomenon of temporal oscillations in chemical systems was first observed by Belousov in 1951 in the reaction now known as the Belousov–Zhabotinskii (BZ) reaction (for details see Field & Berger 1985). The classical BZ reaction consists of oxidation by bromate ions in an acidic medium catalysed by metal ion oxidants. For example, the oxidation of malonic acid in an acid medium by bromate ions, BrO_3^- , and catalysed by cerium, which has two states Ce^{3+} and Ce^{4+} . With other metal ion catalysts and appropriate dyes, the reaction can be followed by observing changes in colour. This system is capable of producing a spectacular array of spatiotemporal dynamics, including two-dimensional target patterns and outwardly rotating spiral waves, three-dimensional scroll waves and, most recently, two-dimensional inwardly rotating spirals (Vanag & Epstein 2001). All

the steps in this reaction are still not fully determined and understood and, to date, there are of the order of about 50 reaction steps known. Detailed mathematical models have been written down for this reaction (see, for example, Field et al 1972) consisting of several coupled non-linear ordinary differential equations. Remarkably, a vast range of the dynamics of the full reaction can be understood by a simplified model consisting of only three coupled, non-linear differential equations, which can be further reduced to two equations. The reduction arises due to a mixture of caricaturizing certain complex interactions and using the fact that a number of reactions operate on different time scales, so that one can use a quasi-steady-state approach to reduce some differential equations to simpler algebraic equations, allowing for the elimination of certain variables.

A phase-plane analysis of the simplified model leads to an understanding of the essence of the pattern generator within the BZ reaction, namely the relaxation oscillator. This relies on the presence of a slow variable and a fast variable with certain characteristic dynamics (see, for example, Murray 1993). The introduction of diffusion into this model, leading to a system of coupled partial differential equations, allows for the model to capture a bewildering array of the spatiotemporal phenomena observed experimentally, such as propagating fronts, spiral waves, target patterns and toroidal scrolls.

These reduced models have proved to be an invaluable tool for the understanding of the essential mechanisms underlying the patterning processes in the BZ reaction in the way that the study of a detailed computational model would have been impossible. With over 50 reactions and a myriad of parameters (many unknown), the number of simulations required to carry out a full study would be astronomical.

Models for electrical activity

The problem of how a nerve impulse travels along an axon is central to the understanding of neural communication. The Hodgkin–Huxley model for electrical firing in the axon of the giant squid (see, for example, Cronin 1987) was a triumph of mathematical modelling in physiology and they later received the Nobel Prize for their work. The model, describing the temporal dynamics of a number of key ionic species which contribute to the transmembrane potential, consists of four complicated, highly non-linear coupled ordinary differential equations. A well-studied reduction of the model, the FitzHugh–Nagumo model, is a caricature and consists of only two equations (FitzHugh 1961, Nagumo et al 1962). Again, a phase-plane analysis of this model reveals the essential phenomenon of *excitability* by which a neuron ‘fires’ and determines the kinetic properties required to exhibit this behaviour.

Models for aggregation in *Dictyostelium discoideum*

The amoeba *Dictyostelium discoideum* is one of the most studied organisms in developmental biology from both experimental and theoretical aspects and serves as a model paradigm for development in higher organisms. In response to starvation conditions, these unicellular organisms chemically signal each other via cAMP leading to a multicellular aggregation in which the amoebae undergo differentiation into a stalk type and a spore type. The latter can survive for many years until conditions are favourable.

Intercellular signalling in this system, which involves relay and transduction, has been widely studied and modelled. For example, the Martiel & Goldbeter (1987) model consists of nine ordinary differential equations. By exploiting the different timescales on which reactions occur, this model can be reduced to simpler two- and three-variable systems which not only capture most of the experimental behaviour, but also allow one to determine under which parameter constraints certain phenomena arise (Goldbeter 1996). This model turns out to exhibit excitable behaviour, similar in essence to that observed in electrical propagation in nerves.

Such reduced, or caricature models, can then serve as ‘modules’ to be plugged in to behaviour at a higher level in a layered model to understand, for example, the phenomenon of cell streaming and aggregation in response to chemotactic signalling (Höfer et al 1995a,b, Höfer & Maini 1997). Assuming that the cells can be modelled as a continuum, it was shown that the resultant model could exhibit behaviour in agreement with experimental observations. Moreover, the model provided a simple (and counterintuitive) explanation for why the speed of wave propagation slows down with increasing wave number. More sophisticated computational models, in which cells are assumed to be discrete entities, have been shown to give rise to similar behaviour (Dallon & Othmer 1997). Such detailed models can be used to compare the movement of individual cells with experimental observations and therefore allow for a degree of verification that is impossible for models at the continuum level. However, the latter are mathematically tractable and therefore can be used to determine generic behaviours.

Several models, differing in their interpretation of the relay/transduction mechanism and/or details of the chemotactic response all exhibit very similar behaviour (Dallon et al 1997). In one sense this can be thought of as a failure because modelling has been unable to distinguish between different scenarios. On the other hand, these modelling efforts illustrate that the phenomenon of *D. discoideum* aggregation is very robust and has, at its heart, signal relay and chemotaxis.

The Turing model for pattern formation

Diffusion-driven instability was first proposed by Turing in a remarkable paper (Turing 1952), as a mechanism for generating self-organized spatial patterns. He considered a pair of chemicals reacting in such a way that the reaction kinetics were stabilizing, leading to a temporally stable, spatially uniform steady state in chemical concentrations. As we know, diffusion is a homogenizing process. Yet combined in the appropriate way, Turing showed mathematically that these two stabilizing influences could conspire to produce an instability resulting in spatially heterogeneous chemical profiles—a spatial pattern. This is an example of an *emergent property* and led to the general patterning principle of *short-range activation, long-range inhibition* (Gierer & Meinhardt 1972). Such patterns were later discovered in actual chemical systems and this mechanism has been proposed as a possible biological pattern generator (for a review, see Maini et al 1997, Murray 1993).

Turing's study raises a number of important points. It showed that one cannot justifiably follow a purely reductionist approach, as the whole may well be greater than the sum of the parts and that one rules out, at one's peril, the possibility of counterintuitive phenomena emerging as a consequence of collective behaviour. It also illustrates the power of the mathematical technique because, had these results been shown in a computational model without any mathematical backing, it would have been assumed that the instability (which is, after all, counterintuitive) could only have arisen due to a computational artefact. Not only did the mathematics show that the instability was a true reflection of the model behaviour, but also it specified exactly the properties the underlying interactions in the system must possess in order to exhibit the patterning phenomenon. Furthermore, mathematics served to enhance our intuitive understanding of a complex non-linear system.

Discussion

For models to be useful in processes such as drug design, they must necessarily incorporate a level of detail that, on the whole, makes the model mathematically intractable. The phenomenal increase in computing power over recent years now means that very sophisticated models involving the interaction of hundreds of variables in a complex three-dimensional geometry can be solved numerically. This naturally raises a number of questions. (1) How do we validate the model? Specifically, if the model exhibits a counterintuitive result, which is one of the most powerful uses of a model, how do we know that this is a faithful and generic outcome of the model and not simply the result of very special choice of model parameters, or an error in coding? (2) If we take

modelling to its ultimate extreme, we simply replace a biological system we do not understand by a computational model we do not understand. Although the latter is useful in that it can be used to compute the results of virtual experiments, can we say that the exercise has furthered our understanding? Moreover, since it is a model and therefore, by necessity, wrong in the strict sense of the word, how do we know that we are justified in using the model in a particular context?

In going from the gene to the whole organism, biological systems consist of an interaction of processes operating on a wide range of spatial and temporal scales. It is impossible to compute the effects of all the interactions at any level of this spatial hierarchy, even if they were all known. The approach to be taken, therefore, must involve a large degree of caricaturizing (based on experimental experience) and reduction (based on mathematical analysis). The degree to which one simplifies a model depends very much on the question one wishes to answer. For example, to understand in detail the effect of a particular element in the transduction pathway in *D. discoideum* will require a detailed model at that level. However, for understanding aspects of cell movement in response to the signal, it may be sufficient to consider a very simple model which represents the behaviour at the signal transduction level, allowing most of the analytical and computational effort to be spent on investigating cell movement. In this way, one can go from one spatial level to another by 'modularizing' processes at one level (or layer) to be plugged in to the next level. To do this, it is vital to make sure that the appropriate approximations have been made and the correct parameter space and spatiotemporal scales are used. This comes most naturally via a mathematical treatment. Eventually, this allows for a detailed mathematical validation of the process before one begins to expand the models to make them more realistic.

The particular examples considered in this article use the classical techniques of applied mathematics to help understand model behaviour. Much of the mathematical theory underlying dynamical systems and reaction–diffusion equations was motivated by problems in ecology, epidemiology, chemistry and biology. The excitement behind the Turing theory of pattern formation and other areas of non-linear dynamics was that very simple interactions could give rise to very complex behaviour. However, it is becoming increasingly clear that often in biology very complex interactions give rise to very simple behaviours. For example, complex biochemical networks are used to produce only a limited number of outcomes (von Dassow et al 2000). This suggests that it may be the interactions, not the parameter values, that determine system behaviour and, in particular, robustness. This requires perhaps the use of topological or graph theoretical ideas as tools for investigation. Hence it is clear that it will be necessary to incorporate tools from other branches of mathematics and to

develop new mathematical approaches if we are to make sense of the mechanisms underlying the complexity of biological phenomena.

Acknowledgements

This paper was written while the author was a Senior Visiting Fellow at the Isaac Newton Institute for Mathematical Sciences, University of Cambridge. I would like to thank Santiago Schnell and Dr Edmund Crampin for helpful discussions.

References

- Cronin J 1987 *Mathematical aspects of Hodgkin–Huxley neural theory*. Cambridge University Press, Cambridge
- Dallon JC, Othmer HG 1997 A discrete cell model with adaptive signalling for aggregation of *Dictyostelium discoideum*. *Philos Trans R Soc Lond B Biol Sci* 352:391–417
- Dallon JC, Othmer HG, van Oss C et al 1997 Models of *Dictyostelium discoideum* aggregation. In: Alt W, Deutsch G, Dunn G (eds) *Dynamics of cell and tissue motion*. Birkhäuser-Verlag, Boston, MA, p 193–202
- Field RJ, Burger M 1985 *Oscillations and travelling waves in chemical systems*. Wiley, New York
- Field RJ, Körös E, Burger M 1972 Oscillations in chemical systems, Part 2. Thorough analysis of temporal oscillations in the bromate-cerium-malonic acid system. *J Am Chem Soc* 94:8649–8664
- FitzHugh R 1961 Impulses and physiological states in theoretical models of nerve membrane. *Biophys J* 1:445–466
- Gierer A, Meinhardt H 1972 A theory of biological pattern formation. *Kybernetik* 12:30–39
- Goldbeter A 1996 *Biochemical oscillations and cellular rhythms*. Cambridge University Press, Cambridge
- Höfer T, Maini PK 1997 Streaming instability of slime mold amoebae: An analytical model. *Phys Rev E* 56:2074–2080
- Höfer T, Sherratt JA, Maini PK 1995a *Dictyostelium discoideum*: cellular self-organization in an excitable biological medium. *Proc R Soc Lond B Biol Sci* 259:249–257
- Höfer T, Sherratt JA, Maini PK 1995b Cellular pattern formation during *Dictyostelium* aggregation. *Physica D* 85:425–444
- Maini PK, Painter KJ, Chau HNP 1997 Spatial pattern formation in chemical and biological systems. *Faraday Transactions* 93:3601–3610
- Martiel JL, Goldbeter A 1987 A model based on receptor desensitization for cyclic AMP signaling in *Dictyostelium* cells. *Biophys J* 52:807–828
- Murray JD 1993 *Mathematical biology*. Springer-Verlag, Berlin
- Nagumo JS, Arimoto S, Yoshizawa S 1962 An active pulse transmission line simulating nerve axon. *Proc Inst Radio Eng* 50:2061–2070
- Thompson DW 1992 *On growth and form*. Cambridge University Press, Cambridge
- Turing AM 1952 The chemical basis of morphogenesis. *Philos Trans R Soc Lond B Biol Sci* 3327:37–72
- Vanag VK, Epstein IR 2001 Inwardly rotating spiral waves in a reaction–diffusion system. *Science* 294:835–837
- von Dassow G, Meir E, Munro EM, Odell GM 2000 The segment polarity network is a robust developmental module. *Nature* 406:188–192

DISCUSSION

Noble: We will almost certainly revisit the question of levels and reduction versus integration at some stage during this meeting. But it's important to clarify here that you and your mathematical colleagues are using the term 'reduction' in a different sense to that which we biologists use. Let me clarify: when you 'reduce' the Hodgkin–Huxley equations to FitzHugh–Nagumo equations, you are not doing what would be regarded as reduction in biology, which would be to say that we can explain the Hodgkin–Huxley kinetics in terms of the molecular structure of the channels. You are asking whether we can use fewer differential equations, and whether as a result of that we get an understanding. It is extremely important to see those senses of reduction as being completely different.

Maini: I agree; that's an important point.

Noble: Does mathematical reduction always go that way? I was intrigued by the fact that even you, as a mathematician, said you had to understand how that graph worked, in order to understand the mathematics. I always had this naïve idea that mathematicians just understood! I take it there are different sorts of mathematicians, as well as different kinds of biologists, and some will be able to understand things from just the equations. Presumably, the question of understanding in maths is also an issue.

Maini: What I meant by 'understanding' is that we need to determine what are the crucial properties of the system that make it behave in the way that it does. The easiest method for doing that in this case is a phase-plane analysis. This tells us that the behaviour observed is generic for a wide class of interactions, enabling us to determine how accurately parameters must be measured. My talk focused on the differential equation approach to modelling. However, there may be cases where other forms of modelling and/or analysis — for example, graph theory, networks or topology — may be more appropriate. An issue here is how do we expose these problems to those communities?

Loew: I would assert that the kind of mathematical reduction you were talking about — basically, extending your mathematical insights to produce a minimal model — may provide insights to mathematicians, but in most cases it wouldn't be very useful to a biologist. This is because in creating the minimal model you have eliminated many of the parameters that may tie the model to the actual biology. In the BZ reaction you mentioned, you were able to list all of the individual reactions. A biologist would want to see this list of reactions, and see what happens if there is a mutant that behaves a little differently. What does this do to the overall behaviour? You wouldn't be able to use the model, at least as not as directly, if you had your minimal model instead. I feel that it takes us one step further away from biology if we produce these kinds of minimal models.

Maini: It depends what sort of reduction you do. If you use quasi-steady-state assumptions, the parameters in the reduced model are actually algebraically related to the parameters in the full model, so you can still follow through and compute the effects of changing parameters at the level of the full model. Very little information is lost. My concern about very detailed computational models is that one is replacing a complicated biological system one wishes to understand by a complicated computational model one does not understand. Of course, in the very detailed model one can see the outcome of changing a specific parameter, but how do you know whether the answer is correct if you cannot determine on what processes in the model the outcome depends?

Loew: I think it is important because of the issue Denis Noble raised at the beginning of the meeting: about whether there is the possibility for a theoretical biology. If you can produce minimal equations that you can somehow use in a useful way to describe a whole class of biology, this would be very important. I can see analogies in chemistry, where there are some people who like to do *ab initio* calculations in theoretical chemistry, trying to understand molecular structure in the greatest detail. But sometimes it is more useful to get a broader view of the patterns of behaviour and look at things in terms of interaction of orbitals. There it is very useful. Chemistry has found what you call the 'reductionist' approach very useful. It remains to be seen whether this will be useful in biology.

Maini: I would argue that it has already been shown in Kees Weijer's work that such an approach is very useful. He has beautiful models for *Dictyostelium*. He is an experimentalist, and works with mathematicians in the modelling. When it comes to looking at how the cells interact with each other, he will use reductions such as FitzHugh–Nagumo. His approach has resulted in a very detailed understanding of pattern formation processes in *Dictyostelium discoideum*.

Crampin: One of the things mathematics is useful for is to abstract phenomena from specific models to reveal general properties of particular types of system. For example, if you combine an excitable kinetic system with chemotaxis for cell movement, then you will always get the sorts of behaviour that Philip Maini is describing. In this respect, the biological details become unimportant. However, if you do start with a complicated model and use mathematical techniques to reduce the model to a mathematically tractable form, then you can keep track of where different parameters have gone. Some of the variables will turn out not to have very much bearing on what goes on. These you can eliminate happily, knowing that if the biologist goes away and does an experiment, then changing these parameters is not going to have a strong effect. But the important ones you will keep, and they will still appear in the final equations. You should be able to predict what effect varying these parameters in experiments will have. Reducing the mathematical complexity doesn't necessarily throw out all of the biology.

Hunter: If you accept that both approaches are needed (I think they are complementary), who is doing the process of linking the two? Having got the dispersion relation and the parameter range that leads to instability, how does one map this back to the biological system? And how do we deduce general ways of moving between the state space of 11 equations to the state space of two equations?

Maini: That's an issue we have been trying to tackle. There are certain approaches such as homogenization techniques for looking at these sorts of issues. But most of the homogenization techniques that I have seen in the materials context tend to be very specialized. I think it is a challenging problem. Most mathematicians are more interested in proving theorems and are not really interested in such messy applications. They will happily take the sort of equations that I wrote down and throw out a few more terms, so they can just prove some theorem, without caring where the equations arrived from. That is fine, because good mathematics may come out of it, but it is not mathematical biology. Perhaps it will be the physicists who will help to bridge the gap that exists.

Noble: There are obviously different demands here. Part of what you said in relation to helping the biologists was highly significant. It was determining where there was robustness, which I think is extremely important. This may correspond to part of what we call the logic of life. If, through comparing different reductions and the topology of different models, we can end up with a demonstration of robustness, then we have an insight that is biologically important whether or not anyone else goes on to use those mathematical reductions in any of their modelling. Another success is as follows. Where in our computationally heavy modelling we have come up with counterintuitive results, then going back to the mathematicians and asking them to look at it has proven extremely valuable. One example of this is in relation to investigating one of the transporters involved in ischaemic heart disease, where we came across what still seems to me to be a counterintuitive result when we down-regulated or up-regulated this transporter. We gave this problem to Rob Hinch, to see whether he could look at it mathematically. He demonstrated that it was a necessary feature of what it is that is being modelled. This is another respect in which mathematical reduction (as distinct from the biological kind) must be a help to us where we are puzzled by the behaviour of our more complicated models. So we have some unalloyed successes that we can chalk up, even if people don't go on to use the reductions in their modelling.

Hinch: The idea of all modelling, if it is to be useful and predictive, is for it to come up with some original ideas. If you have a very complex simulation model which comes up with a new idea, you do not know whether that is an artefact of the actual model, or if it is a real mechanism occurring. The power of mathematics and the mathematical analysis where these counterintuitive results come up, is that you

can pinpoint what is causing this novel behaviour to happen. This would be a much better way to direct the experimental work. The idea is that by having these reduced models we can understand the mechanism of this interesting behaviour, which will immediately make it much easier for an experimentalist to see whether this is a real phenomenon, or just an artefact of the modelling.

Crampin: In addition to what Philip Maini said, I want to draw a distinction between on the one hand this type of mathematical reduction (formal ways of moving between complicated models and simpler representations), and on the other hand the ‘art’ of modelling—using scientific insight to do that same process. I am not sure whether there will ever be general formal methods for taking a complicated model and generating a simpler one. In practice one uses a combination of approaches, both formally manipulating the equations and using knowledge of the system you are working on. There is also an interesting difference between simulation models and analytical models. The tradition in applied mathematics is that a model is developed to answer a specific question, just for that purpose. It is unlikely for people to expect that model to be used in all sorts of different contexts. In contrast, if we are talking about generating simulation tools, models must be sufficiently general to be applicable in all sorts of different areas, even if you are building computational tools where you can construct models on an *ad hoc* basis for each problem.

Noble: Yes, the modellers are building a jigsaw.

Loew: I certainly appreciate the value of producing a minimal model, both from the point of view of the mathematical insight that it provides, and also from the practical point of view of being able to use a reduced form of a model as a building block for a more complex model. This is certainly an important modelling technique. But the reason I was deliberately being provocative was because we need to be able to connect to the laboratory biologist. It is important not only to avoid just being mathematicians who prove theorems but also to always be practical about how the models are being used as aids for biology. If they get too abstract, then the biologists get very quickly turned off to what we are doing.

Winslow: There is another sense in which model reduction can be performed. It doesn’t involve reducing the number of equations used to describe a system, but rather involves using computational techniques to study the generic properties of those equations. These approaches have been used with some success. One example is bifurcation theory to understand the generic behaviours of non-linear systems subject to parameter variation. This kind of model reduction is where a complex, oscillating cell may be equivalent to a much simpler oscillating system by virtue of the way in which it undergoes oscillation, perhaps by a half-bifurcation. There is no reduction in the number of equations here, but lumping of systems into those that share these general dynamical properties.

Paterson: Les Loew, you commented that for the lab biologist, we need to present models in a form they see as relevant. There is a whole branch of biology that looks at people as opposed to cells! I have people on my staff who you can show gene expression data until you are blue in the face, but they want to understand a complex disease state such as diabetes where there are huge unanswered questions of integrated physiology that can only be answered by investigations at the clinical level. In terms of tying models to the biology you are right, and for bench scientists working with high-throughput *in vitro* data, I think the types of very detailed models we are talking about are very necessary. But in terms of tying it to extremely relevant data at the clinical level, for understanding the manifestation of disease states, you can't afford to build a model at the gene expression level for a complicated disease state such as diabetes. While gene expression data in key pathways may be relevant, clinical data of the diverse phenotype must be linked as well. How this relates to Peter Hunter's point about the transition, is that biology gives us a wonderful stepping stone—the cell. There is a tremendous amount of detail within the cell. I would be interested to hear estimates of the fraction of the proteins coded by the genome that actually participate in communication outside the cell membrane. My guess is that it is an extremely small fraction. If you look at the cell as a highly self-organized information and resource-processing entity, and consider that it is participating in many different activities taking place in the organism, then there are opportunities to operate at a more highly aggregated level where you are looking at aggregated cellular functions that link up to clinical data. Then you go into the more detailed cellular models to link into *in vitro* and gene expression data. In this way you can have your cake and eat it too. The fact that the cell represents a nice bridging point between these two extremes can help us provide multiple modelling domains that are relevant to molecular cell biologists and clinical biologists.

Cassman: Philip Maini, what did you mean by the term 'robustness'? This is another term that is thrown around a lot. It usually means that the output is insensitive to the actual parameterization of the model. I'm not sure this is what you meant.

Maini: What I meant in this particular context is that in some of these models you could change the parameter values by several orders of magnitude and it would not qualitatively change the outcome.

Noble: There's another possible sense, which I regard as extremely important. Between the different models we determine what is essential, and, having done the mathematical analysis, we can say that the robustness lies within a certain domain and these models are inside it, but another model is outside it.

Berridge: For those of us who are simple-minded biologists, when we come across something like *Dictyostelium* with five or six models all capable of explaining the same phenomenon but apparently slightly different, which one are

we going to adopt? There needs to be some kind of seal of approval so we know which one to opt for.

Crampin: To turn that on its head, as a modeller reading the primary experimental literature, I often find completely conflicting results!

Berridge: One of the nice things about Philip Maini's paper was that he was able to explain this very complicated behaviour of cells aggregating, including complex spiral waves, using just two ideas. One was the excitable medium idea, and the other one was chemotaxis. While he used chemotaxis as part of the model, I don't think there is anything in the model that actually explains the phenomenon of chemotaxis. This is a complex phenomenon, for which I don't think there is a mathematical model. How is it that a cell can detect a minute gradient between its front end and back end? While those working on eukaryotes don't have a good model, people working on bacteria do. This is where we really need some help from the mathematicians, to give us a clue as to the sorts of parameters a cell might use to detect minute gradients and move in the right direction.

Maini: There are mathematicians trying to model the movement of individual cells.

Berridge: It's not the movement I'm referring to, but the actual detection of the gradient.

Shimizu: The gradient-sensing mechanism is very well understood in bacteria. The cell compares the concentration that is being detected at present to the concentration that was detected a few seconds ago in the past. So in bacteria, it is by temporal comparisons that the gradient is measured. This is different from the spatial comparisons that *Dictyostelium* makes.

Berridge: I understand the bacterial system; it is the eukaryotic cell where it isn't clear. There isn't a model that adequately explains how this is done.

On ontologies for biologists: the Gene Ontology — untangling the web

Michael Ashburner* and Suzanna Lewis†

*Department of Genetics, University of Cambridge and EMBL — European Bioinformatics Institute, Hinxton, Cambridge, UK and †Berkeley Drosophila Genome Project, Lawrence Berkeley National Laboratory, University of California, Berkeley, CA 94720, USA

Abstract. The mantra of the 'post-genomic' era is 'gene function'. Yet surprisingly little attention has been given to how functional and other information concerning genes is to be captured, made accessible to biologists or structured in a computable form. The aim of the Gene Ontology (GO) Consortium is to provide a framework for both the description and the organisation of such information. The GO Consortium is presently concerned with three structured controlled vocabularies which can be used to describe three discrete biological domains, building structured vocabularies which can be used to describe the molecular function, biological roles and cellular locations of gene products.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 66–83

Status

The Gene Ontology (GO) Consortium's work is motivated by the need of both biologists and bioinformaticists for a method for rigorously describing the biological attributes of gene products (Ashburner et al 2000, The Gene Ontology Consortium 2001). A comprehensive lexicon (with mutually understood meanings) describing those attributes of molecular biology that are common to more than one life form is essential to enable communication, in both computer and natural languages. In this era, when newly sequenced genomes are rapidly being completed, all needing to be discussed, described, and compared, the development of a common language is crucial.

The most familiar of these attributes is that of 'function'. Indeed, as early as 1993 Monica Riley attempted a hierarchical functional classification of all the then known proteins of *Escherichia coli* (Riley 1993). Since then, there have been other

attempts to provide vocabularies and ontologies¹ for the description of gene function, either explicitly or implicitly (e.g. Dure 1991, Commission of Plant Gene Nomenclature 1994, Fleischmann et al 1995, Overbeek et al 1997, 2000 Takai-Igarashi et al 1998, Baker et al 1999, Mewes et al 1999, Stevens et al 2000; see Riley 1988, Rison et al 2000, Sklyar 2001 for reviews). Riley has recently updated her classification for the proteins of *E. coli* (Serres & Riley 2000, Serres et al 2001).

One problem with many (though not all: e.g. Schulze-Kremer 1997, 1998, Karp et al 2002a,b) efforts prior to that of the GO Consortium is that they lacked semantic clarity due, to a large degree, to the absence of definitions for the terms used. Moreover, these previous classifications were usually not explicit concerning the relationships between different (e.g. 'parent' and 'child') terms or concepts. A further problem with these efforts was that, by and large, they were developed as one-off exercises, with little consideration given to revision and implementation beyond the domain for which they were first conceived. They generally also lacked the apparatus required for both persistence and consistent use by others, i.e. versioning, archiving and unique identifiers attached to their concepts.

The GO vocabularies distinguish three orthogonal domains (vocabularies); the concepts within one vocabulary do not overlap those within another. These domains are molecular_function, biological_process and cellular_component, defined as follows:

- molecular_function: an action characteristic of a gene product.
- biological_process: a phenomenon marked by changes that lead to a particular result, mediated by one or more gene products.
- cellular_component: the part, or parts, of a cell of which a gene product is a component; for this purpose includes the extracellular environment of cells.

The initial objective of the GO Consortium is to provide a rich, structured vocabulary of terms (concepts) for use by those annotating gene products within an informatics context, be it a database of the genetics and genomics of a model organism, a database of protein sequences or a database of information about gene products, such as might be obtained from a DNA microarray experiment. In GO the annotation of gene products with GO terms follows two guidelines: (1) all annotations include the evidence upon which an assertion is based and, (2)

¹Philosophically speaking an ontology is 'the study of that which exists' and is defined in opposition to 'epistemology', which means 'the study of that which is known or knowable'. Within the field of artificial intelligence the term ontology has taken on another meaning: 'A specification of a conceptualization that is designed for reuse across multiple applications and implementations' (Karp 2000) and it is in this sense that we use this word.

the evidence provided for each annotation includes attribution to an available external source, such as a literature reference.

Databases using GO for annotation are widely distributed. Therefore an additional task of the Consortium is to provide a centralized holding site for their annotations. GO provides a simple format for contributing databases to submit their annotations to a central annotation database maintained by GO. The annotation data submitted include the association of gene products with GO terms as well as ancillary information, such as evidence and attribution. These annotations can then form the basis for queries—either by an individual or a computer program.

At present, gene product associations are available for several different organisms, including two yeasts (*Schizosaccharomyces pombe* and *Saccharomyces cerevisiae*), two invertebrates (*Caenorhabditis elegans* and *Drosophila melanogaster*), two mammals (mouse and rat) and a plant (*Arabidopsis thaliana*). In addition, the first bacterium (*Vibrio cholerae*) has now been annotated with GO and efforts are now underway to annotate all 60 or so publicly available bacterial genomes. Over 80% of the proteins in the SWISS-PROT protein database have been annotated with GO terms (the majority by automatic annotation, see below), these include the SWISS-PROT to GO annotations of over 16 000 human proteins (available at www.geneontology.org/cgi-bin/GO/downloadGOGA.pl/gene_association.goa-human). Some 7000 human proteins were also annotated with GO by Proteome Inc. and are available from LocusLink (Pruitt & Maglott 2001).

A number of other organismal databases are in the process of using GO for annotation, including those for *Plasmodium falciparum* (and other parasitic protozoa) (M. Berriman, personal communication), *Dictyostelium discoideum* (R. Chisholm, personal communication) and the grasses (rice, maize, wheat, etc.) (GRAMENE 2002). The availability of these sets of data has led to the construction of GO browsers which enable users to query them all simultaneously for genes whose products serve a particular function, play a role in a particular biological process or are located in a particular subcellular part (AmiGO 2001). These associations are also available as tab-delimited tables (www.geneontology.org/#annotations) or with protein sequences. GO thus achieves *de facto* a degree of database integration (see Leser 1998), one Holy Grail of applied bioinformatics.

Availability

The products of the GO Consortium's work can be obtained from their World Wide Web home page: www.geneontology.org.

All of the efforts of the GO Consortium are placed in the public domain and can be used by academia or industry alike without any restraint, other than they cannot

be modified and then passed off as the products of the Consortium. This is true for all classes of the GO Consortium's products, including the controlled vocabularies, the gene-association tables, and software for browsing and editing the GO vocabularies and gene association tables (AmiGO 2001, DAG Edit 2001). Thus the GO Consortium's work is very much in the spirit of the Open Source tradition in software development (DiBona et al 1999, OpenSource 2001). The GO ontologies and their associated files are available as text files, in XML or as tables for a MySQL database.

The structure of the GO ontologies

All biologists are familiar with hierarchical graphs—the system of classification introduced by Linnaeus has been a bedrock for biological research for some 250 years. In a Linnean taxonomy the nodes of the graphs are the names of taxa, be they phyla or species; the edges between these nodes represent the relationship 'is a member of' between parent and child nodes. Thus the node 'species:*Drosophila melanogaster*' 'is a member of' its parent node 'genus:*Drosophila*'. Useful as hierarchies are, they suffer from a serious limitation: each node has one and only one parental node—no species is a member of two (or more) genera, no genus a member of two (or more) families. Yet in the broader world of biology an object may well have two or more parents. Consider, as a simple example, a protein that both binds DNA and hydrolyses ATP. It is as equally correct to describe this as a 'DNA binding protein' as it is to describe it as a 'catalyst' (or enzyme); therefore it should be a child of both within a tree structure. Not all DNA binding proteins are enzymes, not all enzymes are DNA binding proteins, yet some are and we need to be able to represent these facts conceptually. For this reason GO uses a structure known as a directed acyclic graph (DAG), a graph in which nodes can have many parents but in which cycles—that is a path which starts and ends at the same node—are not allowed. All nodes must have at least one parent node, with the exception of the root of each graph.

Alice replies to Humpty Dumpty's inquiry as to the meaning of her name 'Must a name mean something?' 'Of course it must', replies Humpty Dumpty (Heath 1974). This is as true in the real world as in that through the looking glass. The nodes in the GO controlled vocabularies are concepts: concepts that describe the molecular function, biological role or cellular location of gene products. The terms used by GO are simply a shorthand way of referring to these concepts, which are restricted by their natural language definitions. (At present only 20% of the 10 000 or so GO terms are defined but a major effort to correct this situation will be launched early in 2002.) Each and every GO term has a unique identifier consisting of the prefix 'GO:' and an integer, for example, GO:0036562. But what happens if a GO term changes? A change may be as trivial as correcting a

spelling error or as drastic as being a new lexical string. If the change does not change the meaning of the term then there is no change to the GO identifier. If the meaning *is* changed, however, then the old term, its identifier and definition are retired (they are marked as 'obsolete', they never disappear from the database) and the new term gets a new identifier and a new definition. Indeed this is true even if the lexical string is identical between old and new terms; thus if we use the same words to describe a different concept then the old term is retired and the new is created with its own definition and identifier. This is the only case where, within any one of the three GO ontologies, two or more concepts may be lexically identical; all except one of them must be flagged as being obsolete. Because the nodes represent semantic concepts (as described by their definitions) it is not strictly necessary that the terms are unique, but this restriction is imposed in order to facilitate searching. This mechanism helps with maintaining and synchronizing other databases that must track changes within GO, which is, by design, being updated frequently. Keeping everything and everyone consistent is a difficult problem that we had to solve in order permit this dynamic adaptability of GO.

The edges between the nodes represent the relationships between them. GO uses two very different classes of semantic relationship between nodes: 'isa' and 'partof'. Both the isa and partof relationships within GO should be fully transitive. That is to say an instance of a concept is also an instance of all of the parents of that concept (to the root); a part concept that is partof a whole concept is a partof all of the parents of that concept (to the root). Both relationships are reflexive (see below).

The isa relationship is one of subsumption, a relationship that permits refinement in concepts and definitions and thus enables annotators to draw coarser or finer distinctions, depending on the present degree of knowledge. This class of relationship is known as hyponymy (and its reflexive relation hypernymy) to the authors of the lexical database WordNet (Fellbaum 1998). Thus the term DNA binding is a hyponym of the term nucleic acid binding; conversely nucleic acid binding is a hypernym of DNA binding. The latter term is more specific than the former, and hence its child. It has been argued that the isa relationship, both generally (see below) and as used by GO (P. Karp, personal communication; S. Schultze-Kremer, personal communication) is complex and that further information describing the nature of the relationship should be captured. Indeed this is true, because the precise connotation of the isa relationship is dependent upon each unique pairing of terms and the meanings of these terms. Thus the isa relationship is not a relationship between terms, but rather is a relationship between particular concepts. Therefore the isa relationship is not a single type of relationship; its precise meaning is dependent on the parent and child terms it connects. The relationship simply describes the parent as the more general

concept and the child as the more precise concept and says nothing about how the child specifically refines the concept.

The partof relationship (meronymy and its reflexive relationship holonymy) (Cruse 1986, cited in Miller 1998) is also semantically complex as used by GO (see Winston et al 1987, Miller 1998, Priss 1998, Rogers & Rector 2000). It may mean that a child node concept 'is a component of' its parent concept. (The reflexive relationship [holonymy] would be 'has a component'.) The mitochondrion 'is a component of' the cell; the small ribosomal subunit 'is a component of' the ribosome. This is the most common meaning of the partof relationship in the GO cellular_component ontology. In the biological_process ontology, however, the semantic meaning of partof can be quite different, it can mean 'is a subprocess of'; thus the concept amino acid activation 'is a subprocess of' of the concept protein biosynthesis. It is in the future for the GO Consortium to clarify these semantic relationships while, at the same time not making the vocabularies too cumbersome and difficult to maintain and use.

Meronymy and hyponymy cause terms to 'become intertwined in complex ways' (Miller 1998:38). This is because one term can be a hyponym with respect to one parent, but a meronym with respect to another. Thus the concept cytosolic small ribosomal subunit is both a meronym of the concept cytosolic ribosome and a hyponym of the concept small ribosomal subunit, since there also exists the concept mitochondrial small ribosomal subunit.

The third semantic relationship represented in GO is the familiar relationship of synonymy. Each concept defined in GO (i.e. each node) has one primary term (used for identification) and may have zero or many synonyms. In the sense of the WordNet noun lexicon a term and its synonyms at each node represents a synset (Miller 1998); in GO, however, the relationship between synonyms is strong, and not as context dependent as in WordNet's synsets. This means that in GO all members of synset are completely interchangeable in whatever context the terms are found. That is to say, for example, that 'lymphocyte receptor of death' and 'death receptor 3' are equivalent labels for the same concept and are conceptually identical. One consequence of this strict usage is that synonyms are not inherited from parent to child concepts in GO.

The final semantic relationship in GO is a cross-reference to some other database resource, representing the relationship 'is equivalent to'. Thus the cross-reference between the GO concept alcohol dehydrogenase and the Enzyme Commission's number EC:1.1.1.1 is an equivalence (but not necessarily an identity, these cross-references within GO are for a practical rather than theoretical purpose). As with synonyms, database cross-references are not inherited from parent to child concept in GO.

As we have expressed, we are not fully satisfied that the two major classes of relationship within GO, isa and partof, are yet defined as clearly as we would

like. There is, moreover, some need for a wider agreement in this field on the classes of relationship that are required to express complex relationships between biological concepts. Others are using relationships that, at first sight appear to be similar to these. For example, within the aMAZE database (van Helden et al 2001) the relationships ContainedCompartment and SubType appear to be similar to GO's partof and isa, respectively. Yet ContainedCompartment and partof have, on closer inspection, different meanings (GO's partof seems to be a much broader concept than aMAZE's ContainedCompartment).

The three domains now considered by the GO Consortium, molecular_function, biological_process and cellular_component are orthogonal. They can be applied independently of each other to describe separable characteristics. A curator can describe where some protein is found without knowing what process it is involved in. Likewise, it may be known that a protein is involved in a particular process without knowing its function. There are no edges between the domains, although we realize that there are relationships between them. This constraint was made because of problems in defining the semantic meanings of edges between nodes in different ontologies (see Rogers & Rector 2000, for a discussion of the problems of transitivity met within an ontology that includes different domains of knowledge). This structure is, however, to a degree, artificial. Thus all (or, certainly most) gene products annotated with the GO function term transcription factor will be involved in the process transcription, DNA-dependent and the majority will have the cellular location nucleus. This really becomes important not so much within GO itself, but at the level of the use of GO for annotation. For example, if a curator were annotating genes in FlyBase, the genetic and genomic database for *Drosophila* (FlyBase 2002), then it would be an obvious convenience for a gene product annotated with the function term transcription factor to inherit both the process transcription, DNA-dependent and the location nucleus. There are plans to build a tool to do this, but one that allows a curator to say to the system 'in this case do not inherit' where to do so would be misleading or wrong.

Annotation using GO

There are two general methods for using GO to annotate gene products within a database. These may be characterized as the 'curatorial' and 'automatic' methods. By 'curatorial' we mean that a domain expert annotates gene products with GO terms as the result of either reading the relevant literature or by an evaluation of a computational result (see for example Dwight et al 2002). Automated methods rely solely on computational sequence comparisons such as the result of a BLAST (Altschul et al 1990) or InterProScan (Zdobnov & Apweiler 2001) analysis of a gene product's known or predicted protein sequence. Whatever method is used,

the basis for the annotation is then summarized, using a small controlled list of phrases (www.geneontology.org/GO.evidence.html); perhaps ‘inferred from direct assay’ if annotating on the evidence of experimental data in a publication or ‘inferred from sequence comparison with database:object’ (where database:object could be, for example, SWISS-PROT:P12345, where P12345 is a sequence accession in the SWISS-PROT database of protein sequences), if the inference is made from a BLAST or InterProScan compute which has been evaluated by a curator.

The incorrect inference of a protein’s or predicted protein’s function from sequence comparison is well known to be a major problem and one that has often contaminated both databases and the literature (Kyripides & Ouzounis 1998, for one example among many). The syntax of GO annotation in databases allows curators to annotate a protein as NOT having a particular function despite impressive BLAST data. For example, in the genome of *Drosophila melanogaster* there are at least 480 proteins or predicted proteins that any casual or routine curation of BLASTP output would assign the function peptidase (or one of its child concepts) yet, on closer inspection, at least 14 of these lack residues required for the catalytic function of peptidases (D. Coates, personal communication). In FlyBase these are curated with the ‘function’ ‘NOT peptidase’. What is needed is a comprehensive set of computational rules to allow curators, who cannot be experts in every protein family, to automatically detect the signatures of these cases, cases where the transitive inference would be incorrect (Kretschmann et al 2001). It is also conceivable that triggers to correct dependent annotations could be constructed because GO annotations track the identifiers of the sequence upon which annotation is based.

Curatorial annotation will be at a quality proportional both to the extent of the available evidence for annotation and the human resources available for annotation. Potentially, its quality is high but at the expense of human effort. For this reason several ‘automatic’ methods for the annotation of gene products are being developed. These are especially valuable for a first-pass annotation of a large number of gene products, those, for example, from a complete genome sequencing project. One of the first to be used was M. Yandell’s program LOVEATFIRSTSIGHT developed for the annotation of the gene products predicted from the complete genome of *Drosophila melanogaster* (Adams et al 2000). Here, the sequences were matched (by BLAST) to a set of sequences from other organisms that had already been curated using GO.

Three other methods, DIAN (Pouliot et al 2001), PANTHER (Kerlavage et al 2002) and GO Editor (Xie et al 2002), also rely on comprehensive databases of sequences or sequence clusters that have been annotated with GO terms by curation, albeit with a large element of automation in the early stages of the process. PANTHER is a method in which proteins are clustered into

'phylogenetic' families and subfamilies, which are then annotated with GO terms by expert curators. New proteins can then be matched to a cluster (in fact to a Hidden Markov Model describing the conserved sequence patterns of that cluster) and transitively annotated with appropriate GO terms. In a recent experiment PANTHER performed well in comparison with the curated set of GO annotations of *Drosophila* genes in FlyBase (Mi et al 2002). DIAN matches proteins to a curated set using two algorithms, one is vocabulary based and is only suitable for sequences that already have some attached annotation; the other is domain based, using Pfam Hidden Markov Models of protein domains.

Even simpler methods have also been used. For example, much of the first-pass GO annotation of mouse proteins was done by parsing the KEYWORDS attached to SWISS-PROT records of mouse proteins, using a file that semantically mapped these KEYWORDS to GO concepts (see www.geneontology.org/external2go/spkw2go) (Hill et al 2001).

Automatic annotations have the advantages of speed, essential if large protein data sets are to be analysed within a short time. Their disadvantage is that the accuracy of annotation may not be high and the risk of errors by incorrect transitive inference is great. For this reason, all annotations made by such methods are tagged in GO gene-association files as being 'inferred by electronic annotation'. Ideally, all such annotations are reviewed by curators and subsequently replaced by annotations of higher confidence.

The problems of complexity and redundancy

There are in the biological _process ontology many words or strings of words that have no business being there. The major examples of offending concepts are chemical names and anatomical parts. There are two reasons why this is problematic, one practical and the other of more theoretical importance. The practical problem is one of maintainability. The number of chemical compounds that are metabolized by living organisms is vast. Each one deserves its own unique set of GO terms: carbohydrate metabolism (and its children carbohydrate biosynthesis, carbohydrate catabolism), carbohydrate transport and so on. In the ideal world there would exist a public domain ontology for natural (and xenobiotic) compounds:

```
carbohydrate
  simple carbohydrate
    pentose
    hexose
      glucose
      galactose
    polysaccharide
```

and so on. Then we could make the cross-product between this little DAG (a DAG because a carbohydrate could also be an acid or an alcohol, for example) and this small biological _process DAG:

```
metabolism
  biosynthesis
  catabolism
```

to produce automatically:

```
carbohydrate metabolism
  carbohydrate biosynthesis
  carbohydrate catabolism
  simple carbohydrate metabolism
  simple carbohydrate biosynthesis
  simple carbohydrate catabolism
  pentose metabolism
    pentose biosynthesis
    pentose catabolism
  hexose metabolism
    hexose biosynthesis
    hexose catabolism
    glucose metabolism
      glucose biosynthesis
      glucose catabolism
    galactose metabolism
      galactose biosynthesis
      galactose catabolism
  polysaccharide metabolism
    polysaccharide biosynthesis
    polysaccharide catabolism
```

Such cross-product DAGs may often have compound terms that are not appropriate. For example, the GO concepts 1,1,1-trichloro-2,2-bis-(4'-chlorophenyl)ethane metabolism and 1,1,1-trichloro-2,2-bis-(4'-chlorophenyl)ethane catabolism are appropriate, yet 1,1,1-trichloro-2,2-bis-(4'-chlorophenyl)ethane biosynthesis is not; organisms break down DDT but do not synthesise it. For this reason any cross-product tree would need pruning by a domain expert subsequent to its computation (or rules for selecting subgraphs that are not be cross-multiplied).

Unfortunately, as no suitable ontology of compounds yet exists in the public domain, there is no alternative to the present method of maintaining this part of the biological_process ontology by hand.

A very similar situation exists for anatomical terms, in effect used as anatomical qualifiers to terms in the biological_process ontology. An example is eye morphogenesis, a term that can be broken up into an anatomical component (eye) and a process component (morphogenesis). This example illustrates a further problem, we clearly need to be able to distinguish the morphogenesis of a fly eye from that of a murine eye, or a *Xenopus* eye, or an acanthocephalan eye (were they to have eyes). Such is not the way to maintain an ontology. Far better would be to have species- (or clade-) specific anatomical ontologies and then to generate the required terms for biological_process as cross-products. This is indeed the way in which GO will proceed (Hill et al 2002) and anatomical ontologies for *Drosophila* and *Arabidopsis* are already available from the GO Consortium (<ftp://ftp.geneontology.org/pub/go/anatomy>), with those for mouse and *C. elegans* in preparation (see Bard & Winter 2001, for a discussion). The other advantage of this approach is that these anatomical ontologies can then be used in other contexts, for example for the description of expression patterns or mutant phenotypes (Hamsey 1997).

gobo: global open biological ontologies

Although the three controlled vocabularies built by the GO Consortium are far from complete they are already showing their value (e.g. Venter et al 2001, Jenssen et al 2001, Laegreid et al 2002, Pouliot et al 2001, Raychaudhuri et al 2002). Yet, as discussed in the preceding paragraphs the present method of building and maintaining some of these vocabularies cannot be sustained. Both for their own use, as well as the belief that it will be useful for the community at large, the GO Consortium is sponsoring gobo (global open biological ontologies) as an umbrella for structured controlled vocabularies for the biological domain. A small ontology of such ontologies might look like this:

```
gobo
  gene
    gene_attribute
      gene_structure
      gene_variation
  gene_product
    gene_product_attribute
      molecular_function
      biological_process
      cellular_component
```

```

    protein_family
chemical_substance
  biochemical_substance
    class
    biochemical_substance_attribute
  pathway
    pathway_attribute
developmental_timeline
anatomy
  gross_anatomy
  tissue
  cell_type
phenotype
  mutant_phenotype
  pathology
  disease
  experimental_condition
  taxonomy

```

Some of these already exist (e.g. Taxman for taxonomy; Wheeler et al 2000) or are under active development (e.g. the MGED ontologies for microarray data description; MGED 2001), a trait ontology for grasses (GRAMENE 2002) others are not. There is everything to be gained if these ontologies could (at least) all be instantiated in the same syntax (e.g. that used now by the GO Consortium or in DAML+OIL; Fensel et al 2001); for then they could share software, both tools and browsers, and be more readily exchanged. There is also everything to be gained if these are all open source and agree on a shared namespace for unique identifiers.

GO is very much a work in progress. Moreover, it is a community rather than individual effort. As such, it tries to be responsive to feedback from its users so that it can improve its utility to both biologists and bioinformaticists, a distinction, we observe, that is growing harder to make every day.

Acknowledgements

The Gene Ontology Consortium is supported by a grant to the GO Consortium from the National Institutes of Health (HG02273), a grant to FlyBase from the Medical Research Council, London (G9827766) and by donations from AstraZeneca Inc and Incyte Genomics.

The work described in this review is that of the Gene Ontology Consortium and not the authors—they are just the raconteurs; they thank all of their colleagues for their great support. They also thank Robert Stevens, a user-friendly artificial intelligencer, for his comments and for providing references that would otherwise have evaded them; MA thanks

Donald Michie for introducing him to *WordNet*, albeit over a rather grotty Chinese meal in York.

References

- Adams M, Celniker SE, Holt RA et al 2000 The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ 1990 Basic local alignment search tool. *J Mol Biol* 215:403–410
- AmiGO 2001 url: www.godatabase.org/cgi-bin/go.cgi
- Ashburner M, Ball CA, Blake JA et al 2000 Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25:25–29
- Baker PG, Goble CA, Bechhofer S, Paton NW, Stevens R, Brass A 1999 An ontology for bioinformatics applications. *Bioinformatics* 15:510–520
- Bard J, Winter R 2001 Ontologies of developmental anatomy: their current and future roles. *Brief Bioinform* 2:289–299
- Commission of Plant Gene Nomenclature 1994 Nomenclature of sequenced plant genes. *Plant Molec Biol Rep* 12:S1–S109
- Cruse DA 1986 *Lexical semantics*. New York, Cambridge University Press
- DAG Edit 2001 url: sourceforge.net/projects/geneontology/
- DiBona C, Ockman S, Stone M (eds) 1999 *Open sources: voices from the Open Source revolution*. O'Reilly, Sebastopol, CA
- Dure L III 1991 On naming plant genes. *Plant Molec Biol Rep* 9:220–228
- Dwight SS, Harris MA, Dolinski K et al 2002 *Saccharomyces* Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res* 30:69–72
- Fellbaum C (ed) 1998 *WordNet*. An electronic lexical database. MIT Press, Cambridge, MA
- Fensel D, van Harmelen F, Horrocks I, McGuinness D, Patel-Schneider PF 2001 OIL: An ontology infrastructure for the semantic web. *IEEE (Inst Electr Electron Eng) Intelligent Systems* 16:38–45 [url: www.daml.org]
- Fleischmann RD, Adams MD, White O et al 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269:496–512
- The FlyBase Consortium 2002 The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res* 30:106–108
- The Gene Ontology Consortium 2001 Creating the gene ontology resource: design and implementation. *Genome Res* 11:1425–1433
- GRAMENE 2002 Controlled ontology and vocabulary for plants. url: www.gramene.org/plant_ontology
- Hamsey M 1997 A review of phenotypes of *Saccharomyces cerevisiae*. *Yeast* 1:1099–1133.
- Heath P 1974 (ed) *The philosopher's Alice*. Carroll L, Alice's adventures in wonderland & through a looking glass. Academy Editions, London
- Hill DP, Davis AP, Richardson JE et al 2001 Program description: strategies for biological annotation of mammalian systems: implementing gene ontologies in mouse genome informatics. *Genomics* 74:121–128
- Hill DP, Richardson JE, Blake JA, Ringwald M 2002 Extension and integration of the Gene Ontology (GO): combining GO vocabularies with external vocabularies. *Genome Res*, in press
- Karp PD 2000 An ontology for biological function based on molecular interactions. *Bioinformatics* 16:269–285
- Karp PD, Riley M, Saier M et al 2002a The EcoCyc database. *Nucleic Acids Res* 30:56–58

- Karp PD, Riley M, Parley SM, Pellegrini-Toole A 2002b The MetaCyc database. *Nucleic Acids Res* 30:59–61
- Kerlavage A, Bonazzi V, di Tommaso M et al 2002 The Celera Discovery system. *Nucleic Acids Res* 30:129–136
- Kretschmann E, Fleischmann W, Apweiler R 2001 Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics* 17:920–926
- Kyrpides NC, Ouzounis CA 1998 Whole-genome sequence annotation 'going wrong with confidence'. *Molec Microbiol* 32:886–887
- Laegreid A, Hvidsten TR, Midelfart H, Komorowski J, Sandvik AK 2002 Supervised learning used to predict biological functions of 196 human genes. Submitted
- Leser U 1998 Semantic mapping for database integration — making use of ontologies. url: cis.cs.tu-berlin.de/~leser/pub_n_pres/ws_ontology_final98.ps.gz
- MGED 2001 Microarray Gene Expression Database Group. url: www.mged.org
- Mewes HW, Heumann K, Kaps A et al 1999 MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 27:44–48
- Mi H, Vandergriff J, Campbell M et al 2002 Assessment of genome-wide protein function classification for *Drosophila melanogaster*. Submitted
- Miller GA 1998 Nouns in WordNet. In: Fellbaum C (ed) WordNet. An electronic lexical database. MIT Press, Cambridge, MA, p 23–46
- OpenSource 2001 url: www.opensource.org/
- Overbeek R, Larsen N, Smith W, Maltsev N, Selkov E 1997 Representation of function: the next step. *Gene* 191:GC1–GC9
- Overbeek R, Larsen N, Pusch GD et al 2000 WIT: integrated system for high-level throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res* 28:123–125
- Pouliot Y, Gao J, Su QJ, Liu GG, Ling YB 2001 DIAN: a novel algorithm for genome ontological classification. *Genome Res* 11:1766–1779
- Priss UE 1998 The formalization of WordNet by methods of relational concept analysis. In: Fellbaum C (ed) WordNet. An electronic lexical database. MIT Press, Cambridge, MA, p 179–190
- Pruitt KD, Maglott DR 2001 RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res* 29:137–140
- Raychaudhuri S, Chang JT, Sutphin PD, Altman RB 2002 Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res* 12:203–214
- Riley M 1988 Systems for categorizing functions of gene products. *Curr Opin Struct Biol* 8: 388–392
- Riley M 1993 Functions of the gene products of *Escherichia coli*. *Microbiol Rev* 57:862–952
- Rison SCG, Hodgman TC, Thornton JM 2000 Comparison of functional annotation schemes for genomes. *Funct Integr Genomics* 1:56–69
- Rogers JE, Rector AL 2000 GALEN's model of parts and wholes: Experience and comparisons. Annual Fall Symposium of American Medical Informatics Association, Los Angeles. Hanley & Belfus Inc, Philadelphia, CA, p 714–718
- Schulze-Kremer S 1997 Integrating and exploiting large-scale, heterogeneous and autonomous databases with an ontology for molecular biology. In: Hofstaedt R, Lim H (eds) Molecular bioinformatics — The human genome project. Shaker Verlag, Aachen, p 43–46
- Schulze-Kremer S 1998 Ontologies for molecular biology. *Proc Pacific Symp Biocomput* 3: 695–706
- Serres MH, Riley M 2000 Multifun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products. *Microb Comp Genomics* 5:205–222

- Serres MH, Gopal S, Nahum LA, Liang P, Gaasterland T, Riley M 2001 A functional update of the *Escherichia coli* K-12 genome. *Genome Biol* 2:RESEARCH 0035
- Sklyar N 2001 Survey of existing Bio-ontologies. url: <http://dol.uni-leipzig.de/pub/2001-30/en>
- Stevens R, Baker P, Bechhofer S et al 2000 TAMBIS: transparent access to multiple bioinformatics information sources. *Bioinformatics* 16:184-183
- Takai-Igarashi T, Nadaoka Y, Kaminuma T 1998 A database for cell signaling networks. *J Comp Biol* 5:747-754
- Van Helden J, Naim A, Lemer C, Mancuso R, Eldridge M, Wodak SJ 2001 From molecular activities and processes to biological function. *Brief Bioinform* 2:81-93
- Venter JC, Adams MD, Meyers EW et al 2001 The sequence of the human genome. *Science* 291:1304-1351
- Wheeler DL, Chappey C, Lash AE et al 2000 Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res* 28:10-14
- Winston ME, Chaffin R, Herrman D 1987 A taxonomy of part-whole relations. *Cognitive Sci* 11:417-444
- Xie H, Wasserman A, Levine Z et al 2002 Automatic large scale protein annotation through Gene Ontology. *Genome Res* 12:785-794
- Zdobnov EM, Apweiler R 2001 InterProScan — an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847-848

DISCUSSION

Subramaniam: Sometimes cellular localization drives the molecular function. The same protein will have a particular function in certain places and then when it is localized somewhere else it will have a different function.

Aspburner: I thought about doing this at the level of annotation, in which you could have a conditionality attached to the annotation. I have been lying during my talk, because I have been talking about annotating gene products. For various reasons — partly historical and partly because of resources — none of the single model organism databases we are collaborating with (at least in their public versions) really instantiate gene products in the proper way. That is, if you had a phosphorylated and a non-phosphorylated form of a particular protein, they should have different identifiers and different names. This is what we should be annotating. What in fact we are annotating is genes as surrogates of gene products. I am very aware of this problem. With FlyBase we do have different identifiers for isoforms of proteins, and in theory for different post-translational modifications, but they are not yet readily usable. The difficult ones are proteins such as NF- κ B, which is out there in the cytoplasm when it is bound to IF- κ B, but then the Toll pathway comes and translocates it into the nucleus. I can see theoretically how one can express this, but this is a problem too far at the moment.

Subramaniam: MySQL is not really an object relation database. If you try to get your ontology into an object relation database (we have tried to do this) the cardinality doesn't come out right. What happens is that the definitions get a

little bit mixed up between different tables. This is one of the problems in trying to deal with Oracle.

Asburner: That is worth knowing; we can talk to the database people about that. The choice of MySQL was pragmatic.

Subramaniam: Also, MySQL doesn't scale.

Asburner: These are pretty small databases, with a few thousand lines per table and relatively small numbers of tables.

McCulloch: What degree of interpretation do you allow, for example, in compartmentation of the protein? If you go to the original paper it won't necessarily say that the protein is membrane bound or localized to caveolae: it will probably say that it is found in a particulate fraction, or the detergent-insoluble fraction.

Asburner: We do have a facility for allowing curators to add biochemical fraction information, because biochemists tend not to understand biology that well. I want to emphasize that GO is very pragmatic, although there are places where we are going to have to draw a line.

Noble: In relation to the question of linking modelling and databases together, is it worth asking the question of what the modellers would ideally like to see in a database? Does the GO consortium talk to the modellers?

Asburner: We have a bit. There are some people who are beginning to do this, particularly Fritz Roth at Harvard Medical School. We have a mechanism by which we can talk to the modellers because we have open days. There are other systems out there such as EcoCyc (<http://ecocyc.org/>) that are designed with modelling in mind, for making inference. GO isn't; it's designed for description and querying. I think it will come. GO is being used in ways that we had no concept of initially. For instance, it is being developed for literature mining (see Raychadhuri et al 2002). This could be very interesting.

Kanehisa: When there is the same GO identifier in to organisms, how reliable is it in terms of the functional orthologue?

Asburner: That depends very much on how it is done. It is turning out that when a new organism joins the group, what is normally done is a quick-pass electronic annotation using the annotation in SWISS-PROT. This is done completely electronically, and gives a quick and dirty annotation. Then if they have the resources the groups start going through this and cleaning it up, hopefully coming up with direct experimental evidence for each annotation. For example, after Celera we had about 10 000 electronic annotations in FlyBase, but these have all been replaced by literature curations or annotations derived from a much more reliable inspection of sequence similarity.

Subramaniam: Going back to the issue of ontologies and databases, it is important to ask the question about which levels of ontologies can translate into modelling. If you think of modelling in bioinformatics and computational

biology, the flow of information in living systems is going from genes to gene products to functional pathways and then physiology. What we have heard from Michael Ashburner is concerned with the gene and gene function level. The next step is what we are really referring to, which is not merely finding an ontology for the gene function, but going beyond this to integrated function, or systems level function of the cell. There is currently no ontology available at this level. This is one of the issues we are trying to address in the cell signalling project; it is critical for the next stage of modelling work. This has to be driven at this point: whether or not you make the reverse ontology, at least you should provide format translators such as XML.

Ashburner: GO, of course, is sent around the world in XML.

Noble: How do we move forward on this? A comment you made surprised me: I think you said that it is forbidden to modify GO.

Ashburner: No, it is forbidden to modify it and then sell it as if it were GO. If you took it, modified it and called it 'Denis Noble's ontology', we would be at least mildly pissed off.

Subramaniam: We could call it 'extended GO', so that it becomes 'EGO'!

Ashburner: The Manchester people (C. Groble, R. Stevens and colleagues) have something called GONG: GO the Next Generation!

Boissel: Regarding the issue of databases and modelling, we should first be clear about the functions of the database regarding the purpose of modelling. According to the decision we have made at this stage of defining the purpose of the database, there is a series of specifications. For example, a very general specification such as entities, localization of entities, relationship between entities, and where the information comes from (including the variability of the evidence). There are at least four different chapters within the specification. But first we should be clear why we are constructing a database regarding modelling.

Subramaniam: Let's take specific examples. If you talk about pathway ontology, what are you getting from a pathway database? The network topology. And sometimes kinetic parameters, too. All this will be encompassed in the database and can be translated into modelling. Having said this, we should be careful about discriminating between two things in the database. First, the querying of the database to get information that in turn can be used for modelling. The other is going straight from a database into a computational algorithm, and this is precisely what needs to be done. This is why earlier I said that we currently can't do this in a distributed computing environment. The point really is that we need to be able to compute, instead of having to write all our programming in SQL, which we won't be able to do if we have a complex program. We need to design a database so that it will enable us to communicate directly between the database and our computational algorithm. Beyond the pathway level, when we want to model the

whole system, I don't know whether anyone knows how to do this from a database point of view yet.

Berridge: Say we were interested in trying to figure out the pathways in the heart, and I put 'heart' into your database, what would I get out?

Asbburner: At the moment, whatever the mouse genome informatics group have put in.

Berridge: Would I get a list of all the proteins that are expressed in the heart?

Asbburner: No, but you should get a list of all the genes whose products have been inferred to be involved in heart development, for example. The physiological processes are not yet as well covered in GO as we wish, but we are working on this actively.

Noble: So even if it is expressed in the liver, but it affects the heart, it turns up.

Asbburner: Yes.

Berridge: What questions will people be asking with your database?

Asbburner: If you want to find all the genes in *Drosophila* and mouse involved in a signal transduction pathway, for example. It can't predict them: what you get out is what has been put in. The trick is to add the entries in a rigorous manner.

Berridge: So if I put in Ras I would get out the MAP kinase pathway in these different organisms.

Asbburner: Yes.

Levin: Looking higher than the level of the pathway, you indicated that there were no good disease-based databases in the public domain. Can you give a sense of why this is?

Asbburner: I have no idea. They exist commercially: things like Snomed and ICD-10. Some are now being developed. I suspect this is because so much of the human anatomy and physiology work has been so driven by the art of medicine, rather than the science of biomedicine. Doctors are quite avaricious as a whole, particularly in the USA, and many of these databases are used to ensure correct billing!

Reference

Raychaudhuri S, Chang JT, Sutphin PD, Altman RB 2002 Associating genes with gene ontology codes using a maximum entropy analysis of biomedical literature. *Genome Res* 12:203–214

General discussion I

Model validation

Paterson: One of the challenges in model validation is that unless you have a particular purpose in mind, it can turn into a largely academic conversation about what is meant by validation. In a lot of the applied work we do, it is in the context of making predictions for decision making that validation really comes into its own. I would like to introduce a few concepts and then open things up for discussion (see Fig. 1 [*Paterson*]). In the context of validating a model, we are talking about linking detailed mechanisms to observed phenomena. As all of us in this field know that there are always gaps in our knowledge, even if we are talking about parametric variations within a set of equations. For each of these knowledge gaps, there are multiple hypotheses that may be equally valid, and explain the same phenomena. The question is, each of these hypotheses may yield different predictions for novel interventions, which may then lead me to different decisions. If we think about *in silico* modelling as an applied discipline, one central issue is communicating this reality, and how to manage it properly, to the decision makers. Typically, the modelling teams — the people who understand these issues — are separate from the people who have the resources to decide which is the next project to fund, or in pharmaceutical applications what is the next target to pursue. These two groups may have very different backgrounds, which raises further issues of communication. It is therefore necessary to explain why you have confidence in the model, what you think are the unanswered questions, and the implications of both to upcoming decisions. It is certainly in the context of the resources and time when all this comes into play. If the resources concerned are small and the time it takes to go through an iteration of modelling and data collection are small, such explorations may fit easily within budgets and timelines. However, as you consider applications in the pharmaceutical industry, we are talking about many millions of dollars worth of resources, and years of preclinical and clinical research time. These issues of validation and uncertainty when model predictions are used to support decision-making have driven our approach to modelling. I would be interested in whether there are any perspectives people can share in terms of how they approach modelling as a discipline.

Noble: Let me give you a reaction from the point of view of the academic community. It seems to me that this issue links strongly to the issue of the availability of models to be used by those who are not themselves primarily

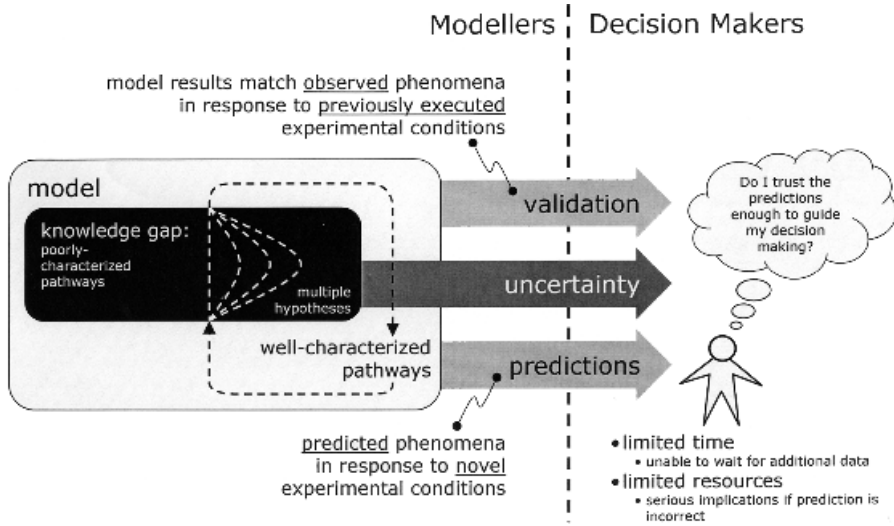


FIG. 1. (Paterson) Validation and uncertainty are key issues when model predictions are used to support decision making.

modellers. In other words, it gets back to this issue of getting the use of models out there among people who are themselves experimentalists. The experience I have is that it is only when people get hands on, where they can feel and play, that they start to get confidence, and that they can get some good explanations of their own data and that the model will help them decide on their next experiment.

Paterson: One complication to your scenario arises from the integrated nature of these models and the diverse expertise represented within them. As the scope of an integrated physiology model increases, the number of researchers that understand that entire scope dwindles. What can happen is that such a model may be used by you as a researcher and your research may be focused on the biology in this one area of the model, but you may be very unfamiliar with the other subsystems. In terms of the data you care about, this model may replicate these data and it is therefore validated from your perspective. However, the context of the other subsystems that you don't have expertise in may be very relevant to those predictions and decisions that will be guided as a result. Part of what the modeller needs to communicate is expertise that may lie beyond the expertise of the researcher using the model.

Noble: I wasn't of course implying that the experimenter who takes modelling on and starts to play around wants to cut links with the experts, as it were.

Loew: I can see where these validation issues are very critical: you need to have a certain amount of confidence in the model before you can go on to influence

decision making in choosing which drugs to take to clinical trials. But from the point of view of an academic modeller, a model is actually most useful if you can prove it to be wrong. You can never prove it to be right. The simplest example here is classical mechanics versus quantum mechanics. Classical mechanics is a very useful model that allows you to make many predictions and decisions, but it was most spectacular when it was proved to be wrong, and in understanding the limits of where it was wrong. This is how the science progresses. From a practical point of view, classical mechanics is great, but from an academic point of view it is really great when you can prove it wrong.

Boissel: We should be careful not to confuse model validation and model dissemination. Getting people to trust and use the model is not the same as model validation. Regarding whether a model is wrong or not, a model is always wrong. The problem is determining just how wrong it is and in which contexts it is right.

Noble: I agree that all models are inevitably wrong, because they are always only partial representations of reality.

McCulloch: Tom Paterson, your diagram does resonate with the academic way of doing things, where the result of the model is really a new hypothesis or set of hypotheses that the decision maker can use to design to new experiment or write a new grant. But is this really the way it works in the pharmaceutical industry? It seems unlikely that the pharmaceutical industry would make a go/no-go decision based on the predictions of a computational model. By the time they are willing to invest large resources, they already have strong proof of principle. The go/no-go decisions are presumably based on more practical considerations. For example, does the antibody cross react in humans? Is the lead compound going to be orally bioavailable? How does the patent position sit? Are there examples today in the pharmaceutical industry where resources are being committed on the basis of *in silico* predictions?

Paterson: Yes, there are many. The key point in answering your question is that it is always the case that pharmaceutical research and development decisions are made under uncertainty about the real causal factors underlying disease pathophysiology. The question is whether they have leveraged all the relevant data to reduce that uncertainty using the model in their heads, or whether they use a computational model. What we are doing isn't that different from the normal scientific process; we are just using different languages, i.e. graphical notations and mathematics, to articulate our hypotheses and to test their consistency. Part of the reason I drew that dotted line in the diagram, which is very critical, was that communication. Why do we have confidence in the model, what are the validation steps, and what are we uncertain about? Research and development decision-making, in general, is less rational and concrete than you would think. Proper use of models can improve

decision-making, and is doing so, but communication of those issues to decision makers is critical.

Subramaniam: Where does the quaint notion of doing a sensitivity analysis of every design step and every parameter come into your diagram?

Paterson: Sensitivity analysis is one way of looking at uncertainty, although it is complicated by the need to remain consistent with the constraints imposed by data. For a particular decision that I am making, there is a certain set of data that our scientist will identify and we will say that we are only going to trust the model if it behaves in these ways under these circumstances consistent with these data. In effect, we define a validation set of experiments that the model needs to perform. Part of what we need to do then, is out of this very large parameter or model space that exists, and given limited time and resources, ask how we can explore this parameter space, given that we may be uncertain about many of those parameters due to the limited availability of data. We may have many competing hypotheses that can be represented within this particular space about how a set of pathways is regulated. As we explore these different pathways we need to ask whether competing hypotheses would make us change our decision. If they don't, then we don't need to invest resources in resolving this uncertainty. If, however, the choice between hypotheses A and B would change our decision, then this is the experiment we want to run.

Subramaniam: Do people routinely do this in industry?

Paterson: No; this is extremely difficult using mental models. My organization is doing this using the models we develop.

Levin: This is not quite correct as there is increasing and routine use of biological modelling in some areas of industry. Models of absorption and metabolism are widely distributed, but they answer very particular, limited questions. The problem that Tom has identified of communicating the value of simulation within an organization is a significant one. The line he describes is less a dotted one than a lead shield in the very traditional pharmaceutical companies. What defines a flexible and innovative organization is one that understands how to cope with transferring new technology while educating its personnel and developing the right management structures to enable and empower change. One second point: the issue of uncertainty and sensitivity analysis is an important one. The question of validation is one that will bedevil many organizations until they understand and learn how to weld biology (in the form of the day-to-day experimentation), fundamental motif formation (at a module level with practical tools at the bench), and then development of protocols to generate appropriate experimental data to iterate between the module and the desired hypothesis. I disagree here with what I think I heard from Jean-Pierre Boissel, in that I think dissemination of a model is linked to validation, but dissemination of the tools and modelling is linked to an understanding of how to link experimentation to models and motif.

Boissel: For sure, you cannot disseminate a model without good validation.

Levin: Another complex issue which is particular to efforts to disseminate models within a large organization (or between organizations and multiple people) is to ensure that you are speaking the same language (using the same ontologies), and you actually have interchangeable models based on common technology and hence permitting researchers to make comparisons. All of these make that dotted line more difficult to cross.

Paterson: Part of the key for adoption is to recognize that there isn't anything we are talking about in this room today that creates that problem. That problem has existed since the pharmaceutical industry began. It is not data that drives decision making, but hypotheses for exploring novel therapeutics. The issue is, whether that hypothesis of the pathophysiology of the disease and the relevance of a particular novel target was developed as part out of a modelling exercise. It is still this process. The promise of what modelling can do is that it makes it more explicit.

Levin: The problems facing those engaged in developing and promulgating modelling are no different from the problems that others developing novel technologies have faced when providing them to the pharmaceutical industry. The line distinguishing decision makers from the (generally younger) scientists at the bench has been there from the start. Whether it be a combinatorial chemist or a genomic scientist — each have faced this in their time, and each have sequentially overcome the managerial resistance in some fashion. In some cases dynamic leadership breaks the ice. But eventually, each segment of science has a particular way of solving the issue. All must overcome similar questions, such as: is this a valid technology, what are the uncertainties relating to it, and how will it affect my decision making? Biology has arrived at a state where there are no easy ways to answer the huge volume of questions precipitated by the genome project and its attendant deluge of data. We no longer can afford to think in the terms that we have done for the last 30 years. We need to solve some very complex high-throughput problems which rest on integrating all of the data and seeking emergent properties. Hypothesis generation of the kind that modelling offers is at least one way of dealing with some key questions that are emerging because of the nature of the pharmaceutical industry. Often, 14 years pass between the initiation and culmination of a project (the release of a new drug), and there is a pipeline of thousands of compounds that have been developed using standard practices and processes. We already know that the overwhelming majority of these compounds will fail to become drugs. By incorporating and modelling the emerging body of data pertaining to the molecular and cell biology function of these compounds, we have a better chance to explain to and point people to where those compounds are likely to succeed.

Boissel: I think we need some type of good model validation practices in order to make our activity more positive for the people who can use it. We need to agree on a series of principles regarding how models should be validated.

Noble: One of the criteria that I would put in would be the number of times that there has been iteration between model and experiment.

Boissel: This is external validity. We also need some principles regarding internal validity. In any validation process, there are three different steps. The first step is to investigate the internal validity, the second is the external validity and the final one is to look at how well the model predictions match what we would have expected at the beginning. The internal validity is whether the model has integrated all the data that we wanted to put in, and really translated what we know in terms of quantitative relationships between the entities and so on. The external validity is what you propose: is the model valuable regarding the data and knowledge which have not been added to it?

Cassman: A model isn't just a representation of elements of some sort, but rather is an embodiment of a theory. There is a long history of how we validate theories. I don't see why it is any different for models than for anything else. Karl Popper has listed characteristics of what constitute good theories: the breadth of information that they incorporate, the relevance to a large set of outcomes, and most importantly predictive value. I don't know that there is anything unique about models as a theory than any other theories. They should be dealt with the same way.

Paterson: There is at least one unique dimension that the quantitative nature of models enables. Particularly when you are talking about developing novel therapies, it is not enough to identify that a particular protein is in a pathway for the disease; you need to know how much leverage it actually has. If I am going to inhibit that protein's activity by 50%, how much of an improvement in the clinical endpoint will I have? Quantitatively, these things make a difference. Even for a single set of equations, the degrees of freedom that you have in the parametric space for complex models relative to the constraints that are imposed by the data is always going to be huge. It is incumbent upon the modeller to explore that uncertainty space, and there are huge benefits to doing this. Instead of giving you one hypothesis I am going to give you a family of hypotheses, all of which have been thoroughly tested for consistency with available data. Different hypotheses may lead to different decision recommendations. In this way, you simultaneously have the opportunity to help make more informed decisions, and if there are time and resources to collect more data you can help identify what is the most important experiment to run. Instead of giving one hypothesis, we give alternatives and show the relevance of these to the decision being made.

Shimizu: One thing I disagree with in your diagram is that it appears to separate the predictions from the validation. I think these are really closely intertwined.

Noble: It's an iteration.

Paterson: Yes, it is completely iterative. But in industry there comes a point where there is no more time for iterations and a decision has to be made. So I have to go with the predictions that come out of the model, or the predictions that come out of the heads of my best researchers in order to push things forwards. At some point the iteration needs to stop.

Shimizu: When I said that they were intertwined, I didn't just mean it in an iterative sense. Of course in general, the more you refine a model by iteration, the better you can expect its predictions to be. But I would call this the accuracy of the model, rather than its validity. The term validity, I believe, should be reserved for discerning whether the type of model you are using is fit to make the desired predictions. In simulating chemical reactions, for example, a set of deterministic equations that beautifully predicts the behaviour of a reaction system might be called a valid model. But there are situations in which such a model can fail. For instance, if you are interested in predicting the behaviour of the same chemical system in a very small volume of space, the behaviour of the system can become stochastic, in which case the deterministic model will break down. So my point is that from the decision maker's point of view, I don't think it's a good idea to have the validation part just as a black box that gives a yes/no result.

Paterson: Absolutely not. You want the decision makers to help you define what the validation criteria are. You also want the decision maker to play a role in what uncertainties you explore and to see how sensitive they are.

McCulloch: If Marv Cassman is correct and logical positivism is the paradigm by which models are best used, this would predict that the decision makers would rely on the models mostly when they decided not to proceed. Is this the case?

Noble: Most grants are turned down, so it must be!

Subramaniam: Falsification is not the only criterion.

McCulloch: Very well, allow me to rephrase the question. Is there an asymmetry in the way that decision makers use the predictions of models? Are they more inclined to accept the model conclusion that it is not going to work than it is?

Paterson: In our experience, as we explore the uncertainty side of the equation to address the robustness, it has probably been easier to show a very robust answer that things will not work versus an extremely robust answer that it will certainly work. In terms of where the pharmaceutical industry is today, in a target-rich environment, then anything you can do to help avoid clinical trial failure by anticipating issues early on is a significant contribution.

The KEGG database

Minoru Kanehisa

*Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji,
Kyoto 611-0011, Japan*

Abstract. KEGG (<http://www.genome.ad.jp/kegg/>) is a suite of databases and associated software for understanding and simulating higher-order functional behaviours of the cell or the organism from its genome information. First, KEGG computerizes data and knowledge on protein interaction networks (PATHWAY database) and chemical reactions (LIGAND database) that are responsible for various cellular processes. Second, KEGG attempts to reconstruct protein interaction networks for all organisms whose genomes are completely sequenced (GENES and SSDB databases). Third, KEGG can be utilized as reference knowledge for functional genomics (EXPRESSION database) and proteomics (BRITE database) experiments. I will review the current status of KEGG and report on new developments in graph representation and graph computations.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 91–103

The term 'post-genomics' is used to refer to functional genomics and proteomics experiments after complete sequencing of the genome, such as for analysing gene expression profiles, protein–protein interactions and 3D protein structures. Systematic experiments have become possible through the development of high-throughput experimental technologies including DNA chips and protein chips. However, the complete cataloguing of genes and proteins by these experimental approaches is only a part of the challenge in the post-genomic era. As illustrated in Fig. 1, a huge challenge is to predict a higher-level biological system, such as a cell or an organism, from genomic information, as is predicting dynamic interactions of the system with its environment (Kanehisa 2000). We have been developing bioinformatics technologies for deciphering the genome in terms of the biological system at the cellular level; namely, in terms of systemic functional behaviours of the cell or the single-celled organism. The set of databases and computational tools that we are developing is collectively called KEGG (Kyoto Encyclopaedia of Genes and Genomes) (Kanehisa 1997, Kanehisa et al 2002).

The databases in KEGG are classified into three categories corresponding to the three axes in Fig. 1. The first category represents parts-list information about genes

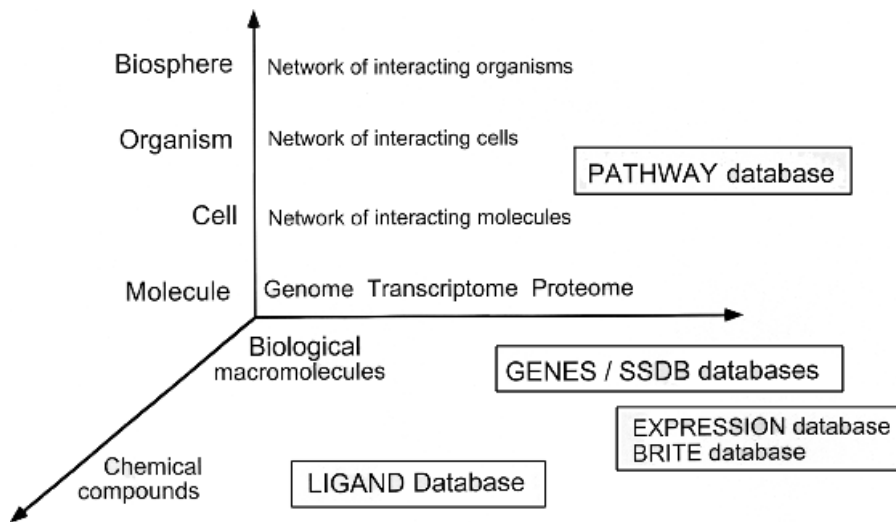


FIG. 1. Post-genomics and KEGG.

and proteins. The gene catalogues of all publicly available complete genomes and some partial genomes are stored in the GENES database, which is a value-added database containing our assignments of EC (Enzyme Commission) numbers and KEGG orthologue identifiers as well as links to SWISS-PROT and other databases. Selected experimental data on gene expression profiles (from microarrays) and protein–protein interactions (from yeast two-hybrid systems) are stored in the EXPRESSION and BRITE databases, respectively. In addition, the sequence similarity relations of all protein-coding genes in the GENES database are computationally generated and stored in the SSDB database. The second category represents computerized knowledge on protein interaction networks in the cell, such as pathways and complexes involving various cellular processes. The networks are drawn by human efforts as graphical diagrams in the PATHWAY database. The third category represents chemical information. The LIGAND database contains manually entered entries for chemical compounds and chemical reactions that are relevant to cellular processes. Chemical compounds include metabolites and other compounds within the cell, drugs, and environmental compounds, while chemical reactions are mostly enzymatic reactions.

Graph representation

A graph is a mathematical object consisting of a set of nodes (vertices) and a set of edges. It is general enough to represent various objects at different levels of abstraction. For example, a protein molecule or a chemical compound can be

Graph objects at different levels of abstraction

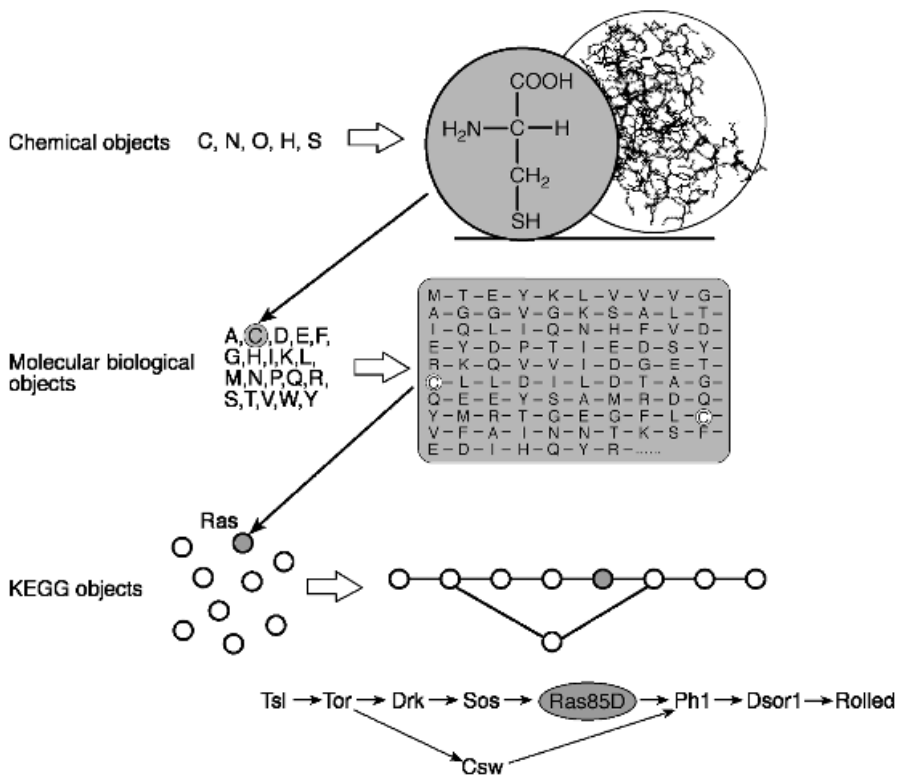


FIG. 2. Graph objects at different levels of abstraction.

viewed as a chemical object, which is represented as a graph consisting of atoms as nodes and atomic interactions as edges. A protein sequence or a DNA sequence can be viewed as a molecular biological object, which is represented as a graph consisting of monomers (amino acids or nucleotides) as nodes and covalent bonds for polymerization (peptide bonds or phosphodiester bonds) as edges. As illustrated in Fig. 2, a molecular biological object is at a higher level of abstraction than a chemical object, because the graph of a chemical object, such as an amino acid, is considered as a node in a molecular biological object. Then, at a still higher level of abstraction, the graph of a molecular biological object can be considered as a node in, what we call, a KEGG object. A KEGG object thus represents interactions and relations among proteins or genes.

Computational technologies are relatively well developed for analysing the molecular biological objects of sequences and the chemical objects of 3D

TABLE 1 KEGG objects representing interactions and relations among genes and proteins

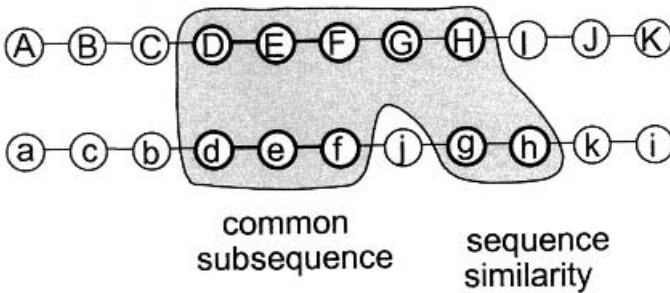
<i>Database</i>	<i>KEGG object</i>	<i>Node</i>	<i>Edge</i>
GENES	Genome	Gene	Adjacency
EXPRESSION	Transcriptome	Gene	Expression similarity
BRITE	Proteome	Protein	Direct interaction
SSDB	Protein universe	Protein	Sequence similarity (orthology, etc.) 3D structural similarity
PATHWAY	Network	Gene product or subnetwork	Generalized protein interaction (direct interaction, gene expression relation, or enzyme–enzyme relation)
LIGAND	Chemical universe	Compound	Chemical reaction

structures, largely because the databases are well developed: GenBank/EMBL/DDBJ for DNA sequences; SWISS-PROT for protein sequences; and PDB for protein 3D structures, among others. In order to analyse higher-level interactions and relations among genes and proteins, it is extremely important to first computerize relevant data and knowledge and then to develop associated computational technologies. KEGG aims at a comprehensive understanding of interaction networks of genes, proteins, and compounds, based on graph representation of biological objects (see Table 1 for the list of KEGG objects), and graph computation technologies (Kanehisa 2001).

Graph computation

The graph computation technologies of interest to us are extensions of the traditional technologies for sequence and 3D structure analyses. First, the sequence comparison and the 3D structure comparison are generalized as the graph comparison, which is utilized to compare two or more KEGG objects in Table 1 for understanding biological implications. Second, feature detection — e.g. for sequence motifs or 3D structure motifs — can be extended as the graph feature detection, which is utilized to analyse a single graph to find characteristic connection patterns, such as cliques, that can be related to biological features. Third, the big challenge of network prediction, which is to predict the entire protein interaction network of the cell from its genome information, can be compared in spirit with the traditional structure prediction problem, which involves predicting the native 3D structure of a protein from its amino acid sequence.

Sequence comparison



Graph comparison

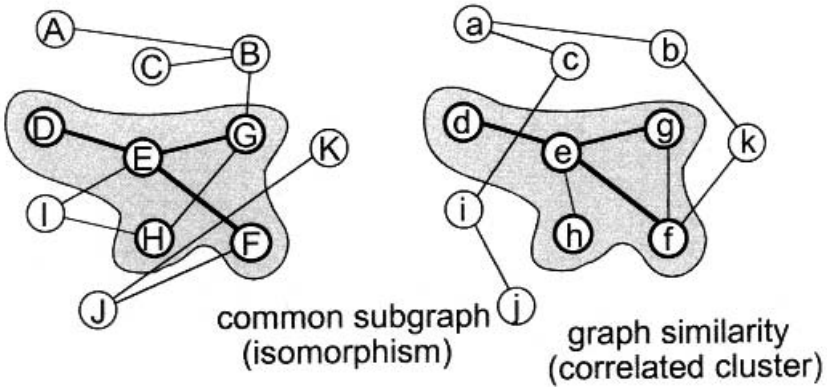


FIG. 3. Sequence comparison and graph comparison.

A simple way to compare two sequences is to search for common sub-sequences. In Fig. 3 two sequences of alphabet letters, one in upper case and the other in lower case, are compared to identify case-insensitive matches. The common subsequence is DEF and def, which consist of a stretch of matching letters: D-d, E-e, and F-f. Note that in addition to the matching nodes (letters) the common subsequence implicitly contains matching edges, in this case DE-de and EF-ef. Now let us generalize the common subsequence in sequence comparison to the common subgraph in graph comparison, which is an isomorphic pair of subgraphs consisting of the same number of nodes connected in the same way. In Fig. 3 the common subgraph is (D, E, F, G, DE, EF, EG) and (d, e, f, g, de, ef, eg), which consist of matching nodes and edges: D-d, E-e, F-f, G-g, DE-de, EF-ef, and EG-eg.

TABLE 2 The top two levels of the KEGG network hierarchy

Metabolism	Environmental information processing
Carbohydrate metabolism	Membrane transport
Energy metabolism	Signal transduction
Lipid metabolism	Ligand–receptor interaction
Nucleotide metabolism	
Amino acid metabolism	Cellular processes
Metabolism of other amino acids	Cell motility
Metabolism of complex carbohydrates	Cell cycle and cell division
Metabolism of complex lipids	Cell death
Metabolism of cofactors and vitamins	Development
Metabolism of secondary metabolites	
Degradation of xenobiotics	Human diseases
	Neurodegenerative disorder
Genetic information processing	
Transcription	
Translation	
Sorting and degradation	
Replication and repair	

In practice, during protein and nucleic acid sequence comparisons perfect matches of common sub-sequences are too restrictive to identify interesting biological features. Thus, sequence comparison algorithms have been developed to find subtle sequence similarities containing gaps and mismatches. In the sequence comparison of Fig. 3, the two sub-sequences DEFGH and defjgh are similar by considering the node *j* as a gap. More precisely, in the subgraphs (D, E, F, G, H, DE, EF, FG, GH) and (d, e, f, j, g, h, de, ef, fj, jg, gh), the edge FG is matched to the pair of edges fj and jg to introduce a gap. In the graph comparison, the two subgraphs (D, E, F, G, I, H, DE, EF, EG, EI, IH) and (d, e, f, g, h, de, ef, eg, eh) are defined as similar subgraphs, because the pair of edges EI and IH can be matched to the edge eh to introduce a gap node I. We have developed a heuristic algorithm to find this type of graph similarity, which is called a correlated cluster (Ogata et al 2000).

Knowledge-based network prediction

The problem of 3D structure prediction has become feasible and practical because of the accumulated body of experimental data on actual protein 3D structures determined by X-ray crystallography and NMR. Empirical relationships between amino acid sequences and protein 3D structures have been analysed and utilized for prediction, for example, in terms of potential functions for threading and libraries of oligopeptide structures. The KEGG/PATHWAY database is our attempt to computerize current knowledge on protein interaction networks based on graph

representation and to understand empirical relations between genomes and networks based on graph computations.

The PATHWAY database is hierarchically categorized. The top two levels are shown in Table 2. The third level corresponds to a pathway diagram, such as the lysine biosynthesis pathway shown in Fig. 4. The pathway diagram represents a protein interaction network where gene products (proteins) are the nodes that are connected by three types of edges. The edge in the metabolic pathway is called the enzyme–enzyme relation consisting of two enzymes catalysing successive reaction steps. The other two types of edges are the direct protein–protein interaction (such as binding, phosphorylation, and ubiquitination), and the gene expression relation between a transcription factor and a target protein product. The protein interaction network is a compound (or nested) graph which allows nodes to contain graphs. For example, an enzyme complex with a single EC number is a node in the metabolic network but it is also a graph consisting of multiple gene products.

Each pathway diagram is manually drawn gathering knowledge and information from published literature. This reference knowledge can then be utilized for network prediction, as illustrated in Fig. 5. By matching genes in the genome and gene products in the reference network according to the assigned KEGG orthologue identifiers, an organism-specific network is computationally generated. Figure 4 is a result of this matching where the nodes (boxes) are shaded grey when genes are found in the genome, in this case in the *E. coli* genome. The connection pattern of coloured boxes then indicates the presence of the bacteria-type lysine biosynthesis pathway. Thus, given the reference network, a metabolic capability of the organism can be predicted from the genome information.

The knowledge-based prediction has an inherent limitation; when the reference knowledge does not exist, the prediction is not possible. To overcome this limitation, additional experimental data and/or computational results are incorporated in the prediction procedure as illustrated in Fig. 6. For example, the data obtained by yeast two-hybrid systems suggest possible protein–protein interactions, and the data obtained by microarrays suggest possible relations of coexpressed genes. These data are thus represented as a set of binary relations, which is essentially a graph. By making use of the graph comparison algorithm, multiple graphs are superimposed to identify possible extensions of the network graph.

Network dynamics

The predicted network according to the protocol shown in Fig. 6 is a static network indicating the constituent nodes and their connection patterns. The next step is to predict the network dynamics. We think small perturbations around the

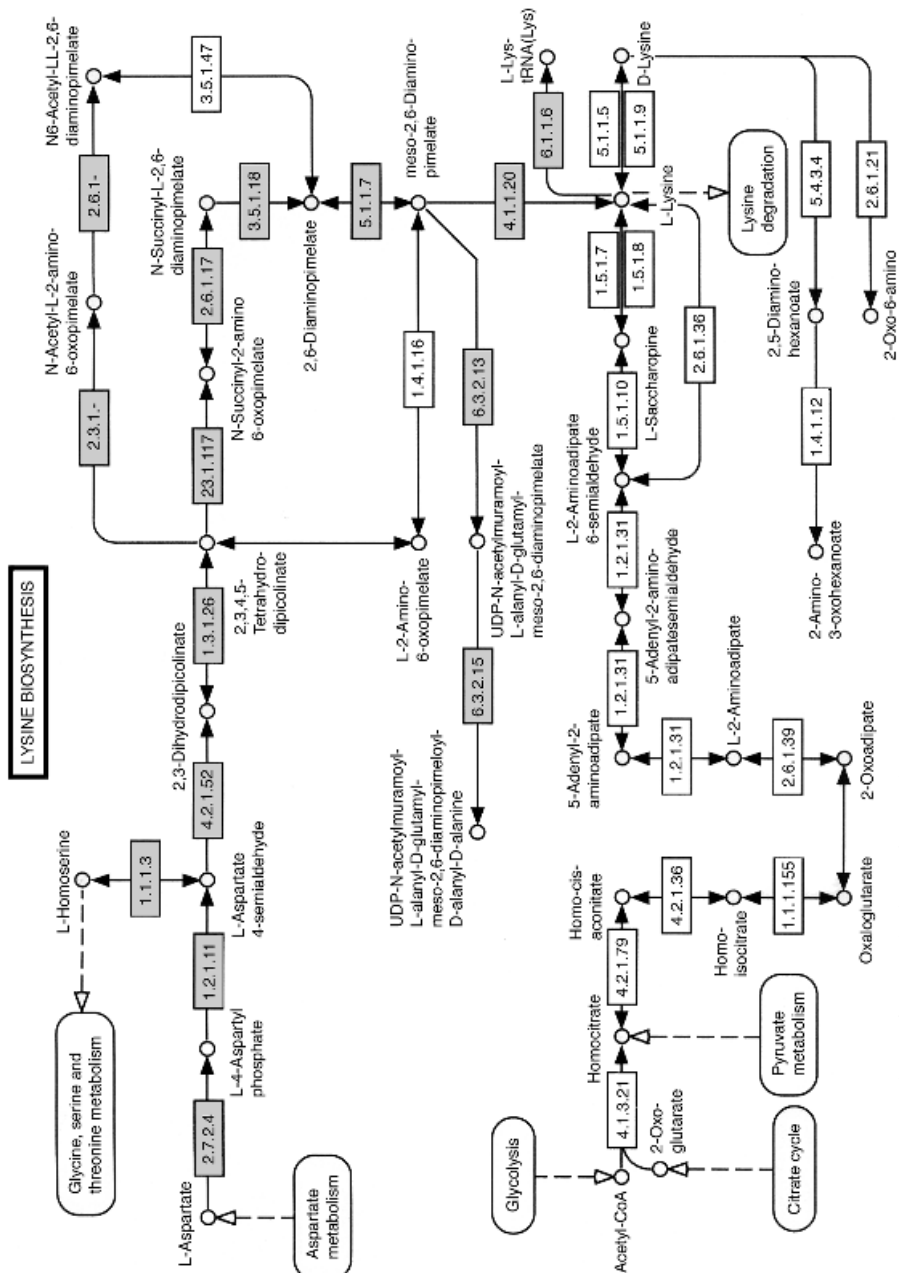


FIG. 4. KEGG pathway diagram for lysine biosynthesis.

The KEGG Reference Network
 Matching genes in the genome against gene products in the network

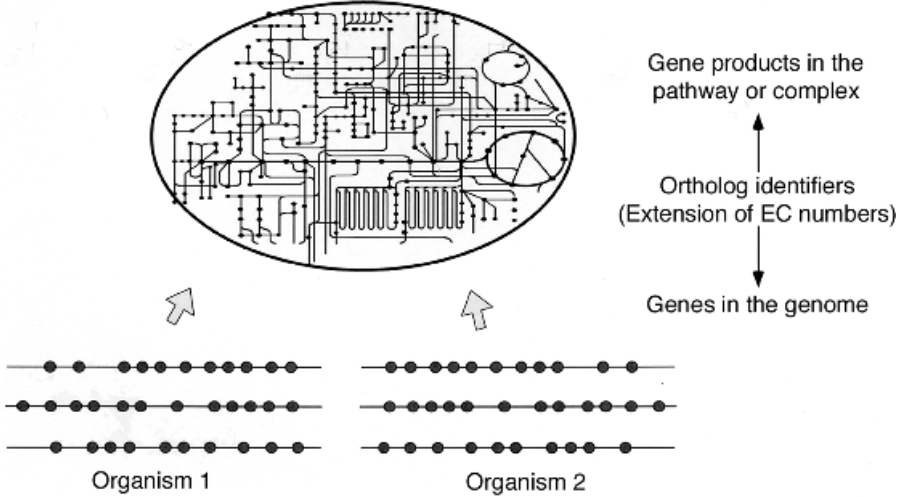


FIG. 5. KEGG reference network for knowledge-based prediction.

native network are computable, which is like computing small perturbations around the native structure of a protein. However, the dynamics of cell differentiation, for example, would be extremely difficult to compute, which is like computing the dynamics of protein folding from the extended chain to the native structure. A perturbation to the network may be internal or external. An internal perturbation is a genomic change such as a gene mutation or a molecular change such as a protein modification, and an external perturbation is a change in the environment of the cell.

Although we do not yet have a proper way to compute dynamic responses of the network to small perturbations, a general consideration can be made. Figure 7 illustrates the basic system architecture that results from the interactions with the environment. The basic principle of the native structure formation of a globular protein is that it consists of the conserved hydrophobic core to stabilize the globule and the divergent hydrophilic surface to perform specific functions. The protein interaction network in the cell seems to have a similar dual architecture. It consists of the conserved core such as metabolism for the basic maintenance of life and the divergent surface such as transporters and receptors for interactions with the environment. The subnetwork of genetic information processing may also have a dual architecture: the conserved core of RNA polymerase and ribosome and the divergent surface of transcription factors. In both cases the core is encoded by a set of orthologous genes that are conserved among organisms, and the surface is

1. Generate networks by matching against reference knowledge
2. Extend networks by incorporating additional experimental data

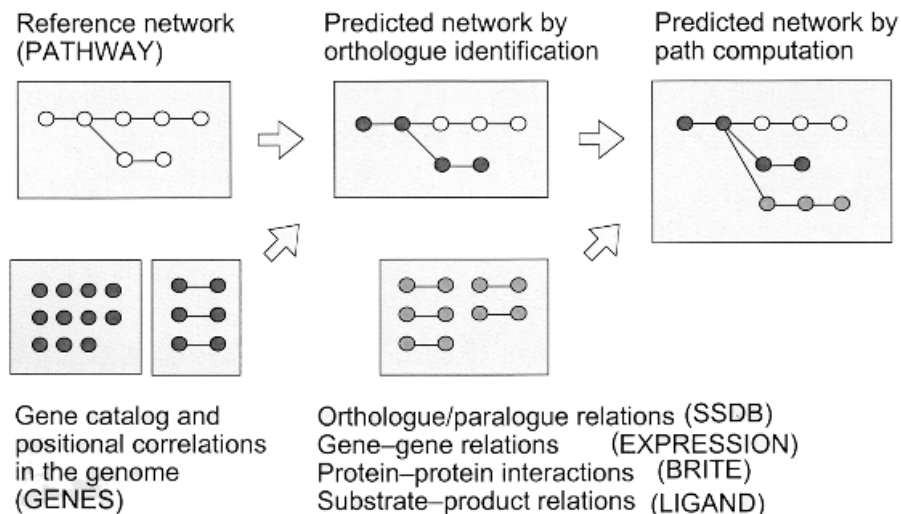


FIG. 6. Network prediction protocol in KEGG.

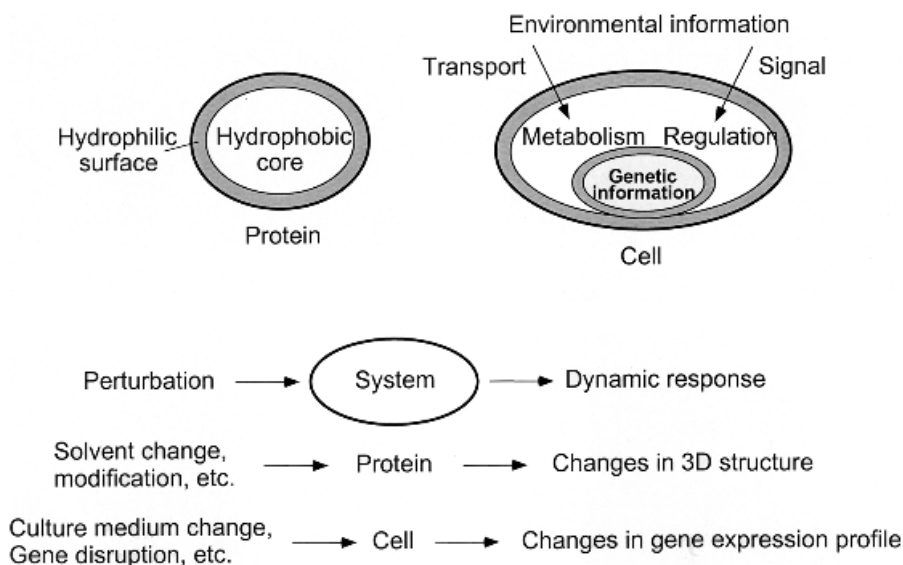


FIG. 7. System architecture that results from interactions with the environment.

encoded by sets of paralogous genes that are dependent on each organism. Thus, we expect that the genomic compositions of different types of genes in different organisms reflect the environments which they inhabit and also the stability of the network against environmental perturbations. By comparative analysis of a number of genomes, together with experimental data observing perturbation–response relations such as by microarray gene expression profiles, we hope to come up with a ‘conformational energy’ of the protein interaction network, which would then be utilized to compute a perturbed network by an energy minimization procedure.

Acknowledgements

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology of Japan, the Japan Society for the Promotion of Science, and the Japan Science and Technology Corporation.

References

- Kanehisa M 1997 A database for post-genome analysis. *Trends Genet* 13:375–376
Kanehisa M 2000 Post-genome informatics. Oxford University Press, Oxford
Kanehisa M 2001 Prediction of higher order functional networks from genomic data. *Pharmacogenomics* 2:373–385
Kanehisa M, Goto S, Kawashima S, Nakaya A 2002 The KEGG databases at GenomeNet. *Nucleic Acids Res* 30:42–46
Ogata H, Fujibuchi W, Goto S, Kanehisa M 2000 A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res* 28:4021–4028

DISCUSSION

Subramaniam: How would one go about making comparisons of microarray data with yeast two-hybrid data, which have different methods of interaction distance assessment and completely different metrics?

Kanehisa: At the moment we don’t include a numerical value. We just say whether the edge is present or not. It is a kind of logical comparison. If we start including the metrics we run into the problem of how we balance two different graphs. We would need to normalize them.

Subramaniam: When you draw networks by analogy, using your graph-related methods, if you have more nodes adding on going from a pathway in one organism to a pathway in another organism, it is not a problem because you can add more nodes. But what if the state of the protein is different in the two pathways? We have a good example with receptor tyrosine kinases: there are two different phosphorylation states of this. In one case there are two tyrosines phosphorylated, in another there are four. How do you deal with this distinction in the state-dependent properties of the graph?

Kanebisa: At the moment we don't distinguish different states. We are satisfied with just relating each node to the genomic information. As long as we have the box coloured, which means that the gene is present, that is sufficient — our interest is to obtain a rough picture of the global network, not details of individual pathways.

Reinhardt: Take the following scenario. I am trying to predict a protein–protein interaction from expression profiles. I take two different genes, look at them across a number of experiments and construct and compare the vectors. I find that one of the genes has two biochemical roles, and is shuttling between two compartments. Then what I would need, when I try to speak in the language of sequence analysis, is a local alignment. Currently, all we do in expression profiling is to compute a global alignment. We are in the Stone Age. Have you any idea of how to address this need for local alignment? Given your concluding Pearson correlation coefficient of 0.97, it wouldn't work if you have multifunctional proteins. How do you address this?

Kanebisa: Again, just looking at expression data it is very difficult to find the right answer. But we have an additional set of data, including yeast two-hybrid data. Integration of different types of data is the way we want to do the screening. Together with an additional data set we can find the local similarity when we do the graph comparison.

Crampin: How do you go about incorporating data other than just connectivity, for example the strengths of interactions between components of a network? Obviously, if you are describing atoms within a protein molecule, this is not of such great importance. But if you are looking at networks at the signalling level, the strengths of interactions may be crucial. Interestingly, there are some modelling results suggesting that for some gene networks it is the topology and not the strengths of connections that is responsible for the behaviour of the network (von Dassow et al 2000).

Kanebisa: We see this database as the starting point of giving you all candidates. By using this database and then screening it is possible to identify subsets of candidates. If you have additional information, this may help identify subsets among the results. Then you can start incorporating kinetic parameters and so forth.

Crampin: As you go up in scale from purely molecular data, you also need to include spatial information. Are there clear ways of doing this?

Kanebisa: This can be done. We showed the distinction of organism-specific pathways by colouring. The spatial information can be included by different colouring or by drawing different diagrams.

Subramaniam: From your graphs can you define modules for pathways that can then be used for modelling at higher levels? Is there an automatic emergence of the natural definition of 'module'.

Kanebisa: Yes. The reason why we are able to find graph features such as hubs and cliques is that the graph can be viewed at a lower resolution. We are trying to find a composite node or a module that can be used as a higher-level node in modelling.

Berridge: So if you put Ras into your model, would it predict the MAP kinase pathway?

Kanebisa: Yes.

McCulloch: Would you be able to predict this without the reference information?

Kanebisa: No.

Subramaniam: With reference to your modules, can they be used for kinetic modelling such as the sort of thing that Andrew McCulloch does? Or can they be used as a central node for doing control-theory-level modelling?

Kanebisa: I'm not sure. First, we need a kinetics scheme among modules, which is not present in our graph. But maybe we can tell you which modules to consider.

Reinhardt: As an example of how this approach might be used, if you have a protein and you don't know what it does, you can ask this system to give it its biological context. If you think about it, half of the genes in the genome are of unknown function. In the future we will have whole genome Affymetrix-style chips, and this will be a very important tool. We can go to this 50% of unknown genes, run it across a series of tissue samples and then try to see which pathways these genes are involved with and which proteins they are interacting with. This would give us a rough idea of the biological context of these unknown genes.

Reference

von Dassow G, Meir E, Munro EM, Odell GM 2000 The segment polarity network is a robust developmental module. *Nature* 406:188–192

Bioinformatics of cellular signalling

Shankar Subramaniam and the Bioinformatics Core Laboratory

Departments of Bioengineering and Chemistry and Biochemistry, The University of California at San Diego and The San Diego Supercomputer Center, La Jolla, CA 92037, USA

Abstract. The completion of the human genome sequencing provides a unique opportunity to understand the complex functioning of cells in terms of myriad biochemical pathways. Of special significance are pathways involved in cellular signalling. Understanding how signal transduction occurs in cells is of paramount importance to medicine and pharmacology. The major steps involved in deciphering signalling pathways are: (a) identifying the molecules involved in signalling; (b) figuring out who talks to whom, i.e. deciphering molecular interactions in a context specific manner; (c) obtaining the spatiotemporal location of the signalling events; (d) reconstructing signalling modules and networks evoked in specific response to input; (e) correlating the signalling response to different cellular inputs; and (f) deciphering cross-talk between signalling modules in response to single and multiple inputs. High-throughput experimental investigations offer the promise of providing data pertaining to the above steps. A major challenge, then, is the organization of this data into knowledge in the form of hypothesis, models and context-specific understanding. The Alliance for Cellular Signaling (AfCS) is a multi-institution, multidisciplinary project and its primary objective is to utilize a multitude of high throughput approaches to obtain context-specific knowledge of cellular response to input. It is anticipated that the AfCS experimental data in combination with curated gene and protein annotations, available from public repositories, will serve as a basis for reconstruction of signalling networks. It will then be possible to model the networks mathematically to obtain quantitative measures of cellular response. In this paper we describe some of the bioinformatics strategies employed in the AfCS.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 104–118

The response of a mammalian cell to input is mediated by intracellular signalling pathways. Such pathways have been the focus of extensive research ranging from mechanistic biochemistry to pharmacology. The availability of the complete genome sequences portends the potential to provide a detailed parts list from which all signalling networks can eventually be constructed. However, the genome merely provides the constitutive genes and carries no information on the on the exact state of the protein that manifests function.

In order to map signalling networks in mammalian cells it is desirable to obtain an inventory of the contents of the cell in a spatiotemporal context, such that the presence and concentration of every species is mapped from cellular input to

response. The 'functional states' of proteins and their interactions then can be constituted into a network which can then serve as a model for computation and further experimental investigations (Duan et al 2002).

The Alliance for Cellular Signaling (AfCS) (<http://www.afcs.org>), is a multi-institutional, multi-investigator effort aimed at parsing cellular response to input in a context-dependent manner. The major objectives of this effort are to carry out extensive measurements of the parts list of the cell involved in cellular signalling to answer the question of where, when and how proteins parse signals within cells leading to a cellular response. The measurements include ligand screen experiments that provide snapshots of the concentrations of the intracellular second messengers, phosphorylated proteins and gene transcripts after the addition of defined ligand inputs to the cell. Further, protein interaction screens provide a detailed list of interacting proteins and fluorescent microscopy provides the location within the cell where specific events occur. These measurements in conjunction with phenotypic measurements such as movement of B cells in the presence of chemoattractants and contractility in cardiac myocyte cells can provide insights into the intracellular signalling framework.

The ligand screen experiments are expected to provide a measure of similarity of cellular response to different inputs and as a consequence provide insights into the signalling network. The data are publicly disseminated prior to analysis by the AfCS laboratories through the AfCS website (<http://www.afcs.org>). Further experiments include a variety of interaction screens including yeast two-hybrid and co-immunoprecipitation. It is expected that the combined data from these experiments will provide the input for reconstruction of the signalling network.

Reconstruction of biochemical networks is a complex task. In metabolism, the task is somewhat simplified because of the nature of the network, where each step represents the enzymatic conversion of a substrate into a product (Michal 1999). This is not the case in cellular signalling. The role of each protein in a signalling network is to communicate the signal from one node to the next, and to accomplish this the protein has to be in a defined signalling 'state'. The state of a signalling molecule is characterized by covalent modifications of the native polypeptide, the substrates/ligands bound to the protein, its state of association with other protein partners, and its location in the cell. A signalling molecule may be a receptor, a channel, an enzyme, or several other functionally defined species, depending on its state. In the process of parsing a signal, a molecule may undergo a transition from one functional state to another. We define the Molecule Pages database which will provide a catalogue of states of each signalling molecule, such that one can begin to reconstruct signalling pathways with molecules in well-defined states functioning as nodes of a network. Interactions within and between functional states of molecules, as well as transitions between functional states, provide the building blocks for reconstruction of a signalling network. The

AfCS experiments will test and validate such interactions and transitions in specific cells of interest.

The Molecule Pages database

'Molecule Pages' are the core elements of a comprehensive, literature-derived object-relational (Oracle) database that will capture qualitative and quantitative information about a large number of signalling molecules and the interactions between them. The Molecule Pages contain data from all relevant public repositories and curated data from published literature entered by expert authors. Authors will construct Molecule Pages by entry of information from the literature into Web-based forms designed to standardize data input. The principal barrier on constructing a database such as this lies in the complex vocabulary used by biologists to define entities relating to a molecule. The database can only be useful if it is founded on a structured vocabulary along with defined relationships between objects that constitute the database (Carlis & Maguire 2001). The building of this 'schema' thus is the first step towards the reconstruction of signalling networks. The schema for sequence and other annotation data obtained from public data repositories is presented below. A detailed schema for the author-curated data will be presented elsewhere.

Automated data for Molecule List and Molecule Pages

The automated data component of each Molecule Page comprises information obtained from external database records related in some way to the specific AfCS protein. This includes SwissProt, GenBank, LocusLink, Pfam, PRINTS and Interpro data as well as Blast analysis results from comparing against a non-redundant set of sequence databases (created by the AfCS bioinformatics group).

Generation of Protein List sequences

Protein and nucleic numbers are read on a nightly basis from the AfCS Protein List (by a Perl program), and they are used to scan the NCBI Fasta databases to find the sequences. A tool that reports back information and any discrepancies (based on the GI numbers that were assigned) is available for use by the Protein List editors. Fasta files for all AfCS proteins and nucleotides are generated, with coded headers that allow us to tie each sequence to its AfCS ID. The Fasta files as well as a text file containing a spreadsheet-like view of the AfCS Protein List can be downloaded by the public from an anonymous ftp server. The Fasta protein file is used as the basis for further analysis.

All AfCS data are stored in Oracle tables, keyed on the Protein GI number. Links are provided to NCBI. A database is used to store information to allow each

sequence to be imported the Biology Workbench for further analysis. This process is run about once a month, and consists of a set of PERL programs, which launch the various jobs, parse the output, and load the parsed output into the Oracle database.

Supporting databases for Molecule Pages

In order to support all the annotation, entire copies of each relevant database are mirrored in flat file form on the Alliance Information Management System. These databases include Genbank, Refseq, SwissProt/TrEMBL/TrEMBLnew, LocusLink, MGDB (Mouse Genome Database from Jackson Laboratories), PIR, PRINTS, Pfam, InterPro, and the NCBI Blastable non-redundant protein database 'NCBI-NR'. These databases are updated every day, if changes in the parent repositories are detected. Some of the databases (or sections of the databases) are converted to a relational form and uploaded to the Oracle system to make the analysis system more efficient.

The NCBI-NR database contains all the translations from Genbank, PIR sequences, and SwissProt sequences. It does not contain information on TrEMBL sequences, however, and many public databases contain SwissProt/TrEMBL references exclusively. This necessitated the construction of an in-house combined non-redundant database, called 'CNR' for short.

In addition to database links, title information and the sequence, CNR database contains date information (last update of the sequence) and NCBI taxonomy ID where available. The database also contains the sequences SwissProt/TrEMBL classify as splice variants, variants and conflicts (these are generally features within those records, so a special parser provided by SwissProt is used to generate those variant sequences). A Perl program constructs this database on a weekly basis, and a combination of a Perl/DBI script and Oracle sqldr is used to load the database to the Alliance Information Management Oracle System.

The interface pages are logical groups of the automated data, and are subject to rearrangement and reclassification. Making changes will have no effect on the underlying schema or the methods for obtaining the data. Examples of schema for automated data, employed in the molecule page database, for annotating GenBank, SwissProt, LocusLink and Motif and Domain data are shown in Figs 1–3.

Design of the Signalling Database and Analysis System

The Molecule Pages will serve as a component of the large Signalling Database and Analysis System. This system would have the capability to compare automated and experimental data to elucidate the network components and connectivities in a context-dependent manner. Thus, we can use our biological knowledge of the

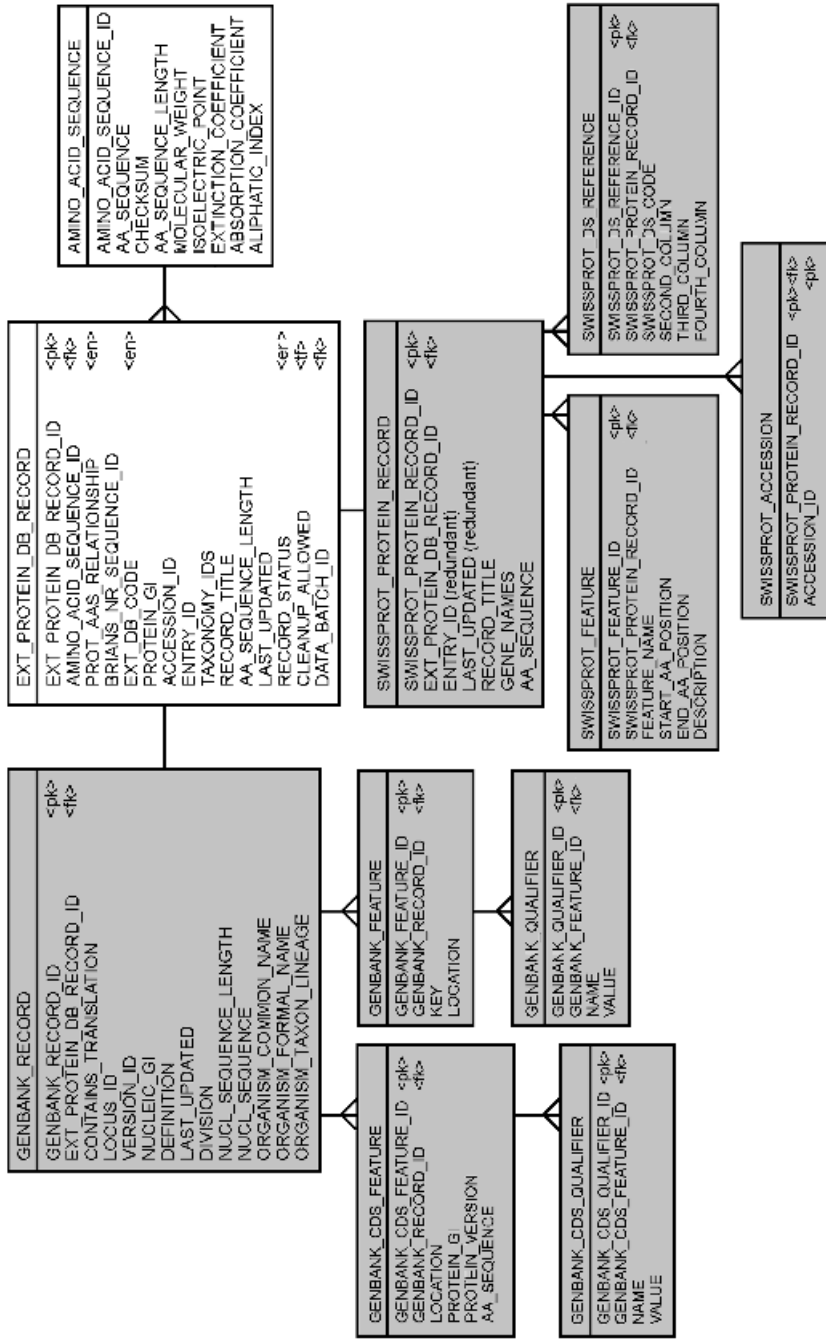


FIG. 1. A data model for GenBank and SwissProt records annotated in the Molecule Pages.

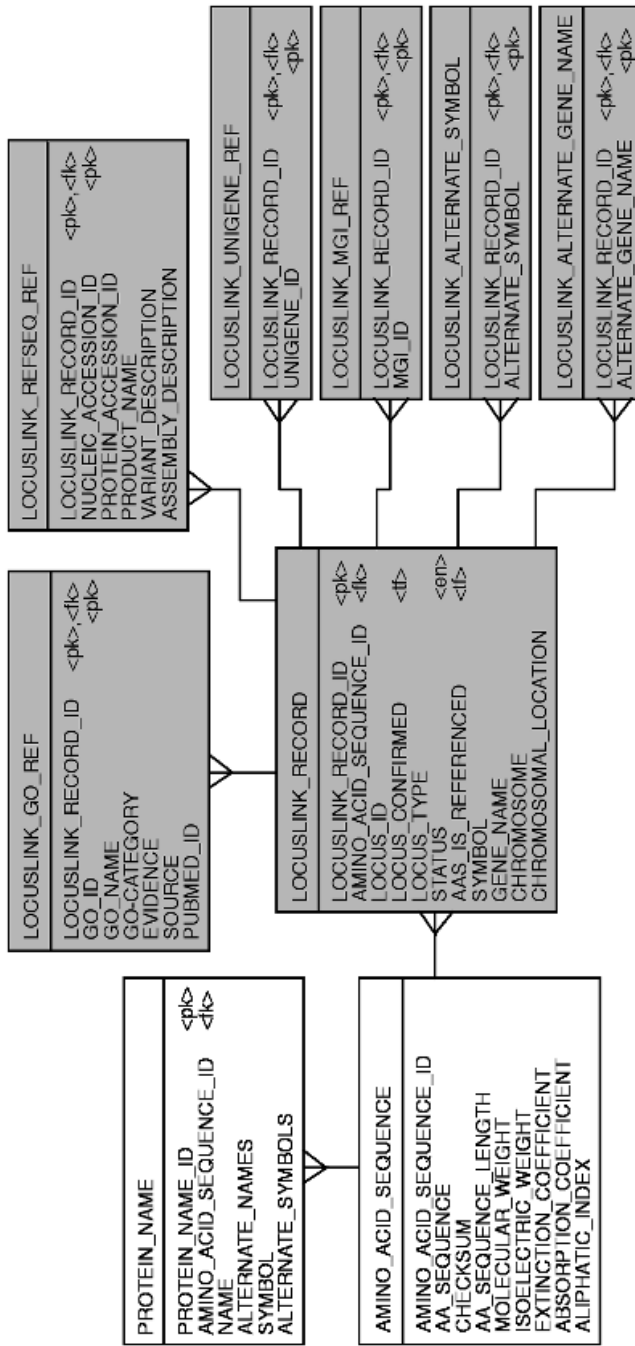


FIG. 2. A data model for Locus Link records annotated in the Molecule Pages.

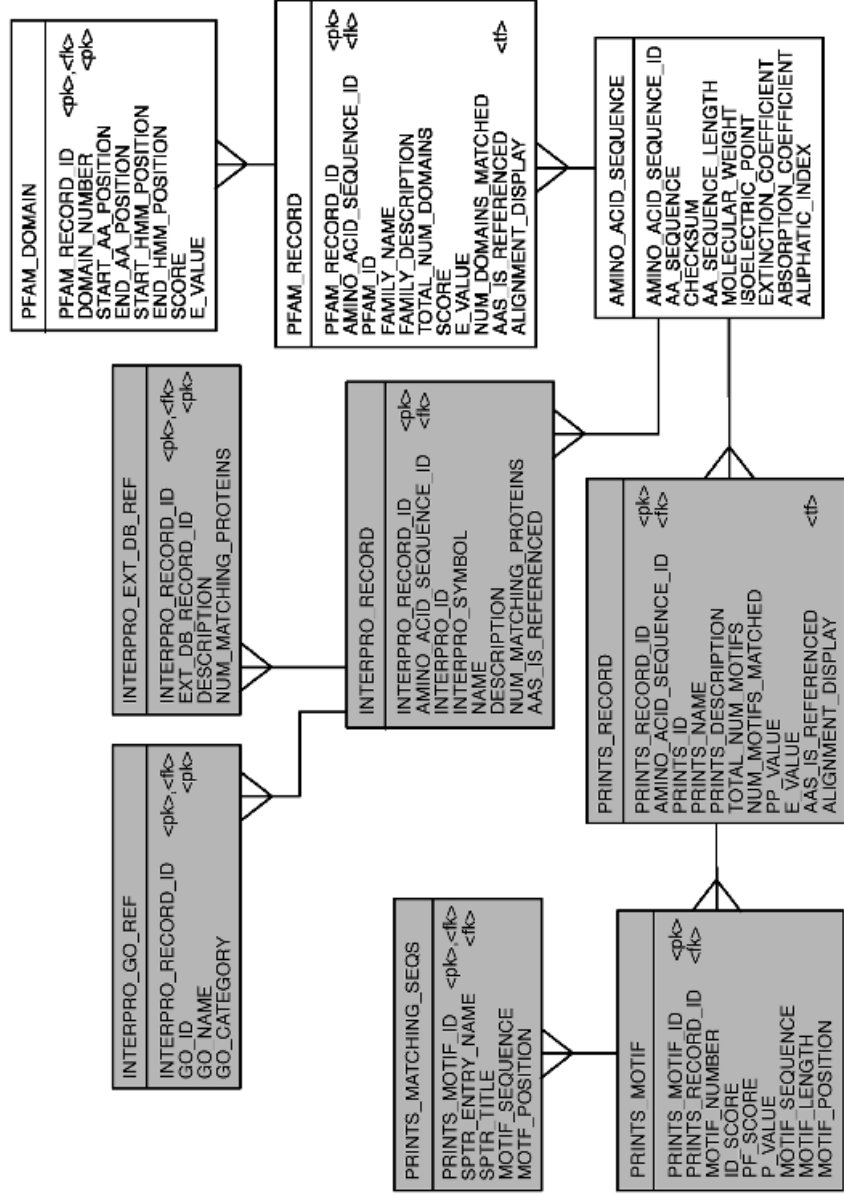


FIG. 3. A data model for PRINTS, Pfam and InterPro records annotated in the Molecule Pages.

putative signalling pathways and concomitant protein interactions to interrogate large-scale experimental data. The analysis of the data can then serve to form a refined pathway hypothesis and, as a consequence, suggest new experiments.

The process of construction of pathway models requires the assembly of an extended signalling database and analysis system. The main components of such a system are a pathway graphical user interface (GUI) for representing both legacy and reconstructed pathways, an underlying data structure that can parse the objects in the GUI into database objects, a signalling pathway database (in Oracle), analysis links between the signalling GUI and other databases, and links to systems analysis and modelling tools.

The components of the Signalling Database and Analysis System include:

- (a) Creation of an integrated signalling GUI and database system
- (b) Design of a system for testing legacy pathways against AfCS experimental data
- (c) Reconstruction of signalling pathways
- (d) Creation of tools for validation of pathway models

An overview of an integrated signalling database environment is presented in Fig. 4.

Computer science strategies

Development of an integrated system of this nature requires the amalgamation of four separate pieces, namely Java, Oracle, Enterprise Java Beans (EJB) and XML (eXtensible Markup Language). We envision an application based on a three-tier paradigm, consisting of the following components.

System architecture. The system is based on a three-tier architecture (Tsichritzis & Klug 1978), as illustrated in the following diagram (Fig. 5). An Oracle 9i database server is connected through a middle tier, Oracle application server (OAS) 9i from a client web browser or a stand-alone application using Java swing. OAS 9i can reduce the number of database connections from client by combination and then connect to the database server. Java Servlets, Java Server Page (JSP), Java Beans and/or EJB are used to separate business logic and presentation for a dynamic web interface. In the business logic middle tier, Java Beans and EJB are used. With Object Oriented features and component-oriented programming, Java API benefits our interface development.

Communication between swing client and middle tier will be through EJB components or via HTTP by talking to servlet/JSP. The latter allows easy navigation through firewalls, while the former allows the client to call the server using intuitive method names, obviates the need for XML parsing, and automatically gives remote access and load-balancing. XML (Quin 2001) will be used for

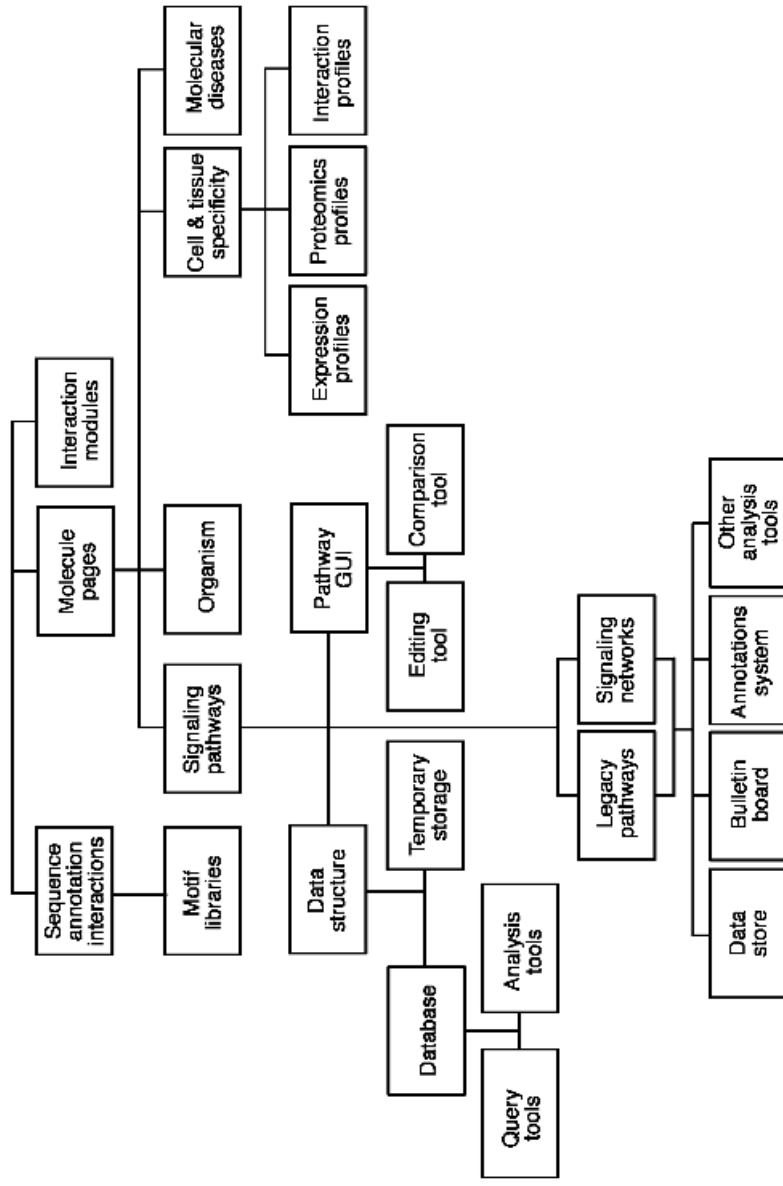


FIG. 4. A schematic diagram for the signalling database and analysis infrastructure. The links show how the signalling graphical user interface is linked to a data structure which communicates with all pertinent databases. In addition the interface links to experimental data that can be invoked through analysis tools provided in the analysis tool kit.

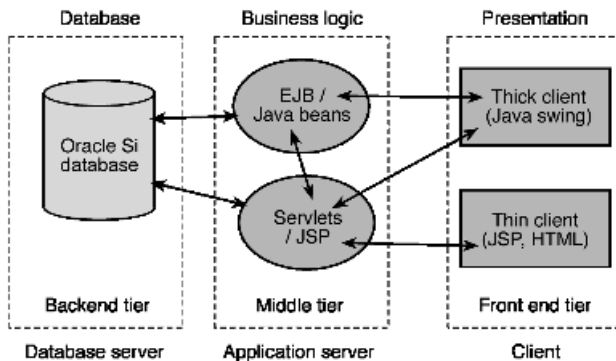


FIG. 5. A schematic view of the three-tier diagram. The three-tier architecture is common to most modern databases (Tsichritzis & Klug 1978).

building the pathway model to store locally and send back to the server. We will explore SBML (Systems Biology Markup Language, (<http://xml.coverpages.org/sbml.html>), and CellML (Cell Markup Language) (http://www.cellml.org/public/specification/cellml_specification.html) for this purpose. EJB/Java Beans middle tier enables query of the relational database, creation of the XML model, and export to the client for display purposes.

Database structure. The Molecule Page database will serve as a core starting point for the Pathway Database System. This database will communicate with other AfCS experimental and annotation databases. The functional states of signalling proteins created in the Molecule Page database will be used to build signalling pathways. A digital signature corresponding to each functional state of a protein has been established in the Molecule Pages to determine whether states described in two distinct Molecule Pages are the same. This digital signature captures the state of the protein in terms of its interactions, covalent modifications, and subcellular localizations. The digital signature enables direct comparisons across nodes in two distinct pathways. Thus, if the digital signature of protein kinase A in two different pathways is the same, then the kinase is in the same functional state in the two pathways.

Middle tier. The middle tier will be composed of both EJB and regular Java classes and is based on Enterprise Java technology. Enterprise Java technology provides common services to the applications, ensuring that these applications are reasonably portable and can be used with little modification on any application server. The specifications cover many areas including:

- HTTP communication: a simple interface is presented for the interrogation of requests from web browsers and for the creation of the response.

- HTML formatting: Java Server Pages (JSP) provide a formatting-centric way of creating web pages with dynamic content.
- Database communication: Java database connectivity (JDBC) is a standard interface for talking to databases from application code. Many large database vendors provide their own implementations of the JDBC specifications.
- Database encapsulation: The EJB specification defines a way to declare a mapping between application code and database tables using an XML file, as well as additional services such as transaction control.
- Authentication and access control: many of the Enterprise Java specifications define standard mechanisms for authenticating users and restricting the content that is available to different users.
- Naming services: the Java naming and directory interface (JNDI) specification defines a way for application code to consistently obtain references to remote objects (i.e. those in another tier) based on names defined in XML files.

The motivation for the development of a middle tier is to isolate the client tier from changes in the database by forcing communication through a consistent interface featuring the objects that we know are present in our system, but for which the schema still occasionally changes. The use of a middle tier also allows both Java swing and web clients to efficiently obtain information from the database. The middle tier can take care of the ‘business logic’ and database access on behalf of other clients. A typical task for the middle tier is that of intercepting requests from client and querying the database for node list, reaction list, localization information, and model meta data, and then returning instances of Java classes that encapsulate the requested information in an object-oriented manner. It can also return to the client an XML document that describes the pathway model.

GUI applications: testing pathway models against AfCS and other data. The primary objective of the GUI will be to extract and display visual representation of pathways. The user will be able to make selection(s), changes, and extensions to the representations in an interactive session. In addition to invoking existing pathways and drawing/editing pathways, the user will be able to launch queries and applications from the GUI. Some examples of interactive queries the user can pose are:

- has the inserted node been seen in any canonical pathways in the legacy databases?
- are the ensuing interactions already known based on protein interaction databases or interaction screen data?
- is a module present in other pathways?
- are two states of a molecule similar and, if so, to what extent?

Reconstruction of pathways

We use a combination of state-specific information from the Molecule Pages and AfCS experimental data to reconstruct pathways. The GUI will provide the graphical objects for the visual assembly editing and scrutiny of the pathways. Existing pathway models can also be invoked and edited to build models that are consistent with the AfCS data. We plan to provide two strategies for reconstruction. In the first, the author will be able to manually invoke specific signalling proteins in assigned states from the Molecule Pages and build appropriate connections. At any intermediate stage, the user can utilize the tools provided to check/validate the connections (as described previously). In the second strategy, the user will be able to utilize the knowledge of pair-wise interactions in specific contexts to automatically build networks that can be further edited. For example, if a user wants to map the interaction partners for a particular protein in a state dependent manner, the user will need to select a protein and its state from the Molecule Page database and make another selection to find the interacting partner. The protein and its interacting partners will be displayed as nodes on the GUI. Each node can now act as a further starting point, and the interaction diagram can be expanded dynamically to build an entire pathway. The existing annotation about the each node in the diagram, which represents a state of the protein, can be obtained by clicking at the node. It will also be possible for the user to incorporate other data that is not available in the Molecule Page database. The user will be able to save the interaction diagram as an XML file, which can be read back into the application or stored in the Oracle database. Other tools available on the GUI will enable the user to compare signalling pathways in relationship to expression or proteomic profiles.

Validation of pathways

We embed three combined approaches to validate pathways. In the first, we can test our pathway models against AfCS experimental measurements. Ca^{2+} and cAMP assays are expected to provide insight at a coarse-grained level into modules and pathways invoked by a ligand input. The immunoblot assays will indicate some of the proteins implicated in the pathway, as will the 2D phosphoprotein gels. The interaction screens will yield information on interaction partners, while the expression profiles are expected to show levels of similarity in response to different inputs. A pathway model can thus be tested against the AfCS data. We note that a more quantitative test of the pathway models will only be feasible when detailed experiments where a system is perturbed to achieve loss or gain of function (e.g. systematic RNAi experiments based on initial pathway models) are carried out and intermediate activities and endpoints are measured.

In the second approach, the pathways can be validated against existing data managed in AfCS databases. Comparative analysis of similar pathways across

cells from different tissues and from different species has been proven to be valuable for both testing the pathway models as well as providing insight into other putative players that have a role in the pathway. The presence of all legacy databases (sequence, interaction and pathways) will allow the user to query them interactively from the interface page.

In the third approach, network analysis tools are employed to investigate the role and sensitivity of each node in the network (Schilling et al 1999). We provide tools for constructing a discrete state network model and perform sensitivity analysis to test the importance and strength of each node and connection in the network. To test the robustness and correctness of our model of the signalling network, we will develop tools that will perturb individual nodes and their interactions to understand the sensitivity of the network to perturbation. The Signalling Database and Analysis System makes these tools accessible through the GUI.

Acknowledgements

The Alliance for cellular signalling is a multi-institutional research endeavour spearheaded by Dr Alfred Gilman at the University of Texas Southwest Medical Center. The participating laboratories include Core Laboratories at UT Southwest Medical Center, University of California San Francisco, Caltech, Stanford and University of California San Diego. The Alliance is a multi-investigator effort. The material presented here describes a collaborative effort across these laboratories. The AfCS is funded primarily through a Glue Grant by the National Institute for General Medical Sciences. Other funding sources include other Institutes at NIH and a number of pharmaceutical and biotechnology companies.

References

- Duan XJ, Xenarios I, Eisenberg D 2002 Describing biological protein interactions in terms of protein states and state transitions. THE LiveDIP DATABASE. *Mol Cell Proteomics* 1:104–116
- Michal G (ed) 1999 *Biochemical pathways*. Wiley, New York
- Carlis JV, Maguire JD 2001 *Mastering data modeling: a user-driven approach*. Addison-Wesley, Boston, MA
- Quin L 2001 Extensible Markup Language (XML). W3C Architecture: <http://www.w3.org/XML>
- Schilling CH, Schuster S, Palsson BO, Heinrich R 1999 Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog* 15:296–303
- Tsichritzis D, Klug A 1978 The ANSI/X3/SPARC DBMS framework report of the study group on database management systems. *Inf Syst* 3:173–191

DISCUSSION

Winslow: What kinds of analytical procedures are you using, particularly with the gene expression data, to deduce network topology?

Subramaniam: Currently we are using gene expression data to characterize the state of each cell. At this point in time we are not doing pathway derivation from this. Having said this, for characterizing the state of each cell, we are focusing very specifically on comparing across different inputs. We are taking 50 different inputs, at five time points with three repetitions. This is 750 microarray data sets. The analysis is done by using ANOVA, which cleans up the statistics, and then we analyse the profiles. The third thing we do is to relate each of the things that come out of this to our biological data. Our hope is that once we have a state-dependent knowledge of the molecule tables, then we can go back and tighten the pathways. Once we know the network, then we can ask the question about how it is related back in the gene expression profile database. One caveat is that we are looking at G protein-coupled receptor (GPCR) events, which are very rapid. During this short time-scale, gene expression changes don't happen, so we are doing the same types of things with proteomic data, which come from 2D gels and mass spectrometry.

McCulloch: What are the five time points?

Subramaniam: For mouse B cells, the time points are zero, 30 minutes, 1 h, 2 h and 4 h. We haven't started yet with the cardiac myocytes. These were chosen on the basis of preliminary experiments.

Hinch: How do you deal with conflicting experimental results? What if two people do the same experiment and get different results?

Subramaniam: It depends on whether these are Alliance experiments or outside experiments. For the Alliance experiments we have a number of repeats. We want to make sure that we have some level of confidence in everything that we do. We have experimental protocols for every experiment included. Even given this, there will be variation in gene expression data. In this case we take into account an average index, and this is where ANOVA is important. With outside data it is different. If we are doing state-dependent tables, for example, we have two different authors. Don't forget that many times these experimental output data are gathered under different conditions. We list all the different conditions. If for some reason for the same conditions there are two different results, then we cite them both.

Asburner: I have a question about the functional states. If there is a protein that has five different states at which phosphorylation might occur, then theoretically we have 2^5 functional states. Do you compute all of these?

Subramaniam: This is where the author-created interactions come into the picture. We ask the author to define all functional states for which data are available, both qualitative and quantitative. If they are not available, we don't worry about the potential functional states. For example, if tyrosine kinases have 16 phosphorylation sites, we are not trying to define 2^{16} possibilities. We recognize just the phosphorylation states that have been characterized.

Paterson: I was interested in how you come up with the perturbations to this system. You are using mouse B cells, and there are lots of interesting signalling events taking place in B cells that are not G protein-related, such as cytokines, Fas–Fas ligand interactions, differentiation and isotype switching. Are you looking at perturbations through cytokine and other receptors?

Subramaniam: GPCRs are our first line of investigation, but we are going to explore all signalling pathways that are coupled to GPCRs in one way or another. This includes cytokines and growth factor signalling.

Paterson: What is the process the Alliance uses for interacting with others in terms of conversations about what sort perturbations actually occur?

Subramaniam: That is an important question. This is why we have a steering committee. We don't want to work for a company, but we would like to solicit input from various pharmaceutical companies as to what they find interesting and exciting. We also have a bulletin board on the Alliance information management system. We encourage people to communicate with the Alliance at whatever of level of detail they choose. This is a community project.

Reinhardt: If many people submit data to your system, how do you deal with the problem of controlling the vocabulary?

Subramaniam: This is one thing we are not socialistic about. We are not going to allow everyone to submit data to this system: it's not that type of database. Where the public input comes in is to alter the shape of molecular pages. The authorship will be curated, peer reviewed and so forth. In terms of the Alliance data, we will post this but it doesn't mean we won't cite references to external data where they are relevant.

Reinhardt: Something you said early on in your talk caught my attention: while the community today mostly relies on relational databases in biology, it is appreciated and understood that this concept is not good enough to model the complexity of biological data. You said you are moving towards object relational databases. What are the object-oriented features of this database?

Subramaniam: This is a very important question. The original article is more relational than object relational. We have entered into a collaboration with Oracle and have decided that we are going to go with an object-relational database. In fact, if you look at our ontology, everything is an object-driven ontology definition. Oracle is now coming up with 10i, which will be a completely object-relational database. We explored four different database formats before we arrived at this, including msSQL, postgresSQL (which I like a lot, but we don't have enough people available to do programming for this) and sybase. It is our firm conclusion, based on hard evidence, that the only thing that has the features of scalability, flexibility and the potential for middle-tier interactions is Oracle.

General discussion II

Standards of communication

Hunter: I am going to talk about the development of CellML, which is a project that originally grew out of our frustrations in dealing with translating models published in papers into a computer program. We decided that the XML (eXtensible Markup Language) developed by the W3C (World Wide Web Consortium) was the appropriate web-browser compliant format for encapsulating the models in electronic form. In conjunction with Physiome Sciences Inc., Poul Nielsen of the Auckland Bioengineering group has led the development of an XML standard for cell models called CellML. It uses MathML, the W3C approved standard for describing mathematical equations on the web and a number of other standards for handling units and bibliographic information, etc. A website (www.cellML.org) has been established as a public domain repository for information about CellML and it contains a rapidly expanding database of models which can be downloaded free of charge and with no restrictions on use. These are currently mainly electrophysiological models, signal transduction pathway models and metabolic models, but the CellML standard is designed to handle all types of models. A similar effort is underway at Caltech for SBML (Systems Biology Markup Language) and the two groups are keeping closely in touch. A number of software packages are being developed which can now, or will soon be able to, read CellML files. Authoring tools are available from Physiome Sciences (free for academic use). Our hope is that the academic journals dealing with cell biology will eventually require models to be submitted as CellML files. This will make it easier for referees to test and verify the models and for scientists to access and use the published models.

Loew: I think the people at Caltech are going to link SBML to Genesis. If the two merge, this would be one of the consequences.

Winslow: There aren't many people—whether they are modellers or biologists—who can write XML applications. You referred to the development of these authoring tools: do you see these being publicly available as open source for the entire community to use to create the kind of CellML that you showed here?

Hunter: One source of these has been Jeremy Levin; he may want to comment on this.

Levin: Part of what we will be doing is actually making some of these tools available publicly and openly.

Winslow: On the flip side, once you have these descriptions of models that are available publicly, what are your plans for converting them to code? How will the community use those descriptions to generate code?

Hunter: There are several ways this can currently be done. For example, you read MathML into Mathematica or MathCAD. These are standard programs that can churn out code from MathML. The cell editor from Physiome Sciences can read CellML files or create new ones. These can then be exported in various languages. In Auckland we are also working on exactly this issue for our own codes, so that we can just take CellML files and generate the code that we can then run in the bigger continuum models.

McCulloch: One of the features of XML is that it is extensible. Some of the models that you cited are actually extensions of previous models. They are often not simple extensions, but people have taken a previous model and made some specific modifications, such as changing some of the parameters and adding a channel. Then the next model came along and took the previous one as a subset. Have you tried an example of actually composing a higher-order model from the lower-order models?

Hunter: It is very tempting to do this. One reason I wanted to illustrate the historical development of electrophysiological models, from Denis Noble's early ones to the latest versions, was to use this almost as a teaching tool, demonstrating the development of models of increasing sophistication. Each one has been based on a published paper. The CellML file is deliberately intended to reflect the model as published in the paper. CellML certainly has the concept of reusability of components where you could do exactly as you are saying. The initial intent has just been to get these models on the website corresponding to the published versions.

Loew: The big problem there would be vocabulary.

Paterson: In my experience, one of the first things you want to do when you share a model is that there may be a variety of behaviours that you want to point out. One question I have for the CellML standard is the following: part of what I would want to give to someone with the model would be various parametric configurations of that model. So I can say that this is a configuration that mimics a particular experiment, or a configuration where you see a particular set of phenomena that come out of the model. I may have many of these to show how the model behaves in different regimes. Is there a facility within CellML to capture different parametric configurations of the basic cellular equations, and then perhaps to annotate to end-users, looking for the behaviour that comes out under these circumstances?

Hunter: In a way this is more to do with the database issue. Once you have a CellML version of a published model, you can then run that model with different

initial conditions and parameters. Some of the Physiome modelling software will allow you to do this and archive those particular parameter sets for those runs in the database. This is a little bit separate from CellML itself.

Paterson: I guess that is the key question. I am not sure whether what I am asking is purely in the domain of the environment, or whether CellML itself as a standard captures some of those. It seems that the answer is that it is more the environment.

Levin: CellML facilitates at least the two topics you are talking about here. For example, one is that by using our software we can automatically sweep through model parameters and store them. This is a different issue to CellML itself, but it is related to the ability to use it easily. Perhaps more importantly, because of the common format CellML actually allows us to merge different types of models together. For example, we can combine an electrophysiological model and a signalling model. This is made feasible only by the use of CellML.

Hinch: In CellML is there a way to link back to the original experimental data?

Hunter: Yes.

Semantics and intercommunicability

Boissel: I have prepared a short list of words for which there is uncertainty regarding their meaning, both in the field and—perhaps more importantly—outside our community (Fig. 1 [*Boissel*]). We want to communicate with people outside the field, in particular to convince them that the modelling approach is important in biology. I propose to go systematically through this list discussing the proper meaning we should adopt for each of these terms.

First, what are the purposes of a model? It is either descriptive, explanatory or predictive. These three functions are worth considering.

Levin: I would also say that a model is integrative. Its job is also to integrate data. If this fits under the heading of ‘descriptive’, then I agree with you, but I’m not sure that this is what you are encompassing.

Boissel: So you are proposing we add integrative as a fourth function?

Paterson: I would think that integrative would cut across all three: it is almost orthogonal.

Subramaniam: I don’t understand the difference between ‘descriptive’ and ‘explanatory’. Can you give an example of the difference between the two?

Boissel: We may decide to model something just for the sake of putting together the available knowledge, to make this knowledge more accessible. This is a description. In contrast, if you are modelling in order to explain something, you are doing the model in order to sort out what the important components are, in order to explain the outcome of interest. This adds something to a purely descriptive purpose.

Drivers

entity or function that 'explains' the outcome/output

Integration

incorporating components through quantitative relations
combining all available knowledge and evidence

Model components

biochemical entities
functional entities
both
either one

Modularity and modules

breaking down the problem in N autonomous and homogenous sub-problems
extension: is life a combination of modules?

Ontologies

models
data (knowledge, evidence,...)

Purposes of a model

descriptive
explanatory
predictive

Reductionism

explaining the outcome or the phenomenon by the play of a single biochemical or functional entity
reducing the complexity to a more workable level reducing complexity to a level relevant to the objective of modelling

Robustness

insensitivity to parameter values
insensitivity to uncertainty

Validation

objective driven
standard
both

FIG. 1. (*Boissel*) A short list of words commonly used in biological modelling whose meaning is uncertain.

Asbburner: Surely the description of Hodgkin–Huxley within CellML is a descriptive model of that model.

Noble: It is interesting that you have taken the Hodgkin–Huxley model as an example. The title of that paper is very interesting. It isn't, 'A *model* of a nerve impulse'. It does not even go on to say, 'A *theory* of a nerve impulse'. It says, 'A *description* of ionic currents, and their application to conduction and

excitation in nerve'. This title has been ingrained in my head since 1952! This raises an important question: what is explanatory, like beauty, is in the eye of the beholder. Is a description already an explanation? Obviously Andrew Huxley and Alan Hodgkin were operating in a biological environment in 1952 that did not accept the idea that there was a theory. Certainly, the *Journal of Physiology* would not have accepted that you publish a theory—I know from hard experience! I also think that what is explanatory to me or you, Jean-Pierre Boissel, is not explanatory to Philip Maini. What he regards as an explanation is something he can get his mathematical mind around, be it graphically or in terms of seeing an insight in certain equations. I would suggest that we will all have our different ways of seeing something as being satisfactorily explanatory. In relation to communicating to the outside world, it would be very likely that the cross-section of what we regard as explanatory in this room will not be the cross-section of what is regarded as explanatory in the outside world.

McCulloch: You are really talking about the relationship between observation and theory. In the case of 'descriptive', you mean a mathematical formulation that merely parameterizes observation, without attempting to gain any further insight from it. The explanatory goes beyond merely parameterizing observation to be able to compute or predict results that have already been observed independently. Predictive modelling is applying that same sort of process to come up with a result that has not yet been observed, but in principle could be.

Aspburner: Consider a model of the universe at the time of the big bang. This model has to be all three: descriptive, explanatory and predictive.

Boissel: I believe that it is difficult to delineate a clear boundary between these different classes of models. Nonetheless, when you start to do your modelling process, you have an objective in mind, and it will probably primarily be concerned with one of these.

Cassman: I would like to give another dichotomy. The models that I have seen tend to fall into two categories. One is archival. Essentially, all they do is represent existing information in some kind of a form that is easily visible and that can then be traced back to some fundamental piece of information. These are purely archival, but they can't be used in many cases for any kind of predictive approaches. There is no embedded information that can be extracted in a way that allows us to manipulate them: they are static, and could just as easily be on a page as in a computer. The big difference between that and models that can be manipulated is that in some way or other the latter are predictive, even if they are only predictive of something that people already know.

Paterson: At least for me, what I have found in working with mainstream biologists is what I would mark as a distinct cutting line between descriptive and explanatory models, in that statistical models are descriptive. They are telling us that something happened; they are making no attempt to say why it happened. The

minute we start modelling the underlying dynamic processes that can give rise to these data, then we have stepped across the line into explanation, and then whether or not it is explanation or prediction is a question of our state of information. For most biologists, ‘model’ means statistics.

Boissel: There seems to be a certain consensus here, so we can move on to model components. I am not sure how important this is. A model is composed of components, which might be biochemical or functional entities. One question is what we should call the various pieces we put in the model between the entities. Should we call them component pieces? Perhaps this isn’t important.

Noble: It is important: the way in which we structure our languages and software is in part an attempt to respect the ontology, what we think are the components. The dilemma is that the world is not divided up in a way that is given. We have to divide the world up and see what the entities are.

Subramaniam: We need to define the elements or components of the model before we can go beyond that point. Ontology comes into the picture here as well, because we need to define the elements fairly precisely, so that any two people who are using the model will have the same understanding of the components and their relationships.

Boissel: What term do you prefer: ‘component’ or ‘element’?

Subramaniam: They are interchangeable in this case. The main point is that when we define the elements we are not defining the relationships between the elements. It is in the ontology that we define the relationships with the elements.

Boissel: I think we are dealing with two types of ontologies. There is one ontology for the models, and one for the data used to design and parameterize the model.

Subramaniam: With reference to models, there is an issue of representation which becomes integrally tied in with the ontology to some extent. With reference to data, sometimes there are representation issues, but many times these can be extraneous, a middle layer. When we talk about models, the representation becomes integral to the model itself. This is an issue we need to be aware with when dealing with ontologies for models.

Boissel: The next issue is modularity and modules. I propose two different definitions. The first is breaking down the problem into autonomous and homogeneous sub problems. These are the modules. An extension to this definition is the answer to the question ‘is life a combination of modules’?

Subramaniam: That is far too philosophical! However, I’d like to focus on a question that is raised by Raimond Winslow (Winslow et al 2002, this volume) and Les Loew’s (Loew 2002, this volume) papers: how do we define signalling modules and pathway modules?

Boissel: It is important to be clear about what we mean when we say that a particular model has ‘modules’. For me the definition of ‘module’ is an operational

one: it is only defined as the process of modelling progresses, rather than by an a priori definition.

Noble: Combination, of course, is a kind of logic. Another way of putting your question is one that I hope that we will return to in the concluding discussion: could there be a 'theoretical biology' in the grand sense? This has to do with the question, is there a logic of life? I once edited a book called 'The logic of life' (Boyd & Noble 1993). The main criticism I received of this book is that the title was wrong. It should have been 'The logics of life', on the grounds that what has happened is that evolution has found various ways of making combinations that happen to work. This means, of course, that they get selected for. Since this is a haphazard affair, there won't be one logic. I think it is an open question, but it is one perhaps we should return to later when we discuss whether a theoretical biology exists.

Boissel: Are you satisfied with the vague definition of modules we have so far?

Subramaniam: I don't like the terms 'autonomous' and 'homogeneous'. We don't need these for defining modules. You can have heterogeneous or non-autonomous modules.

Loew: Ideally, you would want them to be autonomous because this makes the world much easier to deal with, but practically they are not. The real question is how do we deal with modules that are not autonomous?

Asbburner: There could be dependencies between modules.

McCulloch: Also, there are different types of modules. There are structural modules such as the cells and subcellular compartments, and there are functional modules or subsystems.

Cassman: You can think of them as autonomous in the sense of getting modules that have inputs and outputs that are not dependent on any other external interactions. The architecture that you define defines the output based on a specific input, without regard to any other component. There are always going to be some feedback processes or other interactions that could modulate this, but it can be defined as an integral system that doesn't require other kinds of components.

Asbburner: But autonomy is not a necessary condition for modularity.

Cassman: It depends on what you mean by 'autonomy'.

Asbburner: Lack of dependencies.

Cassman: I think it is lack of dependency for something. It doesn't have to be lack of dependency for everything. For example, bacterial chemotaxis is a modular component that is certainly not removed from the rest of the organism, but within that context of interactions you can understand the output without regard to any other involvement.

Subramaniam: It all depends on how you define it. The MAP kinase cascade is a module, but it is not autonomous, because it can go from one system to another.

Boissel: Tying down a definition for ‘integration’ is likely to be difficult. I have two definitions of this term that stem from what we heard yesterday. The first is incorporating components through quantitative relationships. The second is combining all available knowledge and evidence. Which is best?

Noble: There’s a problem with the second one, in that all models are partial representations of reality. This is necessarily so, because a complete representation of reality is a clone of reality.

Paterson: You are making a very important distinction: there is integration from the perspective of the pieces that make the whole, and there is integration from the perspective of looking at available knowledge, data and evidence. The data or evidence may relate to phenomena that come from a component in isolation, and phenomena that come from the integrated whole. This is describing integration from both the perspective of the components that make up the whole, and then the behaviours and phenomena generated by the components versus those generated by the integrated whole. Having both pieces is useful.

Boissel: So would you keep the two definitions?

Paterson: Yes.

Noble: I like my own integrators! But this flippant remark is only to indicate that there is a problem with the term ‘integration’. I tend to distinguish between levels of modelling, although I take the point made earlier that things can be wrapped together. Nevertheless, in biological work one has to be data rich at some fairly low level of modelling. I see the middle level, before we get to functional interpretation and explanation, as being the integrative level. What you describe as incorporating components with quantitative relations, I would describe as computing the functionality that emerges through those components talking to each other. If I model a pacemaker mechanism, for example, my descriptive data-rich level consists of all the equations for the transporters that are thought to be contributing current to that particular phenomenon. The integrative level will involve connecting those together in the model so that you can literally integrate the equations. What you are also doing is integrating through functionality, and what you hope will emerge out of that will be the oscillation that is the pacemaker phenomenon. Incidentally, when I first asked to use a computer way back in 1960, and I had convinced the bearded computer scientists that I wasn’t going to waste time on their precious machine, the first question they asked was one of puzzlement at the fact that the 1962 Noble equations lacked an oscillator. If I had been sophisticated, rather than a young graduate student, I would have said that this is a phenomenon that is going to emerge by integration. Of course, what integration is doing is to bring out functionality. This will emerge, as presumably it did in evolution, through those interactions.

Boissel: Could we say then that integration is the process of moving from a purely descriptive state to an explanatory one?

Subramaniam: Not necessarily. You can have emergent properties as a consequence of integration.

Noble: And you may even be puzzled as to why. This is not yet an explanation.

Boissel: The next term is ‘robustness’. Yesterday, again, I heard two different definitions. First, insensitivity to parameter values; second, insensitivity to uncertainty. I like the second but not the first.

Noble: In some cases you would want sensitivity. No Hodgkin–Huxley analysis of a nerve impulse would be correct without it being the case that at a certain critical point the whole thing takes off. We will need to have sensitivity to some parameter values.

Boissel: For me, insensitivity to parameter values means that the parameters are useless in the model.

Cassman: In those cases (at least, the fairly limited number where this seems to be true) it is the architecture of the system that determines the output and not the specific parameter values. It seems likely this is only true for certain characteristic phenotypic outcomes. In some cases it exists, in others it doesn’t.

Hinzb: Perhaps a better way of saying this is insensitivity to ill-defined parameter values. In some models there are parameters that are not well defined, which is the case in a lot of signalling networks. In contrast, in a lot of electrophysiology they are well defined and then the model doesn’t have to be robust to a well defined parameter.

Loew: Rather than uncertainty, a better concept for our discussion might be variability. That is, because of differences in the environment and natural variability. We are often dealing with a small number of molecules. There is therefore a certain amount of uncertainty or variability that is built into biology. If a biological system is going to work reliably, it has to be insensitive to this variability.

Boissel: That is different from uncertainty, so we should add variability here.

Paterson: It is the difference between robustness of a prediction versus robustness of a system design. Robustness of a system design would be insensitivity to variability. Robustness of a prediction, where you are trying to make a prediction based on a model with incomplete data is more the uncertainty issue.

Maini: It all depends what you mean by parameter. Parameter can also refer to the topology and networking of the system, or to boundary conditions. There is a link between the parameter values and the uncertainty. If your model only worked if a certain parameter was 4.6, biologically you could never be certain that this parameter was 4.6. It might be 4.61. In this case you would say that this was not a good model.

Boissel: There is another issue regarding uncertainty, which is the strength of evidence of the data that have been used to parameterize the model. This is a difficult issue.

References

- Boyd CAR, Noble D 1993 *The logic of life*. Oxford University Press, Oxford
- Loew L 2002 The Virtual Cell project. In: '*In silico*' simulation of biological processes. Wiley, Chichester (Novartis Found Symp 247) p 151–161
- Winslow RL, Helm P, Baumgartner W Jr et al 2002 Imaging-based integrative models of the heart: closing the loop between experiment and simulation. In: '*In silico*' simulation of biological processes. Wiley, Chichester (Novartis Found Symp 247) p 129–143

Imaging-based integrative models of the heart: closing the loop between experiment and simulation

Raimond L. Winslow*, Patrick Helm*, William Baumgartner Jr.*, Srinivas Peddi†, Tilak Ratnanather‡, Elliot McVeigh‡ and Michael I. Miller†

**The Whitaker Biomedical Engineering Institute Center for Computational Medicine & Biology and †Center for Imaging Sciences, ‡NIH Laboratory of Cardiac Energetics: Medical Imaging Section 3, Johns Hopkins University, Baltimore MD 21218, USA*

Abstract. We describe methodologies for: (a) mapping ventricular activation using high-density epicardial electrode arrays; (b) measuring and modelling ventricular geometry and fibre orientation at high spatial resolution using diffusion tensor magnetic resonance imaging (DTMRI); and (c) simulating electrical conduction; using comprehensive data sets collected from individual canine hearts. We demonstrate that computational models based on these experimental data sets yield reasonably accurate reproduction of measured epicardial activation patterns. We believe this ability to electrically map and model individual hearts will lead to enhanced understanding of the relationship between anatomical structure, and electrical conduction in the cardiac ventricles.

2002 'In silico' *simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 129–143*

Cardiac electrophysiology is a field with a rich history of integrative modelling. A critical milestone for the field was the development of the first biophysically based cell model describing interactions between voltage-gated membrane currents, pumps and exchangers, and intracellular calcium (Ca^{2+}) cycling processes (DiFrancesco & Noble 1985), and the subsequent elaboration of this model to describe the cardiac ventricular myocyte action potential (Noble et al 1991, Luo & Rudy 1994). The contributions of these and other models to understanding of myocyte function have been considerable, and are due in large part to a rich interplay between experiment and modelling—an interplay in which experiments inform modelling, and modelling suggests new experiments.

Modelling of cardiac ventricular conduction has to a large extent lacked this interplay. While it is now possible to measure electrical activation of the epicardium at relatively high spatial resolution, the difficulty of measuring the geometry and fibre structure of hearts which have been electrically mapped has

limited our ability to relate ventricular structure to conduction via quantitative models. We believe there are four major tasks that must be accomplished if we are to understand this structure–function relationship. First, we must identify an appropriate experimental preparation—one which affords the opportunity to study effects of remodelling of ventricular geometry and fibre structure on ventricular conduction. Second, we must develop rapid, accurate methods for measuring both electrical conduction, ventricular geometry and fibre structure in the same heart. Third, we must develop mathematical approaches for identifying statistically significant differences in geometry and fibre structure between hearts. Fourth, once identified, these differences in geometry and fibre structure must be related to differences in conduction properties.

We are pursuing these goals by means of coordinated experimental and modelling studies of electrical conduction in normal canine heart, and canine hearts in which failure is induced using the tachycardia pacing-induced procedure (Williams et al 1994). In the following sections, we describe the ways in which we: (a) map ventricular activation using high-density epicardial electrode arrays; (b) measure and model ventricular geometry and fibre orientation at high spatial resolution using diffusion tensor magnetic resonance imaging (DTMRI); and (c) construct computational models of the imaged hearts; and (d) compare simulated conduction properties with those measured in the same heart.

Mapping of epicardial conduction in normal and failing canine heart

In each of the three normal and three failing canine hearts studied to date, we have, prior to imaging, performed electrical mapping studies in which epicardial conduction in response to various current stimuli are measured using multi-electrode epicardial socks consisting of a nylon mesh with 256 electrodes and electrode spacing of ~ 5 mm sewn around its surface. Bipolar epicardial twisted-pair pacing electrodes are sewn onto the right atrium (RA) and the right ventricular (RV) free-wall. Four to 10 glass beads filled with gadolinium-DTPA (~ 5 mM) are attached to the sock as localization markers, and responses to different pacing protocols are recorded. Figure 1A shows an example of measurement of activation time (colour bar, in ms) measured in response to an RV stimulus pulse applied at the epicardial locations marked in red. After all electrical recordings are obtained, the animal is euthanatized with a bolus of potassium chloride, and the heart is then scanned with high-resolution T1-weighted imaging in order to locate the gadolinium-DTPA filled beads in scanner coordinates. The heart is then excised, sock electrode locations are determined using a 3D digitizer (MicroScribe 3DLX), and the heart is formalin-fixed in preparation for DTMRI.

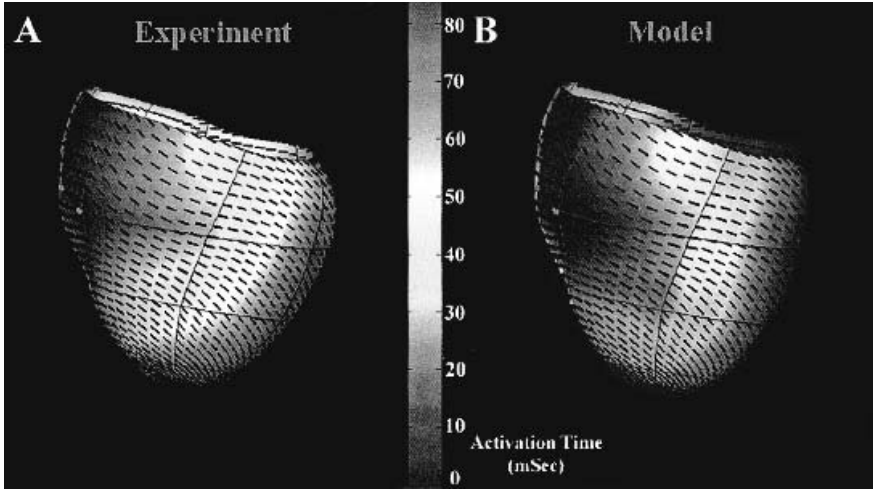


FIG. 1. (A) Electrical activation times (indicated by grey scale) in response to right RV pacing as recorded using electrode arrays. Data was obtained from a normal canine heart that was subsequently reconstructed using DTMRI. Activation times are displayed on the epicardial surface of a finite-element model fit to the DTMRI reconstruction data. Fibre orientation on the epicardial surface, as fit to the DTMRI data by the FEM model, is shown by the short line segments. (B) Activation times predicted using a computational model of the heart mapped in (A).

Measuring the fibre structure of the cardiac ventricles using DTMRI

DTMRI is based on the principle that proton diffusion in the presence of a magnetic field gradient causes signal attenuation, and that measurement of this attenuation in several different directions can be used to estimate a diffusion tensor at each image voxel (Skejskal 1965, Basser et al 1994). Several studies have now confirmed that the principle eigenvector of the diffusion tensor is locally aligned with the long-axis of cardiac fibres (Hsu et al 1998, Scollan et al 1998, Holmes et al 2000).

Use of DTMRI for reconstruction of cardiac fibre orientation provides several advantages over traditional histological methods. First, DTMRI yields estimates of the absolute orientation of cardiac fibres, whereas histological methods yield estimates of only fibre inclination angle. Second, DTMRI performed using formalin-fixed tissue: (a) yields high resolution images of the cardiac boundaries, thus enabling precise reconstruction of ventricular geometry using image segmentation software; and (b) eliminates flow artefacts present in perfused heart, enabling longer imaging times, increased signal-to-noise (SNR) ratio and improved spatial resolution. Third, DTMRI provides estimates of fibre orientation at greater than one order of magnitude more points than possible with histological methods. Fourth, reconstruction time is greatly reduced (~ 60 h versus weeks to months) relative to that for histological methods.

DTMRI data acquisition and analysis for ventricular reconstruction has been semi-automated. Once image data are acquired, software written in the MatLab programming language is used to estimate epicardial and endocardial boundaries in each short-axis section of the image volume using either the method of region growing or the method of parametric active contours (Scollan et al 2000). Diffusion tensor eigenvalues and eigenvectors are computed from the DTMRI data sets at those image voxels corresponding to myocardial points, and fibre orientation at each image voxel is computed as the primary eigenvector of the diffusion tensor.

Representative results from imaging of one normal and one failing heart are shown in Fig. 2. Figures 2A & C are short-axis basal sections taken at approximately the same level in normal (2A) and failing (2C) canine hearts. These two plots show regional anisotropy according to the indicated colour code. Figures 2B & D show the angle of the primary eigenvector relative to the plane of section (inclination angle), according to the indicated colour code, for the same sections as in Figs 2A & C. Inspection of these data show: (a) the failing heart (HF; panels C & D) is dilated relative to the normal heart (N; panels A & B); (b) left ventricular (LV) wall thinning (average LV wall thickness over three hearts is 17.5 ± 2.9 mm in N, and 12.9 ± 2.8 mm in HF); (c) no change in RV wall thickness (average RV wall thickness is 6.1 ± 1.6 mm in N, and 6.3 ± 2.1 mm in HF); (d) increased septal wall thickness HF versus N (average septal wall thickness is 14.7 ± 1.2 mm N, and 19.7 ± 2.1 mm HF); (e) increased septal anisotropy in HF versus N (average septal thickness is 0.71 ± 0.15 N, and 0.82 ± 0.15 HF); and (f) changes in the transmural distribution of septal fibre orientation in HF versus N (contrast panels B & D, particularly near the junction of the septum and RV).

Finite-element modelling of cardiac ventricular anatomy

Structure of the cardiac ventricles is modelled using finite-element modelling (FEM) methods developed by Nielsen et al (1991). The geometry of the heart to be modelled is described initially using a predefined mesh with six circumferential elements and four axial elements. Elements use a cubic Hermite interpolation in the transmural coordinate (λ), and bilinear interpolation in the longitudinal (μ) and circumferential (θ) coordinates. Voxels in the 3D DTMRI images identified as being on the epicardial and endocardial surfaces by the semi-automated contouring algorithms described above are used to deform this initial FEM template. Deformation of the initial mesh is performed to minimize an objective function $F(\underline{n})$.

$$F(\underline{n}) = \sum_{d=1}^D \gamma_d \|\nabla(\underline{\varepsilon}_d) - \mathbf{v}_d\|^2 + \int_{\mathbb{R}^2} \{\alpha \nabla^2 \underline{n} + \beta (\nabla^2 \underline{n})^2\} \partial \underline{\varepsilon}, \quad (1)$$

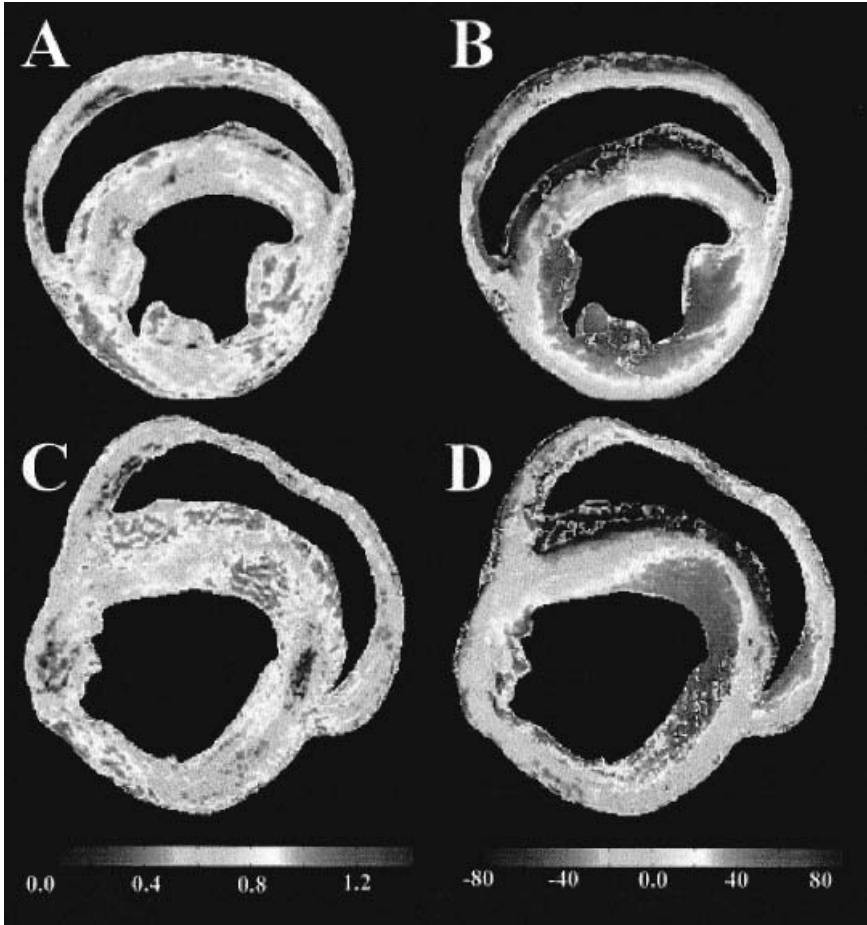


FIG. 2. Fibre anisotropy $\Lambda(\underline{x})$, computed as:

$$\Lambda(\underline{x}) = \sqrt{\frac{[\lambda_1(\underline{x}) - \lambda_2(\underline{x})]^2 + [\lambda_1(\underline{x}) - \lambda_3(\underline{x})]^2 + [\lambda_2(\underline{x}) - \lambda_3(\underline{x})]^2}{\lambda_1(\underline{x})^2 + \lambda_2(\underline{x})^2 + \lambda_3(\underline{x})^2}}$$

where $\lambda_i(\underline{x})$ are diffusion tensor eigenvectors at voxel \underline{x} , in normal (A) and failing (C) canine heart. Fibre inclination angle computed using DTMRI in normal (B) and failing (D) heart. Panels (A) and (B) are the same normal, and panels (C) and (D) the same failing heart.

where \mathbf{n} is a vector of mesh nodal values, \mathbf{v}_d are the surface voxel data, $\mathbf{v}(\epsilon_d)$ are the projections of the surface voxel data on the mesh, and α and β are user defined constants. This objective function consists of two terms. The first describes distance between each surface image voxel (\mathbf{v}_d) and its projection onto the mesh $\mathbf{v}(\epsilon_d)$. The second, known as the weighted Sobelov norm, limits stretching (first

derivative terms) and the bending (second derivative terms) of the surface. The parameters α and β control the degree of deformation of each element. The weighted Sobelov norm is particularly useful in cases where there is an uneven distribution of surface voxels across the elements. A linear least squares algorithm is used to minimize this objective function.

After the geometric mesh is fitted to DTMRI data, the fibre field is defined for the model. Principle eigenvectors lying within the boundaries of the mesh computed above are transformed into the local geometric coordinates of the model using the following transformation.

$$\underline{V}_G = [\underline{F} \ \underline{G} \ \underline{H}]^T [\underline{R}] \underline{V}_S \quad (2)$$

where \underline{R} is a rotation matrix that transforms a vector from scanner coordinates (\underline{V}_S) into the FEM model coordinates \underline{V}_G and \underline{F} , \underline{G} , \underline{H} are orthogonal geometric unit vectors computed from the ventricular geometry as described by LeGrice et al (1997). Once the fibre vectors are represented in geometric coordinates, DTMRI inclination and imbrication angles (α and ϕ) are fit using a bilinear interpolation in the local ε_1 and ε_2 coordinates, and a cubic Hermite interpolation in the ε_3 coordinate. A graphical user interface for fitting FEMs to both the ventricular surfaces and fibre field data has been implemented using the MatLab programming language. Figure 3 shows FEM fits to the epicardial/endocardial surfaces of a reconstructed normal canine heart (Fig. 1A is also an FEM). FEM fits to the fibre orientation data are shown on these surfaces as short line segments.

We have developed relational database and data analysis software named *HeartScan* to facilitate analysis of cardiac structural and electrical data sets obtained from populations of hearts. *HeartScan* enables users to pose queries (in standard query language, or SQL) on a wide range of cardiac data sets by means of a graphical user interface. These data sets include: (a) DTMRI imaging data; (b) FEMs derived from DTMRI data; (c) electrical mapping data obtained using epicardial electrode arrays; (d) model simulation data. Query results are either: (a) displayed on a 3D graphical representation of the heart being analysed; or (b) piped to data processing scripts, the results of which are then displayed visually. Queries may be posed by direct entry of an SQL command into the Query Window (Fig. 4B). This query is executed, and the set of points satisfying this condition are displayed on a wire frame model of the heart being studied (Fig. 4C). Queries operating on a particular region of the heart may also be entered by graphically selecting that region (Fig. 4D). SQL commands specifying the coordinates of the selected voxels are then automatically entered into the Query Window. One example of such a predefined operation is shown in Fig. 4E, which shows computation of transmural inclination angle for the region enclosed by the box in Fig. 4D.

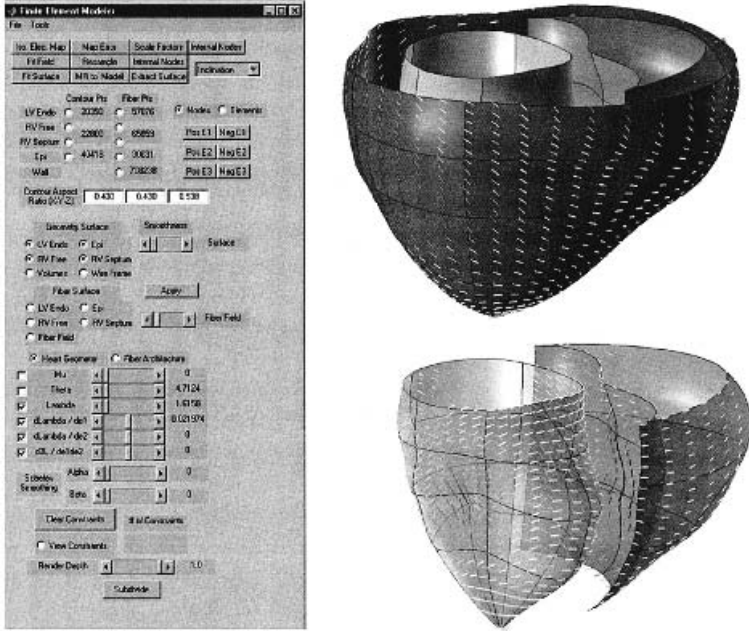


FIG. 3. Finite-element model of canine ventricular anatomy showing the epicardial, LV endocardial and RV endocardial surfaces. Fibre orientation on each surface is shown by short line segments.

Statistical comparison of anatomical differences between hearts

In order to assess anatomical differences between hearts and their effects on ventricular conduction, we must first understand how to bring different hearts into registration, and how to identify statistically significant local and global differences in cardiac structure over ensembles of hearts. Approaches for addressing these issues are being developed in the emerging field of computational anatomy—the discipline of computing transformations ϕ between different anatomical configurations (Grenander & Miller 1998). The transformations ϕ satisfy Eulerian and Lagrangian equations of mechanics so as to generate consistent movement of anatomical coordinates. They are constrained to be one-to-one and differentiable with a differentiable inverse, so that connected sets in the template remain connected in the target, surfaces are transformed as surfaces, and the global relationships between structures are maintained. Transformations can include: (a) translation, rotation and expansion/contraction; (b) large deformation landmark transformations; and (c) high dimensional large deformation image matching transformations. Because of the difficulty in identifying reliable ventricular landmarks as a guide for designing

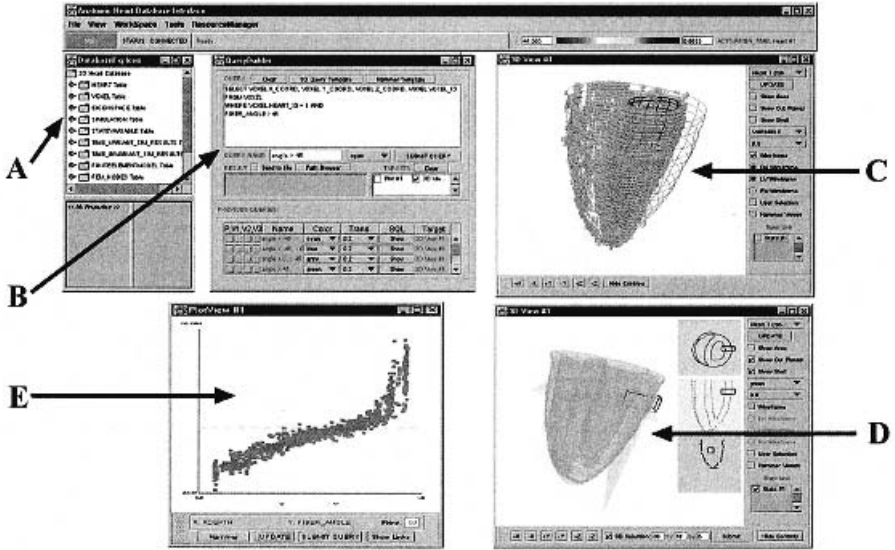


FIG. 4. ‘Screenshot’ of the windows by which the user interacts with *HeartScan*. (A) window for viewing data tables; (B) SQL query window; (C) window for interactive 3D display of heart data; (D) pull-down window for user selection of heart regions to query. (E) statistics display window.

transformations, we use landmark-free transformations that are compositions of rigid and linear motions (a), and that rely on intrinsic image properties such as intensity and connectedness of points (c). These transformations are applied as maps of increasingly higher dimension, generated one after another through composition (Matejic 1997).

The transformations $\phi \in H$ are defined on the space of homeomorphisms constructed from the vector field $\phi : (x_1, x_2, x_3)^3 \mapsto (\phi_1(x), \phi_2(x), \phi_3(x)) \in \Omega$, with inverse $\phi^{-1} \in H$. These transformations evolve in time $t \in [0, 1]$ to minimize a penalty function, and are controlled by the velocity field $v(\cdot, \cdot)$. The flow is given by the solution to the transport equations

$$\begin{aligned} \frac{d\phi(x, t)}{dt} &= v(\phi(x, t), t), & \phi(x, 0) &= x, & \frac{\partial \phi^{-1}(x, t)}{\partial t} &= -\nabla'_x \phi^{-1}(x, t) v(x, t), \\ \phi^{-1}(x, 0) &= x \end{aligned} \quad (3)$$

where

$$\nabla'_x = \left[\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3} \right] \quad (4)$$

The metric distance between two anatomical configurations I_0 and I_1 is given by the geodesic length $\rho(I_0, I_1)$ between them (Trounev 1998, Miller & Younes 2002)

$$\rho(I_0, I_1) = \inf_v \|Lv\|^2 \tag{5}$$

where L is the Cauchy–Navier operator.

Since all the imagery being matched are observed with noise, they are modelled as conditional Gaussian random fields. Take I_0 as the template. The target imagery I_1 is therefore a conditionally Gaussian random field with mean field given by the template composed with the unknown invertible map $I_0 \circ \phi$, and fixed variance. The problem is to estimate the velocity field which matches I_0 to the observable image I_1 , subject to constraints, with minimum penalty. The optimal matching of I_0 to observation I_1 is given by the $d\hat{\phi}/dt = \hat{v}(\hat{\phi})$ from Eq. (3) which satisfies the extremum problem

$$\hat{v}(\square) = \arg \inf_v \|Lv\|^2 + \|I_0 \circ \phi^{-1}(1) - I_1\|^2 \tag{6}$$

The cost is chosen as

$$\|I_0 \circ \hat{\phi}^{-1}(1) - I_1\|^2 = \int_{[0,1]^3} |I_0(\hat{\phi}^{-1}(x,1)) - I_1(x)|^2 dx \tag{7}$$

The Euler–Lagrange equations for the extremum problem for the mapping (Miller & Younes 2002) are then given by:

$$\begin{aligned} (I_1(x) - I_0(\phi(x,1)))\nabla I_0(\phi(x,1))(\nabla\phi)^{-1}(x,1) &= Lv(x,1) \\ \frac{\partial Lv(x,t)}{\partial t} + v \cdot Lv(x,t) + \nabla \cdot vL(x,t) + v \cdot \nabla Lv(x,t) + Lv\nabla v &= 0 \end{aligned} \tag{8}$$

A gradient-based computational algorithm is used to solve the Euler–Lagrange equations.

Figure 5 show preliminary results on computation of transformations ϕ which align a three-dimensional template (failing) and target (normal) cardiac ventricular geometry. In each figure, the left column shows a transverse section from the template (top) and target (bottom). The top panel of the middle column shows the result of applying the forward mapping ϕ to the template in order to map points in this template to points in the target. The bottom panel shows the result of applying the inverse mapping ϕ to the target to take this target back into the template. The right column shows the displacements associated with the transformations ϕ and ϕ^{-1} . These transformations were computed without using any anatomical landmarks to align the images. Note the dilation (indicating by

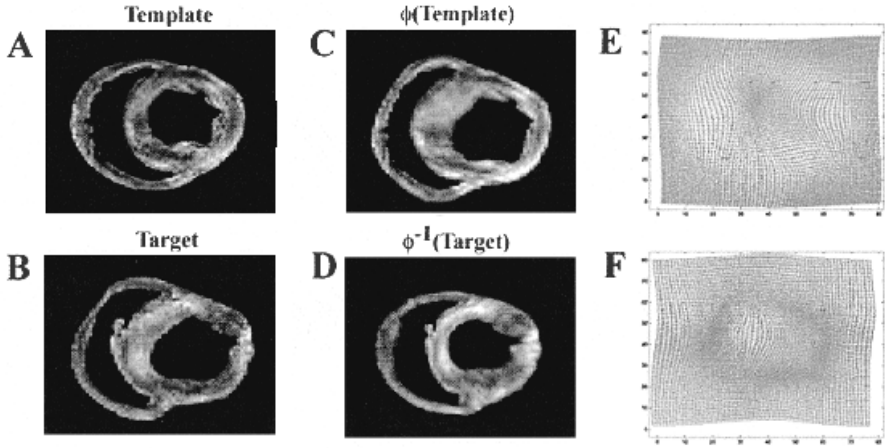


FIG. 5. Transformation of a normal heart (template) transverse section to a failing heart (target). The left column shows the template (A) and target (B), the middle column shows the result of applying the forward mapping ϕ to the template (C), and the inverse mapping ϕ^{-1} to the target (D); the right column shows the grids deformed by the forward mapping ϕ (E) and the inverse mapping ϕ^{-1} (F).

spreading of the lines between grid points) and compression associated with the forward and inverse maps, respectively. Also note that in both figures, the template image is similar to the inverse transformed target image (template $\sim \phi^{-1}$ (target)) and the target image is similar to the forward transformed template (target $\sim \phi$ (template)).

We have not yet reconstructed sufficiently large populations of normal and failing hearts to perform meaningful statistical analyses of anatomic variation. However, the theoretical approach to this problem will be that applied previously to the analysis of hippocampal shape variation, in which anatomical shapes are characterized as Gaussian fields indexed over the manifolds on which the vector fields are defined (Amit & Picconi 1991, Joshi et al 1997, Miller et al 1997, Grenander & Miller 1998).

Three-dimensional modelling of electrical conduction in the cardiac ventricles

Electrical conduction in the ventricles is modelled using the monodomain equation:

$$\frac{\partial v(\underline{x}, t)}{\partial t} = \frac{1}{C_m} \left[-I_{ion}(\underline{x}, t) - I_{app}(\underline{x}, t) + \frac{1}{\beta} \left(\frac{\kappa}{\kappa + 1} \right) \nabla \cdot (M_i(\underline{x}) \nabla v(\underline{x}, t)) \right] \text{ on } H \quad (9)$$

where $I_{ion}(x,t)$ is membrane ionic current as defined in the canine myocyte model of Winslow et al (1999). The conductivity tensors at each myocardial point x are then defined as

$$M_i(x) = P(x)G_i(x)P^T(x), \quad (10)$$

where $G_i(x)$ is a diagonal matrix with elements $\sigma_{1,i}$, $\sigma_{2,i}$ and $\sigma_{3,i}$, ($\sigma_{1,i}$ is longitudinal, and $\sigma_{2,i}$ and $\sigma_{3,i}$ are transverse intracellular conductivities), and $P(x)$ is the transformation matrix from local to scanner coordinates at each point x (Winslow et al 2000, 2001). When working from DTMRI data, the columns of $P(x)$ are set equal to the eigenvectors of the diffusion tensor estimated at point x (Winslow et al 2000, 2001). Coupling conductances are set as in previous models (Henriquez 1993, Henriquez et al 1996), and refined to yield measured epicardial conduction velocities. Presently, coupling conductances are assumed to be transversely isotropic. The reaction–diffusion monodomain equation (Eqs. 9–10) are solved using methods described previously (Yung 2000).

Figure 1 shows the results of applying these methods to the analysis of conduction in a normal canine heart. As described previously, Fig. 1A shows activation time (greyscale, in ms) measured in response to an RV stimulus pulse applied at the epicardial locations marked by the dots. Following electrical mapping, this heart was excised, imaged using DTMRI, and an FEM was then fit to the resulting geometry and fibre orientation data sets. Figure 1A shows activation time displayed on this FEM. The stimulus wave front can be seen to follow the orientation of the epicardial fibres, which is indicated by the dark line segments in Fig. 1A. Fig. 1B shows results of simulating conduction using a computational model of the very same heart that was mapped electrically in Fig. 1A. Results can be seen to agree qualitatively, however model conduction is more rapid in the region where the RV and LV join.

Discussion

In this paper, we have presented a methodology for the electrical mapping, structural modelling and analysis, and electrical modelling of the cardiac ventricles. This methodology is based on the use of high density electrode arrays to measure epicardial conduction properties in response to well defined stimuli, DTMRI to map ventricular geometry and fibre organization, and computational modelling to predict electrical activation in response to the same stimuli used experimentally, all in the same heart. Using these methods, we can now test the hypothesis that the three-dimensional models of the cardiac ventricles can quantitatively reproduce conduction patterns measured in the same hearts that are modelled. While these initial studies have been limited to comparison of

epicardial conduction properties, use of plunge and endocardial basket catheter electrodes will ultimately enable more extensive comparisons of 3D conduction properties between model and experiment. It will also be possible to use MR spin-tagging procedures to collect data on mechanical motion in the same hearts that are electrically mapped and modelled. While there are certainly additional modifications that must be made to the computational models (such as addition of a Purkinje network), we believe the ability to collect such comprehensive data sets in each heart studied will lead to enhanced understanding of the relationship between anatomical structure, electrical conduction, and mechanics of the cardiac ventricles.

Acknowledgements

Supported by NIH RO1 HL60133 and P50 HL52307, The Whitaker Foundation, The Falk Foundation, and IBM Corporation. Owen Faris assisted with electrical mapping studies.

References

- Amit Y, Picconi M 1991 A nonhomogenous Markov process for the estimation of Gaussian Random Fields with non-linear observations. *Ann Prob* 19:1664–1678
- Basser PJ, Mattiello J, LeBihan D 1994 Estimation of the effective self-diffusion tensor from the NMR spin echo. *J Magn Reson B* 103:247–254
- DiFrancesco D, Noble D 1985 A model of cardiac electrical activity incorporating ionic pumps and concentration changes. *Philos Trans R Soc Lond B Bio Sci* 307:353–398
- Grenander U, Miller MI 1998 Computational anatomy: An emerging discipline. *Quart J Mech Appl Math* 56:617–694
- Henriquez CS 1993 Simulating the electrical behavior of cardiac tissue using the bidomain model. *Crit Rev Biomed Eng* 21:1–77
- Henriquez CS, Muzikant AL, Smoak CK 1996 Anisotropy, fiber curvature, and bath loading effects on activation in thin and thick cardiac tissue preparations: simulations in a three-dimensional bidomain model. *J Cardiovasc Electrophysiol* 7:424–444
- Holmes AA, Scollan DF, Winslow RL 2000 Direct histological validation of diffusion tensor MRI in formaldehyde-fixed myocardium. *Magn Reson Med* 44:157–161
- Hsu EW, Muzikant AL, Matulevicius SA, Penland RC, Henriquez CS 1998 Magnetic resonance myocardial fiber-orientation mapping with direct histological correlation. *Am J Physiol* 274:H1627–1634
- Joshi SC, Miller MI, Grenander U 1997 On the geometry and shape of brain sub-manifolds. *Intern J Pattern Recognit Artif Intel* 11:1317–1343
- LeGrice IJ, Hunter PJ, Smail BH 1997 Lamina structure of the heart: a mathematical model. *Am J Physiol* 272:H2466–H2476
- Luo CH, Rudy Y 1994 A dynamic model of the cardiac ventricular action potential: I. Simulations of ionic currents and concentration changes. *Circ Res* 74:1071–1096
- Matejic L 1997 Group cascades for representing biological variability. Brown University, Providence, MA
- Miller MI, Trounev A, Younes L 2002 On the metrics and Euler-Lagrange equations of computational anatomy. *Annu Rev Biomed Eng* 4:375–405

- Miller MI, Banerjee A, Christensen GE et al 1997 Statistical methods in computational anatomy. *Stat Methods Med Res* 6:267–299
- Nielsen PM, LeGrice IJ, Smaill BH, Hunter PJ 1991 Mathematical model of geometry and fibrous structure of the heart. *Am J Physiol* 260:H1365–H1378
- Noble DS, Noble SJ, Bett GC, Earm YE, Ho WK, So IK 1991 The role of sodium-calcium exchange during the cardiac action potential. *Ann N Y Acad Sci* 639:334–353
- Scollan DF, Holmes A, Winslow R, Forder J 1998 Histological validation of myocardial microstructure obtained from diffusion tensor magnetic resonance imaging. *Am J Physiol* 275:H2308–H2318
- Scollan D, Holmes A, Zhang J, Winslow R 2000 Reconstruction of cardiac ventricular geometry and fiber orientation using magnetic resonance imaging. *Ann Biomed Eng* 28:934–944
- Skejskal EA 1965 Spin diffusion measurement: spin echoes in the presence of time-dependent field gradients. *J Chem Phys* 69:1748–1754
- Trounev A 1998 Diffeomorphisms groups and pattern matching in image analysis. *Int J Comput Vis* 28:213–221
- Williams RE, Kass DA, Kawagoe Y et al 1994 Endomyocardial gene expression during development of pacing tachycardia-induced heart failure in the dog. *Circ Res* 75:615–623
- Winslow RL, Rice JJ, Jafri MS, Marban E, O'Rourke B 1999 Mechanisms of altered excitation–contraction coupling in canine tachycardia-induced heart failure, II. Model studies. *Circ Res* 84:571–586
- Winslow RL, Scollan DF, Holmes A, Yung CK, Zhang J, Jafri MS 2000 Electrophysiological modeling of cardiac ventricular function: from cell to organ. *Ann Rev Biomed Eng* 2:119–155
- Winslow R, Scollan D, Greenstein J et al 2001 Mapping, modeling and visual exploration of structure–function relationships in the heart. *IBM Syst J* 40:342–359
- Yung C 2000 Application of a stiff, operator-splitting scheme to the computational modeling of electrical properties in cardiac ventricles. Masters of Engineering Theses, Department of Biomedical Engineering, The Johns Hopkins University, Baltimore MD

DISCUSSION

Hunter: Presumably, if you are looking beyond 20–30 ms you will be reactivating Purkinje networks. Is there any way you can get some assessment from these hearts of the different topology of a Purkinje network? If you are using the mapping data to do this comparison and you try to match the models to it, you are going to be in trouble if you can't deal with the role of Purkinjes involved in that.

Winslow: I am not sure whether there is a precise way. We can change conduction velocity in the entire endocardial surface. This is a crude approximation for a Purkinje network. Or perhaps we could use one of the models that have mapped the conduction network in a particular heart and try to use this. But I don't know any way of specifically marking the Purkinje cells so we could see that network in the same heart that we are imaging.

McCulloch: There are established histological methods.

Hunter: The same question would apply to the sheet structure, whether under those heart failure conditions you would see substantial changes in the second eigenvector. Is there any way you could get information on that?

Winslow: We have this hypothesis that the second and third eigenvectors are within sheet and are the surface normal to a sheet. We came to this hypothesis by taking data in rabbit and plotting these angles of the surface normal. We looked at those angles compared with your histologically reconstructed canine data. Qualitatively, they looked similar. We then passed a data set to Andrew McCulloch. Unfortunately it was one of our very first imaging data sets and was not of high quality. Andrew actually performed a reconstruction of sheet orientation in regions of this data set and the correspondence was partial. The difficulty for us in testing this hypothesis about what these second and third eigenvectors are telling us is our inability to perform these very complicated sheet reconstructions.

McCulloch: It turns out to be much more difficult to do in the rabbit than the dog. It might be better to use the new high-resolution canine data. I have a related question. You described a surprising change in the apparent fibre orientation in the septum in the failing dog hearts. Could this be due to something other than a change in the principal axis of the myocytes, and instead due to some of what the cardiology literature refers to as slippage? This is presumably some sort of shearing between adjacent sheets, as opposed to a genuine change in the vectorial orientation of the myocyte. Or do you think that this really does represent a reorientation of myofibrils and myocytes in that area?

Winslow: I would think that a change in sheet structure would be more reflected in properties of the second and third eigenvectors, if our hypothesis were correct. We haven't paid any attention to these data yet; we have been focused on the information that we think relates to fibre structure. It is a very clear change not so much in the magnitude of diffusion in the direction of the principle eigenvector, but a massive change in the direction of that vector itself. I would think this would have to correspond to a reorientation of the fibres themselves. That reorientation may have something to do with the way in which that heart was paced into failure, with the location of the pacing electrode or the particular pacing rate and parameters. It would probably be worthwhile looking at a different model of heart failure in the dog to test this hypothesis, and also to look at the human data to see whether this feature is still present.

McCulloch: A related question: could it be connected to a remodelling of vascular or microvascular architecture?

Winslow: I don't think it is. The reason why is related to the reason why we switched from imaging in a perfused preparation to a fixed preparation. There were two reasons for changing from a perfused preparation. First, this preparation limited our imaging time: after 10–12 h imaging these hearts would frequently go into contraction and their geometry would change. Second, Ed Hsu and colleagues looked at the effect of turning off the perfusate to these hearts that were being diffusion imaged. They found that when they represented the

diffusion tensor as being formed by a linear combination of two separate diffusion tensors, the second component of the diffusion tensor went away when the perfusate was shut off. There could therefore have been a perfusion artefact in the hearts that we had originally imaged using this preparation. The fact that this imaging artefact goes away argues that the contribution of flow in vessels is minimal. This contribution can be seen when you look at the raw diffusion imaging data on the surface of the heart. You can see regions of isotropic diffusion that seem to agree with the positioning of coronary arteries on the surface of the heart. If there is an effect, it is probably to corrupt our estimate of fibre angle when we encounter a blood vessel, because it is diffusion in that region that is tending to be isotropic.

Noble: It seems to me that the analysis of cardiac arrhythmia is almost a paradigm example of a disease state in which, without integrating all the way from gene through to whole organ physiology, we can't really say that we have a grip on what it is we are trying to understand. There is simply no stage at which we can say there can be a major gap. It leaves one feeling how audacious it was that we have tried over the last 40 years to develop anti-arrhythmic drugs, without all of this knowledge. Of course, it is not too surprising that we haven't been that successful. The dream must be that eventually one can lead the way back into doing this in a much more rational way.

General discussion III

Modelling Ca^{2+} signalling

Noble: I'd like now to switch to general discussion, and focus on one issue — modelling Ca^{2+} signalling, with a view to addressing a general problem, which is the way in which we can interface different levels or types of modelling. I'd like to ask Raimond Winslow to lead off on this.

Winslow: The kinds of models of cardiac myocytes that we and others have constructed so far do a very good job of describing the electrical behaviour of the cell membrane, and are effective at describing long-term Ca^{2+} cycling processes that occur within the myocyte. However, they do a terrible job of describing accurately the detailed properties of Ca^{2+} release from the sarcoplasmic reticulum (SR) and what drives this release. It is surprising that the myocyte models have been able to do so well in their ability to reproduce and even predict data, given that they don't do a good job describing mechanisms of Ca^{2+} release from SR. After all, this is a fundamental property of the myocyte: the amount of Ca^{2+} released from the SR is graded with 'trigger' Ca^{2+} entering the cell normally, through L-type Ca^{2+} channels. This is important for regulating the force of contraction in the heart. These models can't do that at all, yet they have predictive power. We really couldn't understand how these models work so well, given that they have failed dismally to reproduce this fundamental property of the myocyte.

I would like to describe some results showing the importance of this so-called mechanism of graded release. This speaks to the issue of Ca^{2+} cycling in general, and also the issue of integrating across levels of modelling. What I will present is a stochastic model of Ca^{2+} release that needs to be understood and solved concurrently with a differential equation model of the behaviour of the whole cell. Here we have a problem of combining different model types together and simplifying the stochastic component of the model to make it manageable at the level of the whole cell and for whole-heart simulations.

The key observation regarding Ca^{2+} release from the SR is that this release causes inactivation of L-type Ca^{2+} channels. This is not the only thing that inactivates these channels: they are also voltage inactivated. If the membrane is depolarized, L-type Ca^{2+} channels open, but then they go into an inactivated non-conducting state. If Ca^{2+} is released from the SR, this Ca^{2+} can bind to receptors on the inner pore of the channel and also inactivate them. New data are emerging from Dave

Yue's lab suggesting that the balance between Ca^{2+} inactivation and voltage inactivation is radically different from what was suspected. All existing models of the myocyte describe voltage-dependent inactivation of this channel as being the primary mode. We believe that is wrong, and that it is in fact Ca^{2+} inactivation. Our experimental evidence for this comes from recording Ca^{2+} currents in cultured rat myocytes, and comparing situations in which either Ca^{2+} or Ba^{2+} is the charge carrier. Ba^{2+} is used because it knocks out the inactivation of the L-type Ca^{2+} channel. Ryanodine is also used in these cultured cells to empty the SR of Ca^{2+} , so this is not available to be released by the SR. In the absence of this rapid, strong Ca^{2+} inactivation there is a very weak and slow inactivation component that presumably reflects the voltage-dependent properties of inactivation. In an even better experiment, David Yue used the observation that calmodulin appears to be tethered to the L-type Ca^{2+} channel, and it is this that binds the Ca^{2+} and this complex then interacts with the channel to inactivate it. He has fabricated a mutant calmodulin, which is no longer capable of binding Ca^{2+} , and therefore this is a mechanism for ablating the Ca^{2+} -dependent inactivation. In this case, there is a very slow, long inactivation process that presumably reflects this small amount of voltage inactivation.

Linz & Meyer (1998) have further data that argue for this new idea about a shift in balance between Ca^{2+} inactivation and voltage inactivation. They did an AP clamp recording in isolated cardiac myocytes. They showed that there are lots of channels that aren't voltage inactivated, but there aren't many channels that are not Ca^{2+} inactivated. This indicates that Ca^{2+} inactivation in these native myocytes (as opposed to cultured ones) might be primarily controlled by Ca^{2+} .

Current models differ from this significantly. The Jafri-Rice guinea-pig ventricular myocyte model (Jafri et al 1998) is wrong. Our estimate of the not-voltage-inactivated fraction is very low, and the not- Ca^{2+} -inactivated fraction is way too high. This general conclusion holds for all of these other models. The trouble is, when we take these models and shift the balance between voltage and Ca^{2+} inactivation, we find that they all become unstable. The action potentials alternate between long and short values. We think these models become unstable because they are what Micheal Stern referred to as 'common pool' models. All the Ca^{2+} in the SR of the cell is being represented as being in one compartment; all the Ca^{2+} in the diadic space is lumped into one diadic space; and all the L-type Ca^{2+} channels empty into that one diadic space. These models are not capable of reproducing graded release. The problem here is that when you build in these new physiological data, the models don't work. They can't even predict action potentials.

What we have done is to formulate a new model based on the principle of local control, as investigated by many physiologists and theoreticians. In this model of local control we have individual jSR compartments that are

communicating with an L-type Ca^{2+} channel. There is an individual L-type Ca^{2+} channel that is in communication with a small number (four-to-eight) Ca^{2+} -sensing Ca^{2+} release channels. While Ca^{2+} release at the level of this small functional unit may be all or none, it is the ensemble averaging of these units working in an independent fashion throughout the cell that provides the property of a graded release. For any depolarization of the membrane a certain fraction of these channels will open, and for those that open there is regenerative all-or-none release from the functional unit, but it is the averaging of this behaviour that reflects the probability of opening the L-type Ca^{2+} channels. To simulate a model like this, we have done the following. First, to simulate a cell, we have to integrate the system of ordinary differential equations (ODEs) defining the cell model over a time step ΔT . Within each time step we do a Monte Carlo simulation of the gating of this system over some large number of similar systems that we model in an individual myocyte. It is a large calculation that couples Monte Carlo simulation within an ODE integration. The system behaves beautifully and in accordance with experimental data. When we use this local control model as a way of simulating Ca^{2+} release, we can now obtain stable action potentials. We now have a system that is accurately describing the detailed mechanisms of Ca^{2+} release, and these more global properties of their release, yet it is a very complicated simulation model: one that is not really even practical for simulating single cells (we did this on a parallel machine), let alone a myocardial model. There are issues here about the nature of Ca^{2+} release and uptake, and even Ca^{2+} signalling in general in the myocyte that we can discuss. And I think there are issues about integration between different levels of models. What we would now like to do with this system is to find a way to retain the detailed biophysical information about the subsystems, while using a mathematical approach to describe the average behaviour that would be consistent with the principles of local control of Ca^{2+} release. We need to do this to build models of the cardiac myocyte that accurately describe that release. The reason we want to retain a level of biophysical detail is that we know in heart failure that there are changes in the different β subunit compositions of L-type Ca^{2+} channels that can change their gating kinetics. We believe in heart failure that there may be changes in the microstructure of this diadic space. It is not known for sure, but this hypothesis is out there. There may be changes in the phosphorylation state of the ryanodine receptor. All of these things can be addressed with this kind of model. We need a way to move to the more integrative cell model in an efficient fashion.

Noble: If you were to remove the Ca^{2+} -dependent inactivation of the Ca^{2+} channel from the models, you would predict that you would get a massive prolongation of the action potential. I think this is a beautiful case where modelling is clearly leading the way, because quite a lot of the data on this don't show that. It is an interesting point. If you look at Boyett's work on BAPTA-AM,

it doesn't show it (Janvier et al 1997). Nor does Jamie Vandenberg's work on whole hearts (personal communication); again, there is virtually no change in action potential duration. I think we know the answer: a few years ago Jean-Yves Le Guennec and I infused a massive amount of buffer into the cell through the pipette: 30 mM (Le Guennec & Noble 1994). The action potential was doubled in length. I take it that you would agree that the problem lies in the fact that the experiments, although removing the Ca^{2+} transient enough to remove the contraction, are not actually stopping this process.

Winslow: That's right. It has been terribly difficult to control what is happening in that little compartment. I didn't point out that the difference between these channels is 12 nm, so this is a tiny subspace. Dave Yue has done a truly elegant experiment in which he has expressed a mutant CAM in cultured myocytes and looked at action potential duration.

Noble: I don't think one could have unravelled this without modelling. In fact, given the nature of the experiments that have been done with the Ca^{2+} buffers, I think these would have led one in the wrong direction.

Winslow: What Dave observes in this calmodulin mutant myocyte is a ventricular action potential with a duration of about 3 s, as opposed to the 200–300 ms that is normal for the guinea-pig.

Subramaniam: The reason why we are not able to model the SR release of Ca^{2+} efficiently is that the time constants for these processes are very different. This is why a local model is able to do that in a more accurate manner. This goes back to the definition of 'module'. We need to specify the time constants in defining modules appropriately.

Winslow: Even if we define the module in that way, there are 50 000 of these modules in the myocyte. What we need here are mathematical approaches that will enable us to step from the microscopic stochastic behaviour to macroscopic behaviour. I don't know what those approaches are yet, but we desperately need them for the myocyte.

Hinch: This is something that I have been working on. If you look at the results of an individual functional unit, there are many short scale stochastic events. If you look at the overall result, effectively it is flipping between two states. We are working on ways to reduce this very complicated system into a simplified system that has a long time constant. You can go from having millions of Monte Carlo events to having just a few.

Winslow: That is exactly the kind of thing that is necessary. But I would say that whatever technique is used to simplify the system, this technique needs to incorporate the level of biophysical detail that is in the detailed functional unit model. This is so you can change something in this model, such as the properties of the L-type Ca^{2+} channel, reconstruct this simplification and test its consequences on integration.

Hinzb: The simplification is a mathematical derivation. We end up with different transition coefficients but they are functions of what happened before. It is possible.

Shimizu: I am quite interested in this; it is exactly the type of thing that we deal with in bacterial chemotaxis. We do stochastic modelling of localized membrane receptors. You say that you can reduce the model and retain the function. Surely you must lose some information?

Hinzb: The information lost is about what happens at the submillisecond level. The interesting thing that happens is that it switches on at some point and then stays on for a couple of hundred milliseconds. What is happening at the sub-millisecond level is not interesting when you are studying processes at the 100 ms timescale. What you want to know is does it last for 100 ms or 200 ms?

Shimizu: That is fine if you know exactly which features you need to retain to obtain the correct outcome. Obviously in most cases, it is not feasible to have a full stochastic model running in parallel with an ODE system in real time. But what this sort of combined modelling allows people to do is to highlight which experiments need to be done to identify the essential events that occur at the individual molecule level. Many new experimental techniques are becoming available for this type of analysis. Once you have characterized a system at the stochastic individual molecule level, then you can go on to reduce the problem.

Berridge: Many of the things I was going to say have been covered by Raimond Winslow and others. I am not a modeller, but it does seem from hearing what people have been saying that we really need to integrate information between different molecular, structural and physiological elements. While attention has focused on the molecular and structural aspects, the physiological tool kit is also something we need to concentrate on. In fact, the answer to the problem of trying to describe the paradox of graded responses in cardiac cells emerged from a study of the physiological tool kit: breaking the Ca^{2+} signal down into its elementary events led to the realization that individual sparks were associated with the individual SR regions that functioned as autonomous units. These functioned as all-or-none units, and depending on how many are recruited there is a graded response. Such studies on elementary events have been extended to a study of arrhythmias in atrial cells. The structural tool kit is particularly interesting in this cell with regard to the distribution of two key intracellular channels: the ryanodine receptors, and the inositol-1,4,5-trisphosphate (InsP_3) receptors that are particularly strongly expressed in the atrial cell. Staining with anti-ryanodine receptor antibodies lights up striations, which are the individual SR units and these are the modules described earlier. On the other hand, the type 2 InsP_3 receptor is all in the periphery: there is no trace of it on the internal SR. This has led to the idea that there might be a completely novel form of EC coupling in these cardiac cells. The conventional coupling mechanism involves depolarization to activate

the L-type channels to produce the Ca^{2+} sparklet that then fires ryanodine receptors to produce a Ca^{2+} spark. This spark is then amplified by a process of Ca^{2+} -induced Ca^{2+} release and this causes a globalization of the signal. Since endothelin, which is associated with cardiac hypertrophy and heart failure, is known to generate InsP_3 , it is possible that InsP_3 can activate its receptors in the junctional zone to produce trigger Ca^{2+} , which is then able to activate the same kind of amplification units that are used conventionally. Under this condition, therefore, InsP_3 will be acting when there is no depolarization, and essentially will set up an arrhythmia. This example emphasizes the importance of developing a holistic view when trying to understand the Ca^{2+} signalling system.

Noble: I am not a smooth muscle expert, but I believe it's the case that Ca^{2+} oscillations in some forms of smooth muscle also implicate InsP_3 .

Berridge: In the interstitial cells of Cajal, which drive the rhythm in smooth muscle, there seems to be a pacemaker system very similar to the one that has been described for the sinoatrial node in the heart, in that there is an interplay between the intercellular stores and the plasma membrane. Activation of the InsP_3 receptor plays a role in setting up the instability during the pacemaker phase.

Hunter: It occurs to me that there's another sense in which the need for integration is illustrated by the example you have given. The electrotonic coupling between the atrial cell and its adjacent cell will have a big influence on whether this local arrhythmia is able to propagate.

Berridge: Yes, the individual cells must be considered as part of a connected network. What one would imagine is that this kind of spontaneous activity would be distributed throughout the atrial system. If these events coincide in a local area, then the individual effects would sum to drive the depolarization sufficiently to trigger an extra beat.

McCulloch: This phenomenon has been seen in multicellular ventricular preparations by ter Keurs and colleagues. They propagate at about $200 \mu\text{s}^{-1}$.

Berridge: I think the atrial waves are a little slower. It all depends on the sensitivity of the regenerative components. The closer they are the faster the wave goes.

Paterson: Whenever I come to these sorts of meetings I am always impressed by the quality of the modelling, the research that is going into this, and the ability to collect real time data for these kinds of phenomena. A lot of these issues are very unique to this domain. There is a lot of modelling that is perhaps less advanced or has less of a history in other fields such as metabolism and immunology. Some of the broad conclusions in terms of what the key problems and solutions are can be very different. Everything we are talking about here is valid, but it is somewhat coloured by the fact that there is a particular class of problems being worked on by most of the people in this room.

Berridge: I'm not sure that's right: although we work on cardiac cells, we are very interested in T cell activation and how the Ca^{2+} signal is presented there.

Paterson: I didn't say it was irrelevant, but there are certainly issues that I'm aware of in modelling aspects of the immune system that aren't on anyone's radar screens here.

References

- Jafri MS, Rice JJ, Winslow RL 1998 Cardiac Ca^{2+} dynamics: the roles of ryanodine receptor adaptation and sarcoplasmic reticulum load. *Biophys J* 74:1149–1168
- Janvier NC, Harrison SM, Boyett MR 1997 The role of inward Na^{+} - Ca^{2+} exchange current in the ferret ventricular action potential. *J Physiol* 498:611–625
- Le Guennec JV, Noble D 1994 Effects of rapid changes of external Na^{+} concentration at different moments during the action potential in guinea-pig myocytes. *J Physiol* 478:493–504
- Linz KW, Meyer R 1998 Control of L-type calcium current during the action potential of guinea-pig ventricular myocytes. *J Physiol* 513:425–442

The Virtual Cell project

Leslie M. Loew

*Center for Biomedical Imaging Technology, Department of Physiology,
University of Connecticut Health Center, Farmington, CT 06030, USA*

Abstract. The Virtual Cell is a modular computational framework that permits construction of models, application of numerical solvers to perform simulations, and analysis of simulation results. A key feature of the Virtual Cell is that it permits the incorporation of realistic experimental geometries within full 3D spatial models. An intuitive JAVA interface allows access via a web browser and includes options for database access, geometry definition (including directly from microscope images), specification of compartment topology, species definition and assignment, chemical reaction input and computational mesh. The system is designed for cell biologists to aid both the interpretation and the planning of experiments. It also contains sophisticated modelling tools that are appropriate for the needs of mathematical biologists. Thus, communication between these traditionally separate scientific communities can be facilitated. This paper will describe the status of the project and will survey several applications to cell biological problems.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 151–161

The accelerating progress in cataloguing the critical molecular and structural elements responsible for cell function has led to the hope that cell biological processes can be analysed and understood in terms of the interactions of these components. One prerequisite for such analyses is the acquisition and organization of quantitative data on these interactions. These would include biochemical reaction rates, electrophysiological data on membrane transport dynamics, diffusion of cellular species within cellular compartments, and the mechanical properties of cellular structures. But a second prerequisite is the effective synthesis of these often heterogeneous data by constructing models that can then predict the overall behaviour of the biological system. If the model correctly predicts the biological endpoint, one can hypothesize that the elements within the model are sufficient; furthermore, it is often possible to discern which of these elements are the most critical. This can then be tested by further experiments designed to specifically perturb or remove these elements (e.g. gene knockouts). Perhaps more useful, however, is when the model is unable to predict the observed biology. This requires that the elements of the model are either incorrect or incomplete. Analysis

of such faulty models can directly motivate the discovery, via new experiments, of previously unknown critical biochemical or structural features required for the cellular process under investigation.

Despite these clear benefits of the use of modelling as an adjunct to experiment, the difficulties associated with the formulation of mathematical models and the generation of simulations from them has impeded the adoption of this disciplined and quantitative approach to research in cell biology. Because biologists rarely have sufficient training in the mathematics and physics required to build quantitative models, modelling has been largely the purview of theoreticians who have the appropriate training but little experience in the laboratory. This disconnection to the laboratory has limited the impact of mathematical modelling in cell biology and, in some quarters, has even given modelling a poor reputation. The Virtual Cell project aims to address this problem by providing a computational modelling framework that is accessible to cell biologists. It does this by abstracting and automating the mathematical and physical operations involved in constructing models and generating simulations from them. At the same time, the Virtual Cell provides a mathematical interface that allows theoreticians to examine and elaborate models through purely mathematical formulations. This dual interface has the additional benefit of encouraging communication and collaboration between the experimental and modelling communities. This paper will describe the current implementation of the Virtual Cell and briefly review some of the cell biological problems to which it has been applied. The reader is referred to other recent reviews for broader coverage of the field of computational cell biology (Loew & Schaff 2001, Slepchenko et al 2002) and to our website (<http://www.nrcam.uchc.edu>) for a user guide and tutorial.

The problem domain: reaction/diffusion in arbitrary geometries

At its most fundamental level, a cell biological process can be described as the consequence of a complex series of chemical transformations. To understand the process, the relevant molecules have to be identified and their time-varying concentrations and spatial distributions have to be determined. A model, at this molecular level, chooses all the presumed chemical species, assigns them initial concentrations and spatial distributions and connects them with appropriate kinetic expressions. A simulation that predicts the spatiotemporal behaviour of this system has to solve a class of problems known as reaction/diffusion equations. The mathematical problem is summarized by the equations:

$$F_i = -D_i \nabla C_i - z_j \mu_j C_i \nabla \Phi, \quad \mu_i = \frac{D_i F}{RT} \quad (1)$$

$$k + j \longleftrightarrow i \quad R_i = \frac{d[i]}{dt} = k_1[k][j] - k_{-1}[i] \quad (2)$$

$$\frac{\partial C_i}{\partial t} = -\text{div}F_i + R_i \quad (3)$$

The first line is the familiar Nernst–Planck equation that describes the flux, F_i , of a molecule i , driven by its concentration gradient, ∇C_i , and, if it has an ionic charge z_i , the electric field in the system $\nabla\Phi$. The diffusion coefficient, D_i , and the mobility μ_i , are the proportionality constants for these driving forces. The second line portrays a typical reaction that produces molecule i (while consuming j and k). The mass action ordinary differential equation (ODE) for the rate of change of i , R_i , depends on the concentrations of the reactants and products. In general, R_i can depend on the concentrations of any of the molecules in the system and may have a more complex form than the mass action expression shown here. The third line combines the fluxes and reactions into a system of partial differential equations (PDEs) that must be integrated to simulate the behaviour of the molecular species.

The fact that the Virtual Cell is designed to handle any reaction system in any geometry, precludes the formulation of a general analytical solution for the problem. There are two generic approaches to numerical solutions—stochastic and continuous. The continuous approach provides a deterministic description in terms of average species concentration. This approach is effective and accurate so long as the number of molecules in a system is large, such that thermal stochastic fluctuations around average values can be ignored. We have found that the finite volume method (Patankar 1980) for discretization of a system of PDEs is especially well suited for our problem domain—that is reaction/diffusion equations in arbitrary geometries (Schaff et al 1997, 2001, Choi et al 1999). Of course, the software can also solve non-spatial problems corresponding to systems of ODEs describing reactions within well stirred compartments and fluxes across the membranes that separate the compartments. The software provides a choice of several solvers for such compartmental problems including a stiff solver. For both spatial and compartmental problems, we have implemented an automated pseudo-steady approximation that can be invoked by the user when a subsystem of reactions equilibrates rapidly on the timescale of the overall process of interest (Slepchenko et al 2000). The currently available user interface for the Virtual Cell includes full access to these capabilities for numerical solutions of continuous reaction/diffusion equations.

Stochastic fluctuations can become important if the number of molecules involved in a process is relatively small. For fully stochastic problems in which

the number of particles in a reaction/diffusion system is too small to solve with numerical solutions of PDEs, diffusion can be described as Brownian random walks of individual particles and chemical kinetics is simulated as stochastic reaction events. We also need to consider hybrid systems of stochastic differential equations where one can combine the numerical techniques commonly applied to regular differential equations and Monte Carlo methods employing random number generators. In the Virtual Cell, we employ an efficient algorithm in which the probabilities of each reaction are calculated from rate constants and numbers of substrate molecules (Gillespie 1977, 2001). A stochastic method is used to determine which reaction will occur based on their relative probabilities. The time step is then adjusted to match the particular reaction that occurs. After the reaction is complete the numbers of substrate molecules are readjusted prior to the next cycle. When combined with stable accurate numerical schemes developed for the conventional differential equations, they can be applied for numerical solution of stochastic differential equations with discrete random processes. Although this approach has been implemented in our C++ library and has been applied to problems on the dynamics of RNA granule trafficking (Carson et al 2001; <http://www.nrcam.uchc.edu>), the stochastic modelling capabilities of the Virtual Cell are not accessible through the current Java user interface.

The modelling process in the Virtual Cell environment

The Virtual Cell system uses a distributed client-server architecture that permits access over the Internet. The Java client runs through a web browser and is thus compatible with all the common operating systems (Windows, MacOS X and Linux). A numerics server, currently consisting of a cluster of eight dual-processor Alpha nodes, assures the availability of sufficient computational power to the user. The system also includes a database server that maintains user information and ensures the security and integrity of models and simulation results. Through the database structure, users also have the option of ‘sharing’ models with a selected group of collaborators or ‘publishing’ completed models so that they can be accessed by the entire scientific community. Models can be copied and reused or modified through the database as well. In addition to the above benefits, the architecture has the important additional advantage of permitting centralized maintenance and the ready deployment of enhancements.

The modelling process within the Virtual Cell is based on a hierarchical organization that emphasizes reusability. As depicted in Fig. 1, the parent object in a model is a general cell physiological description of the system that we designate the BioModel. The BioModel specifies: the compartmental topology of the system; the identities of molecular species; the compartmental or membrane locations of the species (membranes are automatically defined as the boundaries

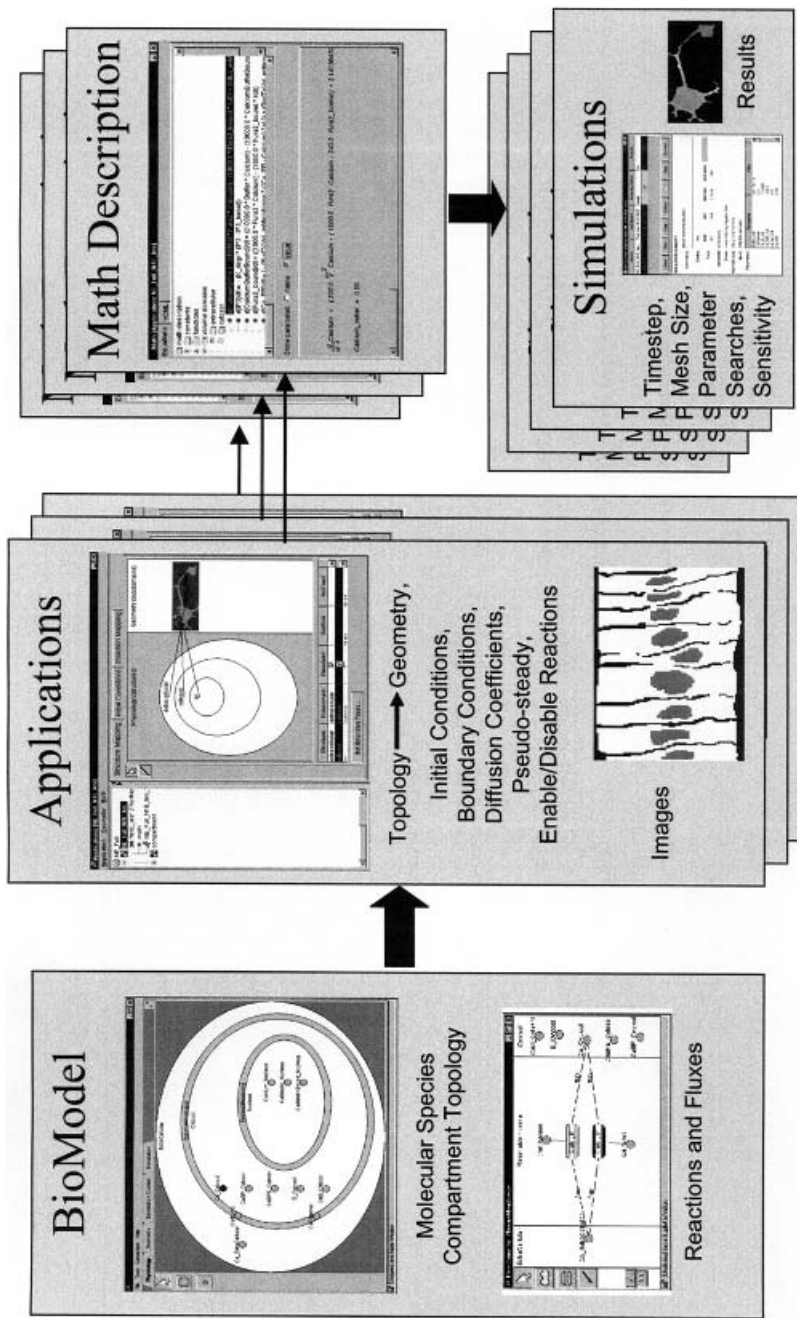
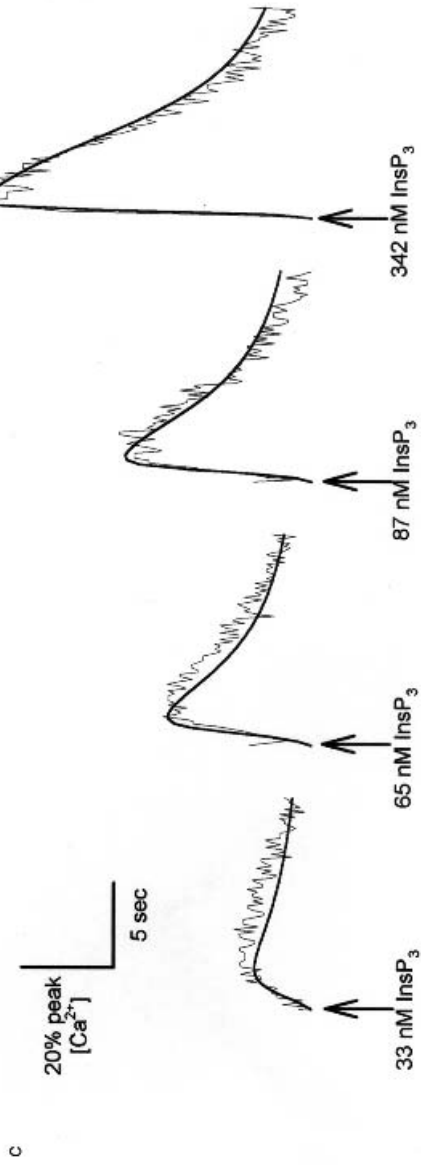
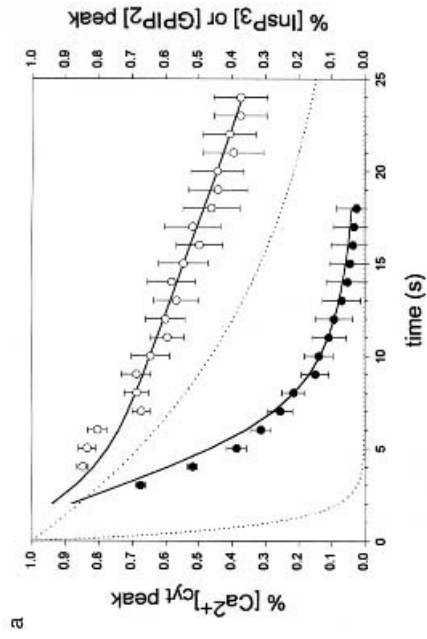
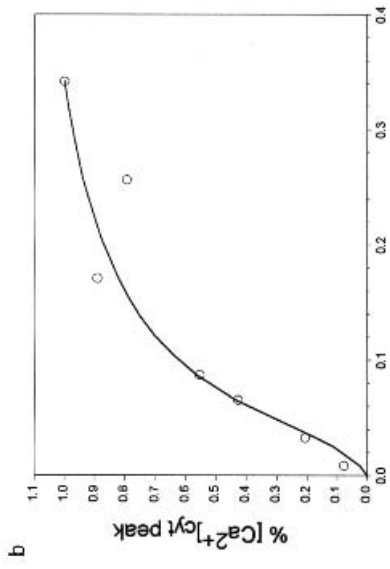


FIG. 1. The hierarchical organization of a Virtual Cell BioModel. Each major component of a model is shown within a rectangle that includes a screenshot of part of the user interface and the model features that are specified. The broad arrows designate a one-to-many relationship and the narrow arrows a one-to-one relationship.



separating compartments); and reactions and membrane transport kinetics. A BioModel can then spawn several ‘Applications’ that each specify a geometry, boundary conditions, default initial concentrations and parameter values, and whether any of the reactions are sufficiently fast to permit a pseudo-steady state approximation. The geometry can be from zero- (i.e. a compartmental model) to three-dimensional and can be derived either by importing a segmented experimental image (e.g. confocal micrographs) or by specifying an analytical geometry. For compartmental models or for compartments that are unresolved within the geometry, volume fractions relative to the parent compartment and surface to volume ratios must be specified. Also at the Application level, individual reactions can be disabled as an aid in determining the proper initial conditions for a pre-stimulus stable state.

An Application together with its parent BioModel is sufficient to completely define the governing mathematics of the model and, accordingly, each application generates its own unique math description expressed in VCMDL (Virtual Cell Math Description Language). VCMDL is a fully declarative language that can be edited independently of the BioModel in a separate MathModel workspace. This can be used to refine models in ways that are more flexible than permitted by the BioModel interface. Indeed, a VCMDL formulation of a model may be created from scratch within the MathModel workspace. The dual BioModel and MathModel interfaces were developed to permit the maximum flexibility in developing a model, but also serve to facilitate interaction between biologists and theoreticians.

The last part of Fig. 1 illustrates the relationship of Applications and MathModels to Simulations. The implementation of a simulation is kept separate from the model specifications and several simulations can be spawned off of a given Application/MathModel. The simulation specifications include the choice of

FIG. 2. Fit of model to experiment for InsP_3 -induced Ca^{2+} dynamics in a smooth muscle cell line. (a) The time course of Ca^{2+} levels following uncaging of either InsP_3 (closed circles) or GPIP_2 (open circles); each point represents the average of 10 experiments, each of which is normalized internally to 1.0 for the peak Ca^{2+} concentration. The fitted lines are the calculated values for the time-course of $[\text{Ca}^{2+}]_{\text{cvt}}$ based on the simulations for InsP_3 and GPIP_2 stimulation (of the same concentrations measured in the experiments). The rate for metabolite degradation was determined for the two conditions to optimize the fit to each set of averaged experimental points; the resultant time constants were 0.8 s for InsP_3 and 13 s for GPIP_2 . For comparison, degradation curves for InsP_3 are also included as dotted curves. (b) The model can also be used to simulate a dose response series for the Ca^{2+} response to varying levels of uncaged InsP_3 in a single cell. The circles are experimental data for a titration of InsP_3 in a single cell; the simulation results are shown as a solid line. (c) Using the same parameters as for the dose-response in (b), we simulated the full time-course for four concentrations of uncaged InsP_3 . Experimental data are light curves and simulation results are shown as heavy curves. (Taken from Fink et al 1999a, with permission of the Biophysical Journal.)

solver, time step, mesh size for spatial simulations, and overrides of the default initial conditions or parameter values. Local sensitivity analysis can be performed within a Simulation to probe for which features of the model are most critical in determining its overall behaviour and also to aid in parameter estimation. Simulation results are displayed as images of the variable values coded in greyscale or pseudocolour and mapped to the simulation geometry. Timeplots at multiple coordinates or intensities along a line or curve within the geometry at a selected time can also be displayed. In addition, results of simulations can be exported in multiple formats, including images, movies and lists of variable values suitable for spreadsheet analysis.

Examples of studies using the Virtual Cell

Our laboratory has applied the Virtual Cell to the analysis of Ca^{2+} dynamics in several cell types. The first paper to appear was a study of inositol-1,4,5-trisphosphate (InsP_3)-induced release of Ca^{2+} from the endoplasmic reticulum (ER) of a smooth muscle cell line (Fink et al 1999a). In that work we used the Virtual Cell to develop a model for the calcium dynamics following uncaging of InsP_3 and a non-hydrolysable analogue, GPIP_2 . The results summarized in Fig. 2 show that the model was able to reproduce both the time-course and dose dependence of the experimentally observed Ca^{2+} release event. The model demonstrated that the behaviour of the system was critically dependent on the degradation of InsP_3 —i.e. that the Ca^{2+} release channel did not significantly inactivate on the timescale of the observed Ca^{2+} dynamics.

This study was followed with a much more extensive investigation of Ca^{2+} release in differentiated N1E-115 neuroblastoma cells (Fink et al 1999b, 2000). This study showed that the neuronal morphology of these cells controlled the spatiotemporal pattern of Ca^{2+} signals following stimulation by bradykinin, a neuromodulator. The modelling activity led us to discover the uneven distribution of ER Ca^{2+} stores within these cells and discern how the interplay of cell shape and receptor distribution assured a Ca^{2+} wave with a uniform amplitude.

In the laboratory of my colleague John Carson, the Virtual Cell has been used to understand the mechanism of RNA granule trafficking (Carson et al 2001). These models require the stochastic simulation capabilities of the C++ library because they attempt to elucidate the behaviour of single granules as they are driven along microtubules by molecular motors. The behaviour of the granules in the Virtual Cell model can be directly compared to the motions of fluorescently labelled RNA granules as visualized through a confocal microscope. A model that includes two opposing motors each with three states corresponding to whether they are unbound to the microtubule track, bound but inactive, or bound and exerting

force is sufficient to describe the behaviour if elastic forces within the granule are also included.

Several other important examples of Virtual Cell applications represent a spectrum ranging from the testing of simple but analytically intractable hypotheses, to the elaboration of complex reaction schemes for well-regulated intracellular processes. Representative of the latter is a study of nucleocytoplasmic transport mediated by the RanGTPase system (Smith et al 2002); in this study the ability of the model to visualize the separate components of the system gave credence to the assertion that the nuclear pore complex did not play an important regulatory role. Also complex is a model that is being developed to understand the influence of mitochondrial morphology as a potential regulator of respiratory efficiency (Mannella et al 2001); this study develops a model for a single mitochondrion based on 3D electron tomography data. A spatially larger system is represented by a model of transepithelial Ca^{2+} transport (Slepchenko & Bronner 2001) that points to the coexistence of two transport systems in the apical membrane. At the level of simpler hypothesis testing, is a study that used the Virtual Cell to demonstrate that nuclear envelope breakdown during mitosis proceeds via an initial breach in the nuclear membrane that progressively widens (Terasaki et al 2001). Finally, the focal photorelease of caged thymosin β , an actin-sequestering molecule, was modelled in order to determine the localization of this perturbation to the cytoskeleton given the diffusion of released molecules from the site of irradiation and their rate of reaction with the pool of g-actin (Roy et al 2001). Thus in the short period that the Virtual Cell has been available, it has been proven useful in quite a variety of cell biological investigations.

Acknowledgements

The author thanks his colleagues James Schaff and Boris Slepchenko who have led the development of the Virtual Cell over the last 6 years. Yung-sze Choi, Ann Cowan, Susan Krueger, Frank Morgan, Ion Moraru, Charles Fink, John Wagner, James Watras and Daniel Lucio are also acknowledged for their many contributions to this work. The NIH National Center for Research Resources has supported this work through grant RR13186.

References

- Carson JH, Cui H, Krueger W, Slepchenko B, Brumwell C, Barbarese E 2001 RNA trafficking in oligodendrocytes. In: Richter D (ed) Cell polarity and subcellular RNA localization. Springer-Verlag, Berlin, p 69–83
- Choi YS, Resasco D, Schaff J, Slepchenko B 1999 Electro-diffusion of ions inside living cells. IMA J Math Appl Med Biol 62:207–226
- Fink CC, Slepchenko B, Loew LM 1999a Determination of time-dependent inositol-1,4,5-trisphosphate concentrations during calcium release in a smooth muscle cell. Biophys J 77:617–628

- Fink CC, Slepchenko B, Moraru II, Schaff J, Watras J, Loew LM 1999b Morphological control of inositol-1,4,5-trisphosphate-dependent signals. *J Cell Biol* 147:929–935
- Fink CC, Slepchenko B, Moraru II, Watras J, Schaff J, Loew LM 2000 An image-based model of calcium waves in differentiated neuroblastoma cells. *Biophys J* 79:163–183
- Gillespie DT 1977 Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81:2340–2361
- Gillespie DT 2001 Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys* 115:1715–1733
- Loew LM, Schaff JC 2001 The Virtual Cell: a software environment for computational cell biology. *Trends Biotechnol* 19:401–406
- Mannella CA, Pfeiffer DR, Bradshaw PC et al 2001 Topology of the mitochondrial inner membrane: dynamics and bioenergetic implications. *IUBMB Life* 52:93–100
- Patankar SV 1980 Numerical heat transfer and fluid flow. Taylor & Francis, London
- Roy P, Rajfur Z, Jones D, Marriott G, Loew LM, Jacobson K 2001 Local photorelease of caged thymosin β 4 in locomoting keratocytes causes cell turning. *J Cell Biol* 153:1035–1048
- Schaff J, Fink CC, Slepchenko B, Carson JH, Loew LM 1997 A general computational framework for modeling cellular structure and function. *Biophys J* 73:1135–1146
- Schaff JC, Slepchenko BM, Choi Y, Wagner JM, Resasco D, Loew LM 2001 Analysis of non-linear dynamics on arbitrary geometries with the Virtual Cell. *Chaos* 11:115–131
- Slepchenko BM, Bronner F 2001 Modeling of transcellular Ca transport in rat duodenum points to the coexistence of two mechanisms of apical entry. *Am J Physiol* 281:C270–C281
- Slepchenko BM, Schaff JC, Choi YS 2000 Numerical approach to fast reaction-diffusion systems: application to buffered calcium waves in bistable models. *J Comp Phys* 162:186–218
- Slepchenko B, Schaff JC, Carson JH, Loew LM 2002 Computational cell biology: spatiotemporal simulation of cellular events. *Annu Rev Biophys Biomol Struct* 31:423–441
- Smith AE, Slepchenko BM, Schaff JC, Loew LM, Macara IG 2002 Systems analysis of Ran transport. *Science* 295:488–491
- Terasaki M, Campagnola P, Rolls MM et al 2001 A new model for nuclear envelope breakdown. *Mol Biol Cell* 12:503–510

DISCUSSION

Berridge: I would like to raise an issue about the work that you did on the cerebellar Purkinje neuron. I am sure you are aware that input-specific modification of the spine depends not on multiple inputs from the parallel fibre but on coincidence between the primary fibre and the parallel fibre. In 1993 I proposed that the InsP_3 receptor was the coincidence detector (Berridge 1993).

Loew: That has been shown experimentally.

Berridge: Have you modelled this example of coincidence detection? Although it is very interesting that the spine can restrict InsP_3 diffusion during repetitive stimuli, the reality is that you only need one pulse to obtain dramatic changes as long as it is connected with another one, as occurs during coincidence detection.

Loew: We have modelled that. The fact is, you can get long-term depression (LTD) with multiple InsP_3 stimulation, and that is what we modelled here. It can also be done the way you have suggested, with one InsP_3 stimulation plus a stimulation from the climbing fibre which activates voltage-dependent channels. We

have the channels in there as well. The tremendous non-linear sensitivity of that system translates to a sensitivity to Ca^{2+} . As you know, the InsP_3 receptor is also activated by Ca^{2+} , and if you put a little bit of Ca^{2+} in there at any point, you can automatically produce a big Ca^{2+} spike. We have modelled this.

Noble: I thought that you highlighted a very important role of modelling in pointing out that you could reveal what the model is saying the InsP_3 levels are doing. This is a feature that can be addressed by modelling in all kinds of different contexts. In addition to revealing parameters that we don't yet have an indicator for (but hopefully one day we will have), we can also do 'gene knockouts' that at the moment aren't possible, pulling components out and putting them back in. This is something that Bernhard Palsson has demonstrated in his impressive metabolic modelling (Edwards et al 2001) and we have also done in relation to some of the work on cardiac modelling. These are aspects of modelling that we need to bring out as one of the great strengths.

Loew: This is sort of equivalent to the idea of lowering the InsP_3 receptor density in the Purkinje cell, or in some way changing its characteristics.

References

- Berridge MJ 1993 Cell signalling. A tale of two messengers. *Nature* 365:388–389
Edwards JS, Ibarra RU, Palsson BO 2001 *In silico* predictions of *Escherichia coli* metabolic capabilities are consistent with experimental data. *Nat Biotechnol* 19:125–130

Modelling the bacterial chemotaxis receptor complex

Thomas Simon Shimizu and Dennis Bray

Department of Zoology, University of Cambridge, Downing Street, Cambridge CB2 3EJ, UK

Abstract. The pathway controlling chemotaxis in *Escherichia coli* is the simplest and most well understood cell signalling system to date. However, quantitative models based on the available data still fail to reproduce important features of the pathway. Most notably, the observed sensitivity of cells to very small changes in stimulus concentrations cannot be reproduced by conventional models based on the measured concentrations, binding affinities and rate constants of the proteins involved. This discrepancy, together with recent experimental findings, drew our attention to the spatial organization of molecules within the cell and in particular to the clusters of receptors localised at the cell poles. A stochastic simulator for chemical reactions, STOCHSIM, was previously developed to model the chemotaxis pathway at the level of individual molecular interactions. This program has now been extended to incorporate a spatial representation that allows the interaction between molecules in a two-dimensional lattice to be simulated. *In silico* 'experiments' using this new version of STOCHSIM demonstrate that lateral interactions between clustered receptors can significantly enhance the excitation response. The adaptation reactions may also exploit the proximity of receptor molecules, and a hypothetical mechanism by which this may occur is currently being tested.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 162–181

The *Escherichia coli* chemotaxis system presents a unique opportunity to identify the principles and to develop the methods required for studying cell signalling *in silico*. It has been the subject of intensive investigation for over three decades as a model cell sensory and signalling system, and an extensive body of literature has developed as a result (for recent reviews, see Bren & Eisenbach 2000, Falke et al 1997). All of the enzymes in the pathway have been characterized kinetically, and a large collection of mutant strains are available for quantitative physiological analysis. Atomic resolution structures have also been determined for nearly all of the involved proteins in recent years, and this has opened the door to a detailed molecular explanation of the mechanisms that account for the observed kinetics. The structure of the pathway is simple, consisting of the chemotactic receptors

and only six cytoplasmic proteins (see Fig. 1A and Table 1), but it shares many features in common with more complicated pathways of eukaryotes, including phosphorylation cascades, covalent modification, multiprotein complexes and clustered receptors. A small number of protein species combine to generate

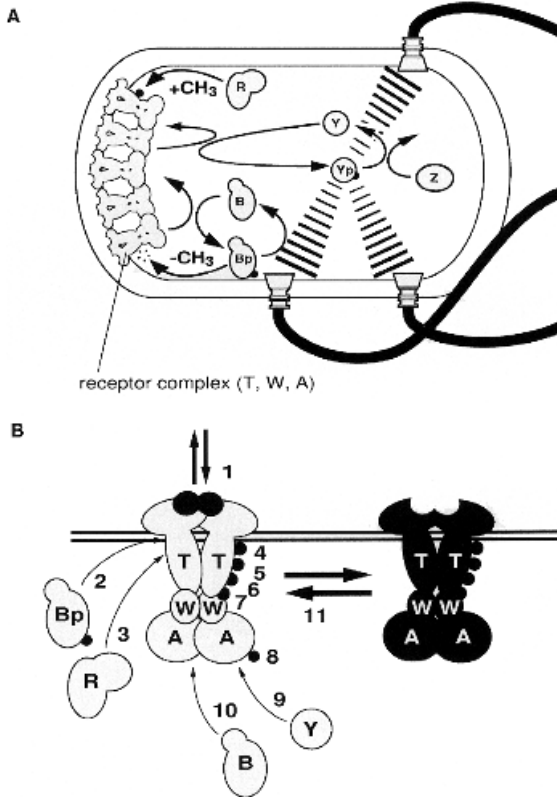


FIG. 1. The bacterial chemotaxis signalling pathway. (A) Overview of the pathway. Chemotactic receptors (T) are clustered primarily at the cell poles, and form stable ternary complexes with the histidine kinase CheA (A) and the linking protein CheW (W). Ligand binding to the receptors influences the rate of phosphotransfer from CheA to the response regulator CheY (Y), the phosphorylated form of which (Yp) interacts with the flagellar motor to control swimming. The steady-state level of this signal is regulated by the antagonistic effects of two- adaptation enzymes, CheR (R) and CheB (B). The reversible phosphorylation of CheB provides negative feedback in the pathway, and CheZ accelerates the dephosphorylation of CheY. See Table 1 for a description of each component. (B) The Tar receptor complex as modelled in STROCHSIM. The state of each receptor complex is represented by eleven binary flags. Ten of these represent the state of binding or modification sites: aspartate binding (1); CheBp binding (2); CheR binding (3); methylation (4–7); phosphorylation (8); CheY binding (9); and CheB (10) binding. Each receptor complex is assumed to be in rapid equilibrium between two conformational states, active (white) and inactive (black), represented by the final flag (11).

TABLE 1 Components of the bacterial chemotaxis pathway

<i>Component (copies per cell)</i>	<i>Description</i>
Receptors (~4000 dimers)	Transmembrane transducers also known as methyl accepting chemotaxis proteins (MCPs). They monitor various attractant and repellent concentrations, as well as temperature and pH. <i>E. coli</i> possesses five MCP species named after the attractants they bind: Tar (aspartate), Tsr (serine), Trg (ribose and galactose), Tap (dipeptides) and Aer (oxygen).
CheW (~8000 monomers)	Scaffolding protein that couples the chemotactic receptors to CheA. It has been shown that CheW is required for polar receptor cluster formation (Maddock & Shapiro 1993).
CheA (~4000 dimers)	Histidine kinase that donates phosphoryl group to CheY and CheB. Its activity is regulated by the chemotactic receptors. Attractant stimuli inhibit CheA activity and repellent stimuli enhance it.
CheY (~17000 monomers)	Response regulator that relays signal from receptor complex to flagellar motors. The phosphorylated form of CheY (CheYp) interacts directly with the switch complex of the flagellar motor to promote CW rotation.
CheR (~200 monomers)	Methyltransferase that adds methyl groups to specific glutamyl residues on the cytoplasmic domain of MCPs. Each added methyl group increases the activity of CheA in complex with the receptor, thereby counteracting the effect of attractant binding.
CheB (~1700 monomers)	Methylesterase/deamidase that counteracts the effect of CheR by removing the added methyl groups and lowering the activity of CheA. Because this activity is strongly enhanced in the phosphorylated form of CheB (CheBp), which in turn is regulated by CheA, it serves as a negative feedback in the pathway.
CheZ (~12000 dimers)	Accelerates the dephosphorylation of CheYp, thereby dramatically increasing the speed at which <i>E. coli</i> cells can respond to stimuli. Only enteric bacteria possess a <i>CheZ</i> gene.
Flagellar motor (~6)	Large protein complex comprising over 100 subunits. In the absence of CheYp, it rotates exclusively counter-clockwise (CCW), causing the cell to swim forward in a straight line (run). The probability of clockwise (CW) rotation, which causes a swimming cell to change direction (tumble), increases with the intracellular concentration of CheYp.

surprisingly sophisticated behaviour including signal detection, integration, amplification and adaptation. The near-completeness of molecular information on this pathway makes it an ideal prototype system for the simulation of cell signalling pathways in general.

A standard method for simulating biochemical pathways is to represent each reaction by a continuous, deterministic rate equation and to numerically integrate the resulting set of equations to obtain the changes in species concentrations over time. This approach has been applied to the chemotaxis pathway for nearly a decade with considerable success. One of the first of such efforts, a computer program named BCT, was developed in an attempt to incorporate the available biochemical data in a coherent simulation. Initially, BCT consisted of 10 ordinary differential equations describing the excitation response, and a simplified model of the flagellar motor to produce a behavioural output (Bray et al 1993). It has since been extended to include the binding reactions leading to the formation of the receptor complex (Bray & Bourret 1995), and a simplified adaptation response (Levin et al 1998). It now consists of 75 differential equations capable of reproducing the chemotactic phenotypes of over 60 mutants, and is actively maintained as a reference model (available for download at <http://www.zoo.cam.ac.uk/comp-cell>). In another application of deterministic equation-based modelling, Barkai & Leibler (1997) have proposed and simulated a mechanism that ensures the robustness of exact adaptation to perturbations in biochemical parameters, and this property was demonstrated later by experiment (Alon et al 1999).

Certain quantitative features of the chemotactic response, however, have proven difficult to reproduce. Cells of *E. coli* display remarkable sensitivity to very small changes in stimulus over a wide range of background concentrations (Mesibov et al 1973, Berg & Tedesco 1975, Segall et al 1986). This combination of high sensitivity and wide dynamic range is not reproduced by BCT or any other simulation based on the measured protein concentrations and rate constants. One possible explanation for this discrepancy was that these models do not fully account for the large number of states that the receptor complex can occupy. For example in BCT, the aspartate receptor (Tar) is modelled with only one methylation site whereas in reality there are four. The Barkai and Leibler model, which does include multiple methylation, ignores the downstream phosphorylation cascade. The full complement of receptor states should include, in addition, the binding of ligand in the periplasm and of the modification enzymes, as well as the activity of the receptor. The deterministic equation-based approach breaks down when one tries to incorporate all of these states as separate molecular species. This is due to the combinatorial explosion in the number of equations that need to be integrated — as additional bindings or modifications are considered, the number of reactions which need to be explicitly represented as rate equations grows exponentially.

To overcome this difficulty, and to study the random fluctuations that may influence the pathway, a novel stochastic simulation program named STOCHSIM was developed by Carl Firth (Morton-Firth 1998) (available for download at <http://www.zoo.cam.ac.uk/comp-cell/StochSim.html>). In STOCHSIM, every molecule

in the reaction system is represented as an individual software object, and a unique algorithm tests a pair of molecules for reaction in every simulation iteration. Because every copy of each molecular species is stored as a software ‘object’ in a separate location of memory, the internal state of each molecule can be encapsulated within a molecule object. This removes the need to represent each state of a protein complex as a separate molecular species, and greatly increases the efficiency of simulation when the number of internal states is large. In addition, because the interaction between discrete particles are computed using reaction probabilities, STOCHSIM is capable of reproducing realistic fluctuations in the concentration of molecules which can be significant when the number of particles of one or more reactant species are very small. This feature has been exploited in a study of the temporal fluctuations in the concentration of the active response regulator CheYp of the chemotaxis pathway (Morton-Firth & Bray 1998).

The STOCHSIM model of chemotaxis

STOCHSIM individual-based algorithm has allowed us to develop a detailed simulation of the chemotaxis pathway in which the Tar receptor complex is modelled with the full complement of known bindings, modifications and conformational states (Fig. 1B). The key assumption of the model is that the signalling output of the receptor complex is determined by a rapid, thermally driven equilibrium between two conformational states, active and inactive. The probability that a receptor complex is in the active state at any instant in time (p) can then be obtained by assigning a free energy difference (ΔG) between the active and inactive states, and using the thermodynamic relationship

$$\Delta G = -RT \ln [p/(1-p)] \quad (1)$$

where R is the gas constant and T is the absolute temperature.

The inputs that modulate this equilibrium are the binding of stimulus ligand in the periplasm, and the methylation state of the receptor dimer. We express their contributions to ΔG by assigning specific energy values based on experimental observations to each ligand binding and methylation event (E_L and E_M , respectively), and assume that their effects are additive so that

$$\Delta G = E_L + E_M \quad (2)$$

This results in a unique value of ΔG for every permutation of ligand-binding and methylation state, some examples of which are depicted in Fig. 2A. Solving Equations 1 and 2 for each of these combinations yields the required set of activation probabilities (p).

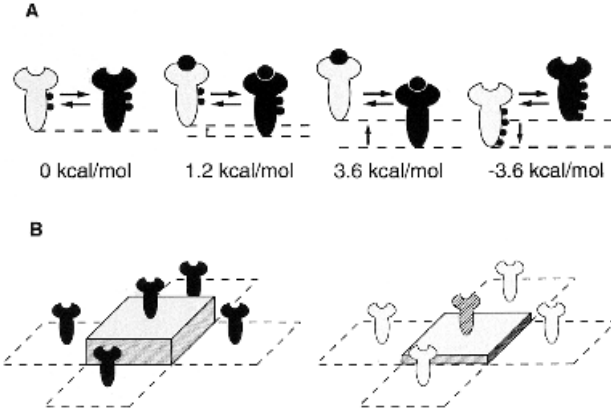


FIG. 2. Free energy-based modelling of receptor activity. White indicates receptors in the active conformation, and black inactive. (A) Dependence of receptor activation energies on ligand binding and methylation. The activation energy is the free energy difference ΔG between the active and inactive states. We set the unliganded receptor with two methyl groups to be the average state with half-maximal activity ($p=0.5$, $\Delta G=0$). The energy of ligand binding can be obtained from the observation that the activity is reduced fourfold when ligand binds (Borkovich & Simon 1990). Solving Eq. 1 for ΔG at 37°C with $p=0.125$ yields 1.2 kcal/mol. Conversely, increasing methylation reduces the activation energy. The two extreme methylation states are depicted (middle right and far right). (B) Activity coupling between nearest neighbours in receptor clusters. Each receptor's activity is influenced by that of its four nearest neighbours, so that the more active neighbours there are, the higher the probability of being active. A small portion (five lattice points) of an extended square lattice is shown here, and the magnitude of ΔG for the receptor at the centre is indicated by the height of the platform. A receptor surrounded by inactive receptors (left panel) has a higher activation energy, and hence a lower probability of being active, than the same receptor when it is surrounded by active neighbours (right panel).

Using these probabilities for receptor activation and experimentally determined rates of the downstream reactions, we constructed the full model of aspartate signalling from ligand binding to CheY phosphorylation. The response time-courses of this model to step stimuli (sudden jumps in concentrations) of aspartate were in good agreement with experiment (Morton-Firth et al 1999). However, the threshold of the response was still much higher than experimental observations, indicating that the sensitivity of the model was still insufficient.

A novel mechanism for signal amplification

This led us to consider the possibility that an as yet unidentified amplification mechanism is responsible for the observed gain in sensitivity. Specifically, we asked whether the spatial organization of molecules in the cell could account for this discrepancy. Significantly, it was shown in 1993 that *E. coli* chemotactic

receptors aggregate in clusters at the cell poles (Maddock & Shapiro 1993). This discovery was particularly striking because it had been previously pointed out that a uniform distribution of receptors over the cell surface would maximize the efficiency of chemoreception (Berg & Purcell 1977). In an attempt to provide an explanation for this observation as well as for the high gain of the system, the idea was put forward that signal amplification could be achieved by interactions between neighbouring receptors in these clusters (Bray et al 1998). Based on this proposal, a Monte Carlo simulation of receptor signalling was developed, and it was shown that a simple mechanism involving nearest-neighbour coupling of activities (Fig. 2B) could enhance the response over a wide dynamic range (Duke & Bray 1999). This model, however, did not include the downstream reactions in the pathway and the receptors were modelled with only one methylation site. To make quantitative comparisons with experimental observations, a more realistic model incorporating these features would be required.

The spatially extended STOCHSIM model





















We therefore sought to extend the STOCHSIM model of the chemotaxis pathway to include a spatial representation of nearest-neighbour interactions in receptor. The original STOCHSIM program did not have any explicit representation of spatial location—implicitly assuming a uniform distribution of molecules throughout the cell volume. The program was therefore modified so as to allow the activity of each receptor to be dependent not only on its own internal state, but also on the state of neighbouring receptors in a cluster. The free energy difference (ΔG) between the active and inactive state of a receptor is now dependent on three inputs, ligand binding (E_L), methylation (E_M) and activity coupling between nearest neighbours (E_J). For simplicity, we assume that contributions from these inputs are independent so that

$$\Delta G = E_L + E_M + E_J \quad (3)$$

E_J takes discrete values determined by the number of active neighbours, so in the case of a square lattice, there are five possible values (with 0, 1, 2, 3 or 4 active neighbours). Solving Equations 1 and 3 for all possible combinations of E_L , E_M and E_J yields the complete set of activation probabilities (p) for the coupled model (Table 2).

The new model of the chemotaxis pathway incorporating these spatial interactions reveals that receptor coupling brings the expected performance much closer to experimental observations. This is readily seen in the impulse response (the response of the system to a brief pulse of stimulus), which provides a succinct phenomenological description of the system's response characteristics.

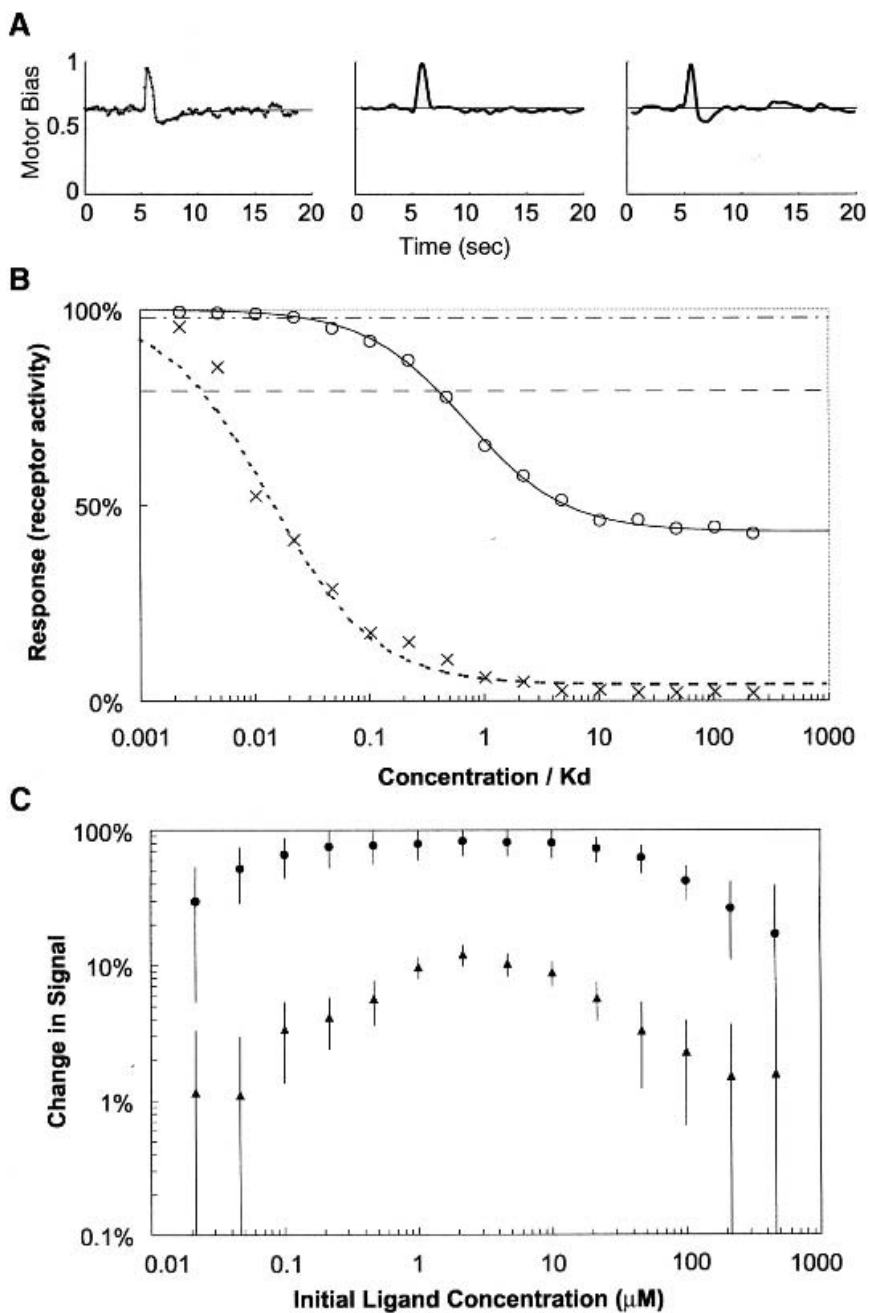
TABLE 2 Free energy changes and activation probabilities for coupled receptors

		<i>Ligand unbound</i>					<i>Ligand bound</i>					
		<i>active neighbours</i>					<i>active neighbours</i>					
	<i>Species</i>	0	1	2*	3	4	<i>Species</i>	0	1	2*	3	4
<i>p</i>		0.00	0.00	0.02	0.31	0.91		0.00	0.00	0.00	0.06	0.59
ΔG		6.22	4.31	2.40	0.49	-1.42		7.42	5.51	3.60	1.69	-0.22
<i>p</i>		0.00	0.01	0.13	0.76	0.99		0.00	0.00	0.02	0.31	0.91
ΔG		5.02	3.11	1.20	-0.71	-2.62		6.22	4.31	2.40	0.49	-1.42
<i>p</i>		0.00	0.04	0.50	0.96	1.00		0.00	0.01	0.13	0.76	0.99
ΔG		3.82	1.91	0.00	-1.91	-3.82		5.02	3.11	1.20	-0.71	-2.62
<i>p</i>		0.01	0.24	0.88	0.99	1.00		0.00	0.04	0.50	0.96	1.00
ΔG		2.62	0.71	-1.20	-3.11	-5.02		3.82	1.91	0.00	-1.91	-3.82
<i>p</i>		0.41	0.94	1.00	1.00	1.00		0.09	0.69	0.98	1.00	1.00
ΔG		0.22	-1.69	-3.60	-5.51	-7.42		1.42	-0.49	-2.40	-4.31	-6.22

* The free energies and activation probabilities for receptors with two active neighbours (shaded column) are equivalent to uncoupled receptors.

Experimentally, it has been shown that the response of *E. coli* cells to short pulses (~ 0.1 s) of aspartate is biphasic (Segall et al 1986, left panel of Fig. 3A). The first phase of the response lasts for approximately one second, over the course of which the motor bias (the probability that the flagellar motor spins in the counter clockwise mode) rapidly jumps to a peak value and then falls below the steady state bias. The second phase of the response is a slower recovery from this undershoot back to the baseline, which lasts approximately four seconds. In the uncoupled STOCHSIM model, the first phase of this response could be reproduced if a sufficiently large pulse of aspartate was applied, but no undershoot could be observed, even in response to a pulse of saturating concentration (middle panel of Fig. 3A). With coupling, however, the STOCHSIM model produces a significant undershoot of a magnitude comparable to the experimentally determined impulse response (right panel of Fig. 3A).

The increased sensitivity due to the activity-coupling mechanism can be observed quantitatively by comparing the dose-response curves of the coupled and uncoupled models. Figure 3B is such a plot which shows the response in receptor activity to steps of aspartate at zero background concentration. There is a noteworthy difference in the shape of the curves (the uncoupled model is satisfactorily fit by a Hill function, whereas the fit to the coupled model is poor). More importantly, there is a ~ 50 -fold difference in the concentration at which



half-maximal inhibition occurs, indicating a significant gain in sensitivity. This enhancement, however, comes at the cost of increased steady-state noise (horizontal lines in Fig. 3B).

The range of ambient concentrations over which the system can respond has been tested by doubling the stimulus concentration after adaptation to an initial stimulus (Fig. 3C). In both the coupled and uncoupled case, the response is masked by the steady-state noise at the high and low extremes of ambient concentration, but the coupled system exhibits a wider range by an order of magnitude. It can also be seen that the response is significantly amplified in the coupled model over the entire range.

Insights from a structural view

One consequence of the way in which the coupled model is formulated is that its performance is very sensitive to the parameter E_J , the energy input due to activity coupling between neighbouring receptors. This suggests that if such a mechanism were to stably provide amplification for the pathway, the receptors would need to be arranged in a well-ordered lattice (Duke & Bray 1999). Another concern in considering amplification is the stoichiometric ratio of receptor to CheW to CheA in the receptor complex. Until recently, this was widely accepted to be 2:2:2, as suggested by binding assays using receptors in membrane preparations

FIG. 3. Performance of the STOCHSIM model with nearest-neighbour coupling between clustered receptors. (A) Impulse response to aspartate. All three panels show the response in motor bias to a brief pulse of aspartate (<0.25 s) at 5 s. The biphasic experimental response (left) reported by Segall et al (1986) could not be reproduced by the STOCHSIM model without receptor coupling, even when pulses of saturating concentration (1 mM) were used (centre). However, when receptor coupling was incorporated into the STOCHSIM model, pulses of comparable size to those used in the experiment ($<0.8 \mu\text{M}$) generated biphasic time-courses of comparable shape and amplitude (right). Motor bias (mb) for the STOCHSIM model was computed from CheYp concentration using the Hill-type equation

$$mb = 1 - \frac{[Yp]^H}{(\langle mb \rangle / (1 - \langle mb \rangle)) [Yp]^H + \langle Yp \rangle^H},$$

where $[Yp]$ is the CheYp concentration, $\langle Yp \rangle$ is the CheYp concentration at steady state, $\langle mb \rangle$ is the motor bias at steady state and the Hill coefficient H was assigned a value of 10 (Cluzel et al 2000). (B) Response to step stimuli at zero background concentration. The initial response prior to adaptation is measured here as the minimum receptor activity encountered within 1 s after stimulus. In the coupled model (crosses), the activity falls off much more rapidly than the uncoupled model (circles) as the step size is increased. (C) Response of system to doubling in concentration after adaptation to an initial stimulus. The response here is measured as the fractional change in receptor activity. It can be seen that the signal is amplified in the coupled model (circles) over the entire range of ambient concentrations tested. Error bars show the steady-state level of noise in the system, which can mask the signal at very low and high ambient concentrations.

and purified cytoplasmic components (Gegner et al 1992). However, a more recent study using soluble receptor fragments has indicated a much higher receptor content, which could have significant consequences on amplification (Liu et al 1997).

These concerns have led us to consider the physical arrangement of receptors in the cluster and the details of how the receptor cytoplasmic domains, CheW and CheA, assemble to form the complex. Fortunately, atomic resolution structures of all three components had been determined in recent years (Kim et al 1999, Bilwes et al 1999, F. W. Dahlquist, personal communication, 2000). We used a somewhat unconventional approach to predict how these structures might assemble into a regular and laterally extendible lattice (Shimizu et al 2000). Briefly, plastic models of all three components were generated using a three-dimensional printer. Guided by mutational data implicating residues which affect the pairwise interactions (Liu & Parkinson 1991, Bass et al 1999) and manual exploration of surface complementarity between these hand-held structures, we were able to assemble a hexagonal lattice composed of trigonal units (Fig. 4). The receptor cytoplasmic domains, which are inserted into the centre of each trigonal unit, are very long coiled-coils of α helices, approximately 26 nm in length. Because it is known that the region of interaction with CheW is at the cytoplasmic end of these 'pillars', the lattice structure of Fig. 4 implies that there will be a significant volume of cytoplasm that is sandwiched between the plasma membrane and the CheA/CheW layer. This may function as an 'adaptation compartment' because all of the receptor residues which are subject to reversible methylation would be located within this region. Sequestration of CheR and/or CheB within such a compartment could have unexpected consequences on adaptation kinetics.

Molecular brachiation: a novel mechanism for adaptation?

The possibility that adaptation kinetics could be affected by the spatial arrangement of molecules in the cell was of particular interest to us: while the kinetics of both adaptation enzymes have been characterized *in vitro* (Simms & Subbaramaiah 1991, Lupas & Stock 1989), a straightforward application of the measured rates had not produced the correct adaptational phenotype in previous models. In the STOCHSIM model, it has been necessary to tune the rate of either CheR or CheB by nearly an order of magnitude to obtain the correct adaptation phenotype (Morton-Firth 1999). In addition, it has been shown recently that both CheR and CheB possess two sites for interacting with the receptors, raising new questions about their kinetic mechanisms. In both CheR and CheB, the first site that binds to the receptors is the catalytic site, which interacts with the methylatable glutamyl residues on the receptors. The second site has an affinity for the C-terminal pentapeptide of the receptors, which is attached to the cytoplasmic domain by

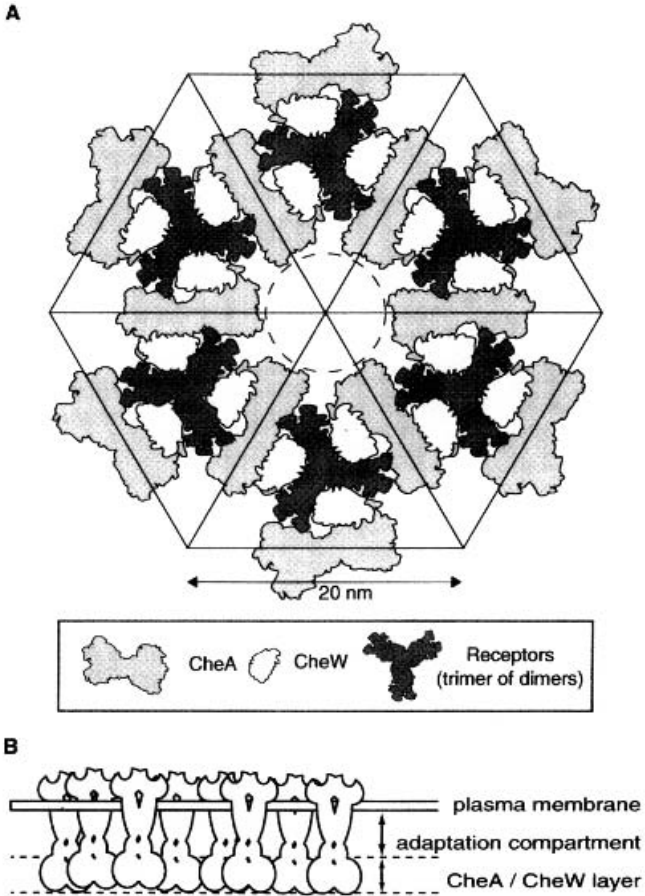


FIG. 4. Hexagonal network consisting of receptors, CheW and CheA predicted from their atomic resolution structures. (A) Plan view, as viewed from the plasma membrane towards the cytoplasm, of a small portion of the lattice. The binding arrangement is such that a network with this geometry could be extended indefinitely in two dimensions. Note the pores at the centre of each hexagon which are large enough (~ 10 nm) for CheR and CheB to pass through. (B) The layer of CheW and CheA, which contains the vertices of the network, is expected to be separated from the plasma membrane by the approximate length of the receptor cytoplasmic domains (~ 26 nm). The cytoplasmic space sandwiched between this layer and the plasma membrane contains all of the methylation sites of the receptors, and thus could serve as an 'adaptation compartment'.

a region of undefined secondary structure. This flexible tether (~ 30 residues) is sufficiently long for a CheR molecule attached at its end to reach the methylation sites of a neighbouring receptor, according to the lattice depicted in Fig. 4, and such inter-receptor methylation has been observed experimentally (Li et al 1997).

We have found that under suitable conditions, the combination of this tethering effect and the proximity of receptors in the lattice could have a significant effect on the movement and localization of CheR (and possibly CheB). By sequential binding and unbinding of the two sites, it is possible that the molecules could move in a hand-over-hand fashion, like an orang-utan swinging through the branches in a jungle (Fig. 5A). We use the term molecular brachiation to characterize this novel mode of movement (Levin et al 2002). Using STOCHSTM, we are currently investigating the feasibility of molecular brachiation as well as

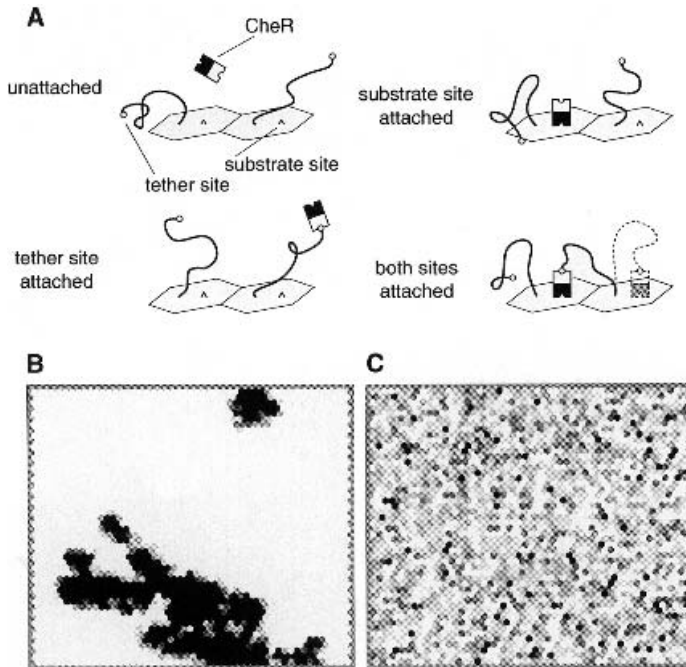


FIG. 5. Molecular brachiation. (A) Schematic illustration of the of brachiation mechanism. Each receptor has two sites to which CheR can bind, one tether site and one site that can undergo methylation. The four panels depict the possible states of binding for the CheR molecule. Note that the presence of two binding sites on the major chemotaxis receptors allows both inter- and intra-receptor binding of CheR. By alternating between states in which one site is attached (upper-right and lower-left panels) and states in which both sites are attached (lower-right panel), CheR could 'brachiate' through a lattice of receptors such as that depicted in Fig. 4. (B) Stochastic simulation of brachiation. A single CheR molecule in a volume of 1.4×10^{-15} l (the approximate volume of a bacterial cell) was allowed to diffuse to a lattice of binding sites and followed over a period of 500 s. (A) Coverage of the lattice by the CheR molecule. Binding sites visited by the molecule are shown in shades of grey, with the intensity indicating the number of repeat visits (1, 2, 3, >4). (C) As for (B) but with the tethers on the receptors removed, so that each receptor has only one binding site for CheR. Brachiation does not occur under these conditions, and the lattice is covered more uniformly, but with fewer return visits to each individual site.

its possible effects on the kinetics of CheR. We predict that such a mechanism would help to sequester CheR in the receptor lattice without compromising its mobility (Fig. 5B, C).

Summary

Because of the unparalleled richness of data regarding its physiology, biochemistry and genetics, many believe that the bacterial chemotaxis pathway is set to become the first cell signalling pathway to be understood ‘completely’. This abundance of information is allowing us to utilise a number of computational methods, including deterministic and stochastic simulations, to reconstruct the pathway *in silico*. While our deterministic model (BCT) allows us to efficiently analyse the broad features of the chemotactic response, the stochastic model (STOCHSIM) is capable of simulating more detailed, physically realistic models. Additionally, the recently determined structures of the component proteins, in conjunction with molecular graphics programs, can be used to explore possible reaction mechanisms and spatial organization.

In general, stochastic simulations are computationally demanding, but for certain types of models noted above, the STOCHSIM algorithm can prove more efficient than its deterministic counterparts. This advantage has been exploited to construct a detailed model of the chemotaxis pathway in which the receptor complex possesses a large complement of molecular states. This model reproduced many features of the physiological response, but singularly failed to reproduce the magnitude of the signal. A two-dimensional spatial structure was implemented in STOCHSIM to reconcile the discrepancy, and the new model with nearest-neighbour coupling produced results that are significantly closer to experimental observations. We also considered the three-dimensional arrangement of the receptor complex and postulated a lattice structure capable of supporting the receptor coupling mechanism. A novel mechanism by which the adaptation enzymes may be sequestered to, but not immobilized at, the receptor cluster has been proposed, and is currently being tested using a combination of deterministic and stochastic simulations.

Acknowledgements

This work was supported by a Glaxo International Scholarship, an Overseas Research Student Award from the Committee of Vice Chancellors and Principals, and a Cambridge Overseas Trust Bursary to TSS. DB is supported by the Medical Research Council.

References

- Alon U, Surette MG, Barkai N, Leibler S 1999 Robustness in bacterial chemotaxis. *Nature* 397:168–171
- Barkai N, Leibler S 1997 Robustness in simple biochemical networks. *Nature* 387:913–917

- Bass RB, Coleman MD, Falke JJ 1999 Signaling domain of the aspartate receptor is a helical hairpin with a localized kinase docking surface: cysteine and disulfide scanning studies. *Biochemistry* 38:9317–9327
- Berg HC, Purcell EM 1977 Physics of chemoreception. *Biophys J* 20:193–219
- Berg HC, Tedesco PM 1975 Transient response to chemotactic stimuli in *Escherichia coli*. *Proc Natl Acad Sci USA* 72:3235–3239
- Bilwes AM, Alex LA, Crane BR, Simon MI 1999 Structure of CheA, a signal-transducing histidine kinase. *Cell* 96:131–141
- Borkovich KA, Simon MI 1990 The dynamics of protein phosphorylation in bacterial chemotaxis. *Cell* 63:1339–1348
- Bray D, Bourret RB 1995 Computer analysis of the binding reactions leading to a transmembrane receptor-linked multiprotein complex involved in bacterial chemotaxis. *Mol Biol Cell* 6:1367–1380
- Bray D, Bourret RB, Simon MI 1993 Computer simulation of the phosphorylation cascade controlling bacterial chemotaxis. *Mol Biol Cell* 4:469–482
- Bray D, Levin MD, Morton-Firth CJ 1998 Receptor clustering as a mechanism to control sensitivity. *Nature* 393:85–88
- Bren A, Eisenbach M 2000 How signals are heard during bacterial chemotaxis: protein–protein interactions in sensory signal propagation. *J Bacteriol* 182:6865–6873
- Cluzel P, Surette M, Leibler S 2000 An ultrasensitive bacterial motor revealed by monitoring signaling proteins in single cells. *Science* 287:1652–1655
- Duke TA, Bray D 1999 Heightened sensitivity of a lattice of membrane receptors. *Proc Natl Acad Sci USA* 96:10104–10108
- Falke JJ, Bass RB, Butler SL, Chervitz SA, Danielson MA 1997 The two component signaling pathway of bacterial chemotaxis: A molecular view of signal transduction by receptors, kinases, and adaptation enzymes. *Annu Rev Cell Dev Biol* 13:457–512
- Gegner JA, Graham DR, Roth AF, Dahlquist FW 1992 Assembly of an MCP receptor, CheW, and kinase CheA complex in the bacterial chemotaxis signal transduction pathway. *Cell* 70:975–982
- Kim KK, Yokota H, Kim SH 1999 Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor. *Nature* 400:787–792
- Levin MD, Morton-Firth CJ, Abouhamad WN, Bourret RB, Bray D 1998 Origins of individual swimming behavior in bacteria. *Biophys J* 74:175–181
- Levin MD, Shimizu TS, Bray D 2002 Binding and diffusion of CheR molecules within a cluster of membrane receptors. *Biophys J* 82:1809–1817
- Li J, Li G, Weis RM 1997 The serine chemoreceptor from *Escherichia coli* is methylated through an inter-dimer process. *Biochemistry* 36:11851–11857
- Liu J, Parkinson JS 1991 Genetic evidence for interaction between CheW and Tsr proteins during chemoreceptor signaling by *Escherichia coli*. *J Bacteriol* 173:4941–4951
- Liu Y, Levit M, Lurz R, Surette MG, Stock JB 1997 Receptor-mediated protein kinase activation and the mechanism of transmembrane signaling in bacterial chemotaxis. *EMBO J* 16:7231–7240
- Lupas A, Stock J 1989 Phosphorylation of an N-terminal regulatory domain activates the CheB methyltransferase in bacterial chemotaxis. *J Biol Chem* 264:17337–17342
- Maddock JR, Shapiro L 1993 Polar location of the chemoreceptor complex in the *Escherichia coli* cell. *Science* 259:1717–1723
- Mesibov R, Ordal GW, Adler J 1973 The range of attractant concentrations for bacterial chemotaxis and the threshold and size of response over this range. Weber law and related phenomenon. *J Gen Physiol* 62:203–223
- Morton-Firth CJ 1998 Stochastic simulation of cell signalling pathways. PhD thesis, University of Cambridge, Cambridge, UK

- Morton-Firth CJ, Bray D 1998 Predicting temporal fluctuations in an intracellular signalling pathway. *J Theor Biol* 192:117–128
- Morton-Firth CJ, Shimizu TS, Bray D 1999 A free-energy-based stochastic simulation of the Tar receptor complex. *J Mol Biol* 286:1059–1074
- Segall JE, Block SM, Berg HC 1986 Temporal comparisons in bacterial chemotaxis. *Proc Natl Acad Sci USA* 83:8987–8991
- Shimizu TS, Le Novere N, Levin MD, Beavil AJ, Sutton BJ, Bray D 2000 Molecular model of a lattice of signalling proteins involved in bacterial chemotaxis. *Nat Cell Biol* 2:792–796
- Simms SA, Subbaramaiah K 1991 The kinetic mechanism of S-adenosyl-L-methionine: glutamylmethyltransferase from *Salmonella typhimurium*. *J Biol Chem* 266:12741–12746

DISCUSSION

Himcb: Your 2D model is the 2D Ising model with a few additional elements added on. The increased sensitivity is very much related to the divergence of the magnetic susceptibility of the Ising model near the critical point, and the increase in background fluctuation is likely to be related to the divergence of the correlation length near the critical point. Both these things very much rely on being close to the critical point. Doesn't that require a very fine tuning of parameters which you do not know?

Shimizu: One thing that I didn't show here is that compared to a standard Ising model, this system with multiple methylation states is less sensitive to the coupling parameter which corresponds to the magnetic susceptibility you mentioned. It is still sensitive, but the border between what you would call the ferromagnetic and paramagnetic behaviours becomes blurred. This is one interesting outcome. But we also think that perhaps there is a mechanism that might account for this in a more robust way.

Subramaniam: The local concentration is going to be very different from overall concentration. This is one of the things that you aren't able to take into account effectively in a model such as yours. You are trying to do this using two dimensions. But this is one of the difficulties in modelling chemotaxis or any other phenomenon like this. There are also some neat experiments that have been done in more complex organisms such as *Dictyostelium*. Have you thought of doing modelling with this?

Shimizu: We'd like to see our program applied to other systems. One thing we think would be very useful is to combine this stochastic algorithm with a deterministic simulator, in a similar fashion to what Raimond Winslow demonstrated earlier with his model of Ca^{2+} channels. This could be used, for example to improve the brachiation simulations. The distributions of sites visited by CheR in the untethered case, as predicted by our present program, is more or less uniform. But more realistically, you would expect this distribution to be more biased, because the CheR molecule performs a three-dimensional random walk

when it is not attached to the lattice. There are efficient analytical expressions for computing this effect (Lagerholm & Thompson 1998), and it would be very interesting to combine such equation-based methods with our individual-based stochastic approach.

Noble: When Raimond Winslow was presenting his work on combining stochastic modelling with differential equation modelling, as I understand it this leads to greatly increased computational times. When I recently heard Dennis Bray present some of this work, he gave the impression that the stochastic computational methods that you are using actually go extremely fast. What is the explanation for this?

Shimizu: If it is the case that there are certain complexes that have a large number of states, so that a large number of equations would need to be integrated at every time point, then stochastic modelling can be faster.

Noble: So it's a matter of whether each of those states were otherwise to be represented by kinetic expressions, rather than by an on-off switch.

Winslow: The reason this is difficult for us is that we are describing stochastic gating of a rather large ensemble of channels in each functional unit. Another confounding variable is the local Ca^{2+} concentration, because this is increasing the total number of states that every one of these channels can be in.

I have a comment. We have now heard about models in three different areas. We have heard about a model of bacterial chemotaxis, neural models that Les Loew described and the cardiac models that Andrew McCulloch and I have talked about. I grant you that in each one of these systems there are different experimental capabilities that may apply, and thereby make the data available for modelling different in each case. But there are a lot of similarities between the mathematics and the computational procedures used in these systems. In each case, we have dealt with issues of stochastic models where the stochastic nature comes in through the nature of channel gating or molecular interactions. We have dealt with ordinary differential equations which arise from systems that are described in laws of mass action, and we have dealt with partial differential equations for systems where there are both reaction and diffusion processes occurring on complicated geometries. Perhaps this is one reason why Virtual Cell is a useful tool for such a community of biologists: it covers so much of what is important in biological modelling. We should see how much overlap there is in these three areas, and whether this is a rather comprehensive class of models defined in these three areas.

Noble: A good way of putting the question would be, 'What is it that is actually missing?' Part of what I suspect is missing at the moment would be the whole field of systems analysis, which presumably can emerge out of the incorporation of pathway modelling into cellular modelling. One of the reasons I regret not having people like Bernhard Palsson here is that we would have seen much more

of that side of things. Are there tricks there that we are missing, that we should have brought out?

Winslow: I would say that this is not a different class of model; it is a technique for analysing models.

Noble: Yes, this could be applicable to a cell or to an immune system.

Subramaniam: I think the missing elements are the actual parameters that can fit in your model at this point, based on the molecular level of detail. We don't have enough of these to do the modelling. Tom Shimizu's paper raised another important point, which is the state dependence. Our lack of knowledge of all the states clearly inhibits us from doing any model that is specific to a system. We are coarse graining all the information into one whole thing.

Winslow: Again, I didn't hear anything in what you just said about a requirement for a new class of models. Rather than new methods of data analysis, you are saying that there may be systems or functionality that we don't yet have powerful experimental tools to fully probe in the same way we can for ion channel function in cardiac myocytes. I agree with that.

Loew: One kind of model that I don't think we have considered here is that of mechanical or structural dynamics, in terms of the physics that controls that. Part of the problem there is also that we don't completely understand that at a molecular level. Virtual Cell deals with reaction–diffusion equations in a static geometry. It isn't so much the static geometry that is the limitation; rather it is that we don't know why that geometry might change. We don't know how to model it because we don't know the physics. We know the physics of reaction–diffusion equations, but the structural dynamics issue is another class of modelling that we haven't done.

Subramaniam: The time-scale is a major issue here. If you want to model at the structural dynamics level, you need to marry different time-scales.

Loew: Getting back to Raimond Winslow's point about the different kinds of modelling, this time-scale by itself does not define a different kind of modelling. The issue is whether the physics is understood.

McCulloch: I agree with both of those points. It seems that what is missing is an accepted set of physical principles by which you can bridge these classes of models, from the stochastic model to the common pool model, and from the common pool model to the reaction–diffusion system. Such physical principles can be found, but I don't think they have been articulated.

Winslow: Yes, we need these rather than our own intuition as to what can be omitted and what must be retained. We need algorithmic procedures for quantifying and performing that.

Paterson: The opportunity to use data at a level above the cell can provide very powerful clues for asking questions of what to explore at the individual cell level. If we are trying to understand behaviour at the tissue, organ or organism level,

this gives us some ways to focus on what mechanisms we may want to investigate at the cellular level. It makes a huge difference in terms of which biologists we work with—for example, whether these are physiologists or clinicians. Many biologists will go on at length about how difficult it is to reproduce *in vivo* environments in *in vitro* experiments. They want to understand things at a higher level.

Winslow: Do you think there is a new class of model at that level, which we haven't considered here yet?

Paterson: No, I think a lot of the issues that we have been talking about are the same at those different levels. In the sort of work my organization does we often run into this issue: if you are starting at the level of biochemical reactions you are much closer to first principles, to the point where if you can actually measure parameters then you can work up to emergent behaviours. But if you are talking with a biologist who studies phenomena significantly above first principles, such as clinical disease, then you have to postulate a hypothesis about what might be responsible for the phenomena and then drill down to see what mechanisms might embody that hypothesis. I'm not sure that there is anything that is fundamentally different, but there are many different domains and specialities in biology, all valuable for providing their unique perspective and data. These perspectives simply change the nature of the conversation.

Crampin: In this discussion of different classes of models, it might also be appropriate to raise the question of different types of algorithms and numerical methods for model solution. The numerical method chosen will of course depend on the sort of models you are dealing with. We have discussed how computer software and hardware will advance over coming years, but we should remember that efforts spent on improving numerical algorithms will pay dividends, especially for more complex problems. Are those people who are developing technologies for biological simulation spending much time considering the different sorts of algorithms that might be used to solve the models? For example, if you are primarily solving reaction–diffusion equations, how much time is spent developing algorithms that run particularly fast for solving the reaction–diffusion models?

Loew: There's a competing set of demands. We use a method called the finite volume method, which is very well adapted to reaction–diffusion equations, but is probably not the best approach. Finite element approaches might be considerably faster. The problem with them, particularly on unstructured grids, is that it is very difficult to create a general-purpose software system that can produce unstructured grids. An experienced modeller would tend to use unstructured grids within a finite element framework; but if we are trying to create a general-purpose software system for biologists, at least so far we haven't been able to think of how to do this.

Subramaniam: Raimond Winslow, with the class of models that you talked about, which are widely applicable, the issues that come up are often boundary

conditions and geometries. How easy is it to develop general-purpose methods that can scale across these? A second issue is that we need to have explosive understanding of feedback regulation coming into the system. It is not obvious to me at this point that this can be taken into account simply by parameterization.

Winslow: The problem with boundary conditions and representing complex geometries is being dealt with rather well by the center for Bioelectric Field Modeling, Simulation and Visualization at the University of Utah (<http://www.sci.utah.edu/ncrr/>). They are building the bio problem-solving environment using Chris Johnson's finite element methods to describe electric current flow in the brain and throughout the body. They have built nice graphical user interfaces for readily adapting these kinds of models. I don't have a sense for whether the applications of those tools have moved to a different and distinct area, but I would offer them as an example of a group that is doing a good job in creating general purpose finite element modelling tools for the community.

Subramaniam: This still doesn't take into account the forces between the different elements that we are dealing with at this point in time. You are doing a stochastic force or a random force. You are not solving Newton's equations, for example. When you try to do this, the complexity becomes quite difficult to deal with, in that it cannot be dealt with in this framework.

Reference

Lagerholm BC, Thompson NL 1998 Theory for ligand rebinding at cell membrane surfaces. *Biophys J* 74:1215–1228

The heart cell *in silico*: successes, failures and prospects

Denis Noble

University Laboratory of Physiology, Parks Road, Oxford OX1 3PT, UK

Abstract. The development of computer models of heart cells is used to illustrate the interaction between simulation and experimental work. At each stage, the reasons for new models are explained, as are their defects and how these were used to point the way to successor models. As much, if not more, was learnt from the way in which models failed as from their successes. The insights gained are evident in the most recent developments in this field, both experimental and theoretical. The prospects for the future are discussed.

2002 'In silico' *simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 182–197*

Modelling is widely accepted in other fields of science and engineering, yet many are still sceptical about its role in biology. One of the reasons for this situation in the case of excitable cells is that the paradigm model, the Hodgkin–Huxley (1952) equations for the squid nerve action potential, was so spectacularly successful that, paradoxically, it may have created an unrealistic expectation for its rapid application elsewhere. By contrast, modelling of the much more complex cardiac cell has required many years of iterative interaction between experiment and theory, a process which some have regarded as a sign of failure. But, in modelling complex biological phenomena, this is in fact precisely what we should expect (see discussions in Novartis Foundation 2001), and it is standard for such interaction to occur over many years in other sciences. Successful models of cars, bridges, aircraft, the solar system, quantum mechanics, cosmology and so on all go through such a process. I will illustrate this interaction in biological simulation using some of the models I have been involved in developing. Since my purpose is didactic, I will be highly selective. A more complete historical review of cardiac cell models can be found elsewhere (Noble & Rudy 2001) and the volume in which that article appeared is also a rich source of material on modelling the heart, since that was its focus.

The developments I will use in this paper will be described in four 'Acts', corresponding to four of the stages at which major shifts in modelling paradigm

occurred. They also correspond to points at which major insights occurred, most of which are now ‘accepted wisdom’. It is the fate of insights that were hard-won at the time to become obvious later. This review will also therefore serve the purpose of reminding readers of the role simulation played in gaining them in the first place.

Act I—Energy conservation during the cardiac cycle: nature’s ‘pact with the devil’

FitzHugh (1960) showed that the Hodgkin–Huxley model of the nerve impulse could generate a long plateau, similar to that occurring during the cardiac action potential, by greatly reducing the amplitude and speed of activation of the delayed K^+ current, I_K . These changes not only slowed repolarization; they also created a plateau. This gave the clue that there must be some property inherent in the Hodgkin–Huxley formulation of the sodium current that permits a persistent inward current to occur. The main defect of the FitzHugh model was that it was a very expensive way of generating a plateau, with such high ionic conductances that during each action potential the Na^+ and K^+ ionic gradients would be run down at a rate at least an order of magnitude too large.

That this was not the case was already evident since Weidmann’s (1951, 1956) results showed that the plateau conductance in Purkinje fibres is very low. The experimental reason for this became clear with the discovery of the inward-rectifier current, I_{K1} (Hutter & Noble 1960, Carmeliet 1961, Hall et al 1963). The permeability of the I_{K1} channel falls almost to zero during strong depolarization. These experiments were also the first to show that there are at least two K^+ conductances in the heart, I_{K1} and I_K (referred to as I_{K2} in early work, but now known to consist of I_{Kr} and I_{Ks}). The Noble (1960, 1962) model was constructed to determine whether this combination of K^+ channels, together with a Hodgkin–Huxley type Na^+ channel could explain all the classical Weidmann experiments on conductance changes. The model not only succeeded in doing this; it also demonstrated that an energy-conserving plateau mechanism was an automatic consequence of the properties of I_{K1} . This has featured in all subsequent models, and it is a very important insight. The main advantage of a low conductance is minimizing energy expenditure.

Unfortunately, however, a low conductance plateau was achieved at the cost of making the repolarization process fragile. Pharmaceutical companies today are struggling to deal with evolution’s answer to this problem, which was to entrust repolarization to the K^+ channel I_{Kr} . A ‘pact with the devil’, indeed! This is one of the most promiscuous receptors known: large ranges of drugs can enter the channel mouth and block it, and even more interact with the G protein-coupled receptors that control it. Molecular promiscuity has a heavy price: roughly US\$0.5 billion per drug withdrawn. Simulation is now playing a major role in attempting

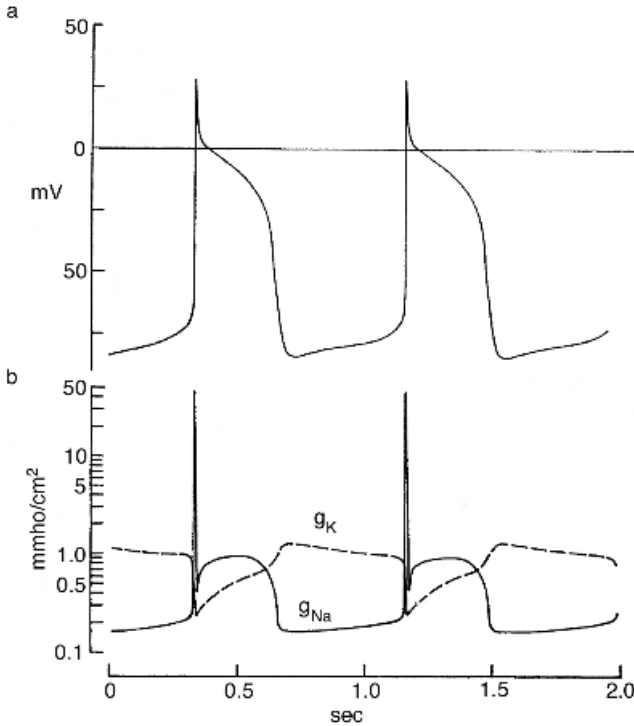


FIG. 1. Na^+ and K^+ conductance changes computed from the 1962 model of the Purkinje fibre. Two cycles of activity are shown. The conductances are plotted on a logarithmic scale to accommodate the large changes in Na^+ conductance. Note the persistent level of Na^+ conductance during the plateau of the action potential, which is about 2% of the peak conductance. Note also the rapid fall in K^+ conductance at the beginning of the action potential. This is attributable to the properties of the inward rectifier I_{K1} (Noble 1962).

to find a way around this difficult and intractable problem (Muzikant & Penland 2002).

Figure 1 shows the ionic conductance changes computed from this model. The ‘emergence’ of a plateau Na^+ conductance is clearly seen, as is the dramatic fall in K^+ conductance at the beginning of the action potential. Both of these fundamental insights have featured in all subsequent models of cardiac cells.

The main defect of the 1962 model was that it included only one voltage gated inward current, I_{Na} . There was a good reason for this. Ca^{2+} currents had not then been discovered. There was, nevertheless, a clue in the model that something important was missing. The only way in which the model could be made to work was to *greatly* extend the voltage range of the Na^+ ‘window’ current by reducing the voltage dependence of the Na^+ activation process (see Noble 1962 [Fig. 15]). In

effect, the Na^+ current was made to serve the function of both the Na^+ and Ca^{2+} channels so far as the plateau is concerned. There was a clear prediction here: either Na^+ channels in the heart are quantitatively different from those in nerve, or other inward current-carrying channels must exist. Both predictions are correct.

The first successful voltage clamp measurements came in 1964 (Deck & Trautwein 1964) and they rapidly led to the discovery of the cardiac Ca^{2+} current (Reuter 1967). By the end of the 1960s therefore, it was already clear that the 1962 model needed replacing.

Act II—Controversy over the ‘pacemaker’ current: the MNT model

In addition to the discovery of the Ca^{2+} current, the early voltage clamp experiments also revealed multiple components of I_K (Noble & Tsien 1969) and that these slow gated currents in the plateau range of potentials were quite distinct from those near the resting potential, i.e. that there were two separate voltage ranges in which very slow conductance changes could be observed (Noble & Tsien 1968, 1969). These experiments formed the basis of the MNT model (McAllister et al 1975).

This model reconstructed a much wider range of experimental results, and it did so with great accuracy in some cases. A good example of this was the reconstruction of the paradoxical effect of small current pulses on the pacemaker depolarisation in Purkinje fibres (see Fig. 2)—paradoxical because brief depolarisations *slow* the process and brief hyperpolarizations greatly *accelerate* it. Reconstructing paradoxical or counterintuitive results is of course a major function of modelling work. This is one of the roles of modelling in unravelling complexity in biological systems.

But the MNT model also contained the seeds of a spectacular failure. Following the experimental evidence (Noble & Tsien 1968) it attributed the slow conductance changes near the resting potential to a slow-gated K^+ current, I_{K2} . In fact, what became the ‘pacemaker current’, or I_f , is an *inward* current activated by hyperpolarization (DiFrancesco 1981) not an *outward* current activated by depolarization. At the time it seemed hard to imagine a more serious failure than getting *both* the current direction *and* the gating by voltage completely wrong. There cannot be much doubt therefore that this stage in the iterative interaction between experiment and simulation created a major problem of credibility. Perhaps cardiac electrophysiology was not really ready for modelling work to be successful?

This was how the failure was widely perceived. Yet it was a deep misunderstanding of the significance of what was emerging from this experience.

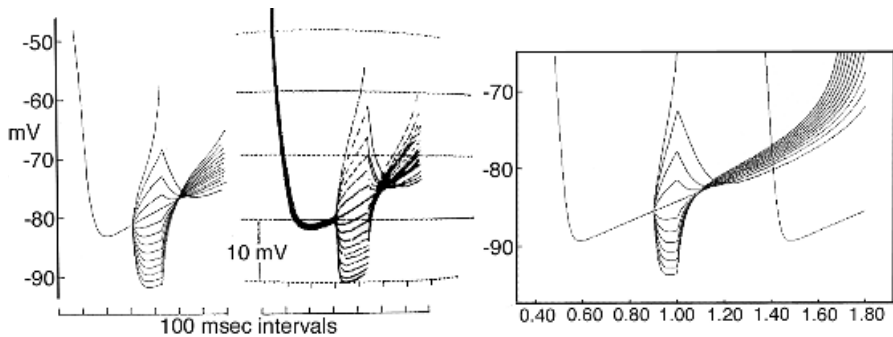


FIG. 2. Reconstruction of the paradoxical effect of small currents injected during pacemaker activity. (Left) Computations from the MNT model (McAllister et al 1975). Small depolarizing and hyperpolarizing currents were applied for 100 ms during the middle of the pacemaker depolarization. Hyperpolarizations are followed by an acceleration of the pacemaker depolarization, while subthreshold depolarizations induce a slowing. (Middle) Experimental records from Weidmann (1951, Fig. 3). (Right) Similar computations using the DiFrancesco–Noble (DiFrancesco & Noble 1985) model. Despite the fundamental differences between these two models, the feature that explains the paradoxical effects of small current pulses survives. This kind of detailed comparison was part of the process of mapping the two models onto each other.

It was no coincidence that both the current direction and the gating were wrong as one follows from the other. And so did much else in the modelling! Working that out in detail was the ground on which future progress could be made.

This is the point at which to make one of the important points about the philosophy of modelling. It is one of the functions of models to be wrong! Not, of course, in arbitrary or purely contingent ways, but in ways that advance our understanding. Again, this situation is familiar to those working in simulation studies in engineering or cosmology or in many other physical sciences. And, in fact, the failure of the MNT model is one of the most instructive examples of experiment–simulation interaction in physiology, and of subsequent successful model development. I do not have the space here to review this issue in all its details. From an historical perspective, that has already been done (see DiFrancesco & Noble 1982, Noble 1984). Here I will simply draw the conclusions relevant to modern work.

First, careful analysis of the MNT model revealed that its pacemaker current mechanism could not be consistent with what is known of the process of ion accumulation and depletion in the extracellular spaces between cells. The model itself was therefore a key tool in understanding the next stage of development.

Second, a complete and accurate mapping between the I_{K2} model and the new I_f model could be constructed (DiFrancesco & Noble 1982) demonstrating how

both models related to the *same* experimental results and to each other. Such mapping between different models is rare in biological work, but it can be very instructive.

Third, this spectacular turn-around was the trigger for the development of models that include changes in ion concentrations inside and outside the cell, and between intracellular compartments.

Finally, the MNT model was the point of departure for the ground-breaking work of Beeler & Reuter (1977) who developed the first ventricular cell model. As they wrote of their model: ‘In a sense, it forms a companion presentation to the recent publication of McAllister et al (1975) on a numerical reconstruction of the cardiac Purkinje fibre action potential. There are sufficiently many and important differences between these two types of cardiac tissue, both functionally and experimentally, that a more or less complete picture of membrane ionic currents in the myocardium must include both simulations.’ For a recent assessment of this model see Noble & Rudy (2001).

The MNT and Beeler–Reuter papers were the last cardiac modelling papers to be published in the *Journal of Physiology*. I don’t think the editors ever recovered from the shock of discovering that models could be wrong! The leading role as publisher was taken over first by the journals of The Royal Society, and then by North American journals.

Act III— Ion concentrations, pumps and exchangers: the DiFrancesco–Noble model

The incorporation not only of ion channels (following the Hodgkin–Huxley paradigm) but also of ion exchangers, such as $\text{Na}^+\text{--K}^+$ exchange (the Na^+ pump), $\text{Na}^+\text{--Ca}^{2+}$ exchange, the SR Ca^{2+} pump and, more recently, all the transporters involved in controlling cellular pH (Ch’en et al 1998), was a fundamental advance since these are essential to the study of some disease states such as congestive heart failure and ischaemic heart disease.

It was necessary to incorporate the $\text{Na}^+\text{--K}^+$ exchange pump since what made I_f so closely resemble a K^+ channel in Purkinje fibres was the depletion of K^+ in extracellular spaces. This was a key feature enabling the accurate mapping of the I_{K2} model (MNT) onto the I_f model (DiFrancesco & Noble 1982). But, to incorporate changes in ion concentrations it became necessary to represent the processes by which ion gradients can be restored and maintained. In a form of modelling ‘avalanche’, once changes in one cation concentration gradient (K^+) had been introduced, the others (Na^+ and Ca^{2+}) had also to be incorporated since the changes are all linked via the $\text{Na}^+\text{--K}^+$ and $\text{Na}^+\text{--Ca}^{2+}$ exchange mechanisms. This ‘avalanche’ of additional processes was the basis of the DiFrancesco–Noble (1985) Purkinje fibre model (Fig. 3).

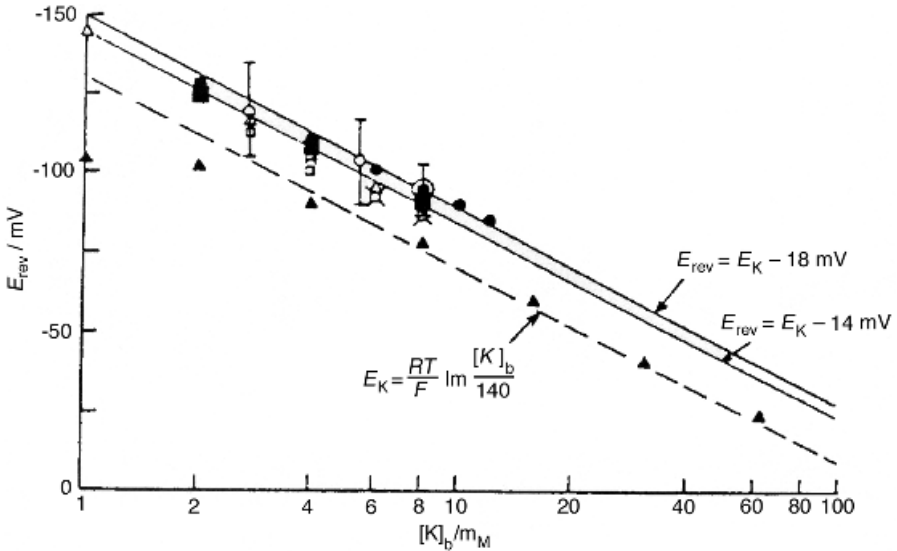


FIG. 3. Mapping of the different models of the ‘pacemaker’ current. The filled triangles show the experimental variation of the resting potential with external bulk potassium concentration, $[K^+]_b$, which closely follows the Nernst equation for K^+ above 4 mM. The open symbols show various experimental determinations of the apparent ‘reversal potential’ for the pacemaker current. The closed circles and the solid lines were derived from the DiFrancesco–Noble (1985) model. The new model not only accounted for the remarkable ‘Nernstian’ behaviour of the apparent reversal potential; it also accounted for the fact that all the experimental points are above (more negative than) the real Nernst potential by around 10–20 mV (the solid lines show 14 and 18 mV discrepancies).

Biological modelling often exhibits this degree of modularity, making it necessary to incorporate a group of protein components together. It will be one of the major challenges of mathematical biology to use simulation work to unravel the modularity of nature. Groups of proteins co-operating to generate a function and therefore being selected together in the evolutionary process will be revealed by this approach. This piecemeal approach to reconstructing the ‘logic of life’ (which is the strict meaning of the word ‘physiology’ — see Boyd & Noble 1993) could also be the route through which a systematic theoretical biology could eventually emerge (see the concluding discussion of this meeting).

The greatly increased complexity of the DiFrancesco–Noble model, which for the first time also represented intracellular events by incorporating a model of calcium release from the sarcoplasmic reticulum, increased both the range of predictions and the opportunities for failure. Here I will limit myself to one example of each.

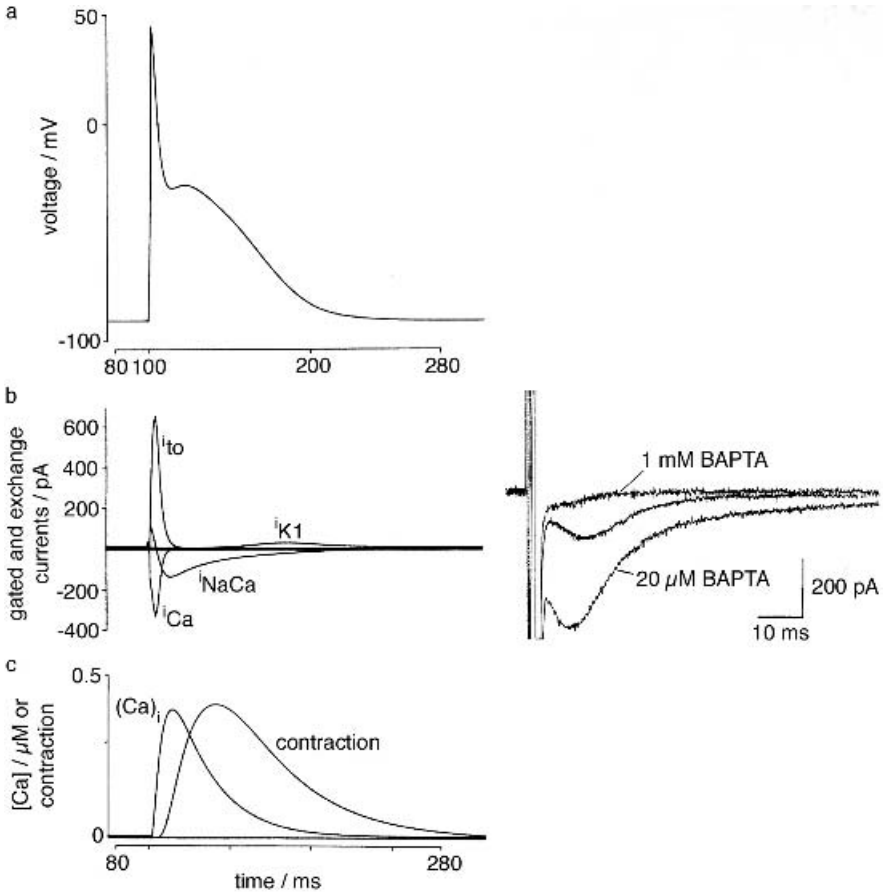


FIG. 4. The first reconstruction of Ca^{2+} balance in cardiac cells. The Hilgemann–Noble model incorporated complete Ca^{2+} cycling, such that intracellular and extracellular Ca^{2+} levels returned to their original state after each cycle and that the effects of sudden changes in frequency could be reproduced. (Left) Simulation using the single-cell version of the model (Earm & Noble 1990). (a) Action potential. (b) Some of the ionic currents involved in shaping repolarization. (c) Intracellular Ca^{2+} transient and contraction. (Right) Experimental recordings of ionic current during voltage clamps at the level (-40 mV) of the late phase of repolarization showing a time course very similar to the computed Na^+-Ca^{2+} exchange current. As the Ca^{2+} buffer (BAPTA) was infused to raise its concentration from 20 μM to 1 mM the current is suppressed (from Earm et al 1990).

Perhaps the most influential prediction was that relating to the Na^+-Ca^{2+} exchanger. In the early 1980s it was still widely thought that the original electrically neutral stoichiometry ($Na^+:Ca^{2+}=2:1$) derived from the early flux measurements was correct. The DiFrancesco–Noble model achieved two

important conclusions. The first was that, with the experimentally known Na^+ gradient, there simply wasn't enough energy in a neutral exchanger to keep resting intracellular Ca^{2+} levels below $1\ \mu\text{M}$. Switching to a stoichiometry of 3:1 readily allowed resting Ca^{2+} to be maintained below $100\ \text{nM}$. This automatically led to the prediction that there must be a current carried by the $\text{Na}^+-\text{Ca}^{2+}$ exchanger and that, if this exchanger was activated by intracellular Ca^{2+} , it must also be strongly time-dependent as intracellular Ca^{2+} varies by an order of magnitude during each action potential. Even as the model was being published, experiments demonstrating the current I_{NaCa} were being performed (Kimura et al 1986) and the variation of this current during activity was being revealed either as a late component of inward current or as a current tail on repolarization.

The main failure was that the intracellular Ca^{2+} transient was far too large. This signalled the need to incorporate intracellular Ca^{2+} buffering.

Act IV — Ca^{2+} balance: the Hilgemann–Noble model

This deficiency was tackled in the Hilgemann–Noble (1987) modelling of the atrial action potential (Fig. 4). Although this was directed towards atrial cells, it also provided a basis for modelling ventricular cells in species (rat, mouse) with short ventricular action potentials. This model addressed a number of important questions concerning Ca^{2+} balance:

- (1) When does the Ca^{2+} that enters during each action potential return to the extracellular space? Does it do this during diastole (as most people had presumed) or during systole itself, i.e. during, not after, the action potential? Hilgemann (1986) had done experiments with tetramethylmurexide, a Ca^{2+} indicator restricted to the extracellular space, showing that the recovery of extracellular Ca^{2+} (in intercellular clefts) occurs remarkably quickly. In fact, net Ca^{2+} efflux is established as soon as $20\ \text{ms}$ after the beginning of the action potential, which at that time was considered to be surprisingly soon. Ca^{2+} activation of efflux via the $\text{Na}^+-\text{Ca}^{2+}$ exchanger achieved this in the model (see Hilgemann & Noble 1987, Fig. 2).
- (2) Where was the current that this would generate and did it correspond to the quantity of Ca^{2+} that the exchanger needed to pump? Mitchell et al (1984) had already done experiments in rat ventricle showing that replacement of Na^+ with Li^+ removes the late plateau. This was the first experimental evidence that the late plateau in action potentials with this shape might be maintained by $\text{Na}^+-\text{Ca}^{2+}$ exchange current. The Hilgemann–Noble model showed that this is what one would expect.

- (3) Could a model of the SR that reproduces at least the major features of Fabiato's (1983, 1985) experiments showing Ca^{2+} -induced Ca^{2+} release (CICR) be incorporated into the cell models and integrate in with whatever were the answers to questions 1–2? This was a major challenge (Hilgemann & Noble 1987). The model followed as much of the Fabiato data as possible, but the conclusions were that the modelling, while broadly consistent with the Fabiato work, could not be based on that alone. It is an important function of simulation to reveal when experimental data needs extending.
- (4) Were the quantities of Ca^{2+} , free and bound, at each stage of the cycle consistent with the properties of the cytosol buffers? The answer here was a very satisfactory 'yes'. The great majority of the cytosol Ca^{2+} is bound so that, although much more calcium movement was involved, the free Ca^{2+} transients were much smaller, within the experimental range.

There were however some gross inadequacies in the Ca^{2+} dynamics. An additional voltage-dependence of Ca^{2+} release was inserted to obtain a fast Ca^{2+} transient. This was a compromise that really requires proper modelling of the subsarcolemmal space where Ca^{2+} channels and the ryanodine receptors interact, a problem later tackled by Jafri et al (1998) (also see recent review by Winslow et al 2000, Noble et al 1998). Another problem was how the conclusions would apply to action potentials with high plateaus. This was tackled both experimentally (Le Guennec & Noble 1994) and computationally (Noble et al 1991, 1998). The answer is that the high plateau in ventricular cells of guinea-pig, dog, human, etc., greatly delays the reversal of the Na^+ – Ca^{2+} exchanger so that net Ca^{2+} entry continues for a longer fraction of the action potential. This property is important in determining the force-frequency characteristics.

I end this historical survey at this point, not because this is the end of the story (see Noble & Rudy 2001), but because these examples deal with the major developments that formed the groundwork for all the current, enormously wide, generation of cellular models of the heart (all cell types have now been modelled, including spatial variations in expression levels), and they illustrate the main conclusions regarding *in silico* techniques that I think are relevant to this meeting.

Finale — Future challenges and the nature of biological simulation

This article has focused on the period up to 1990, which can be regarded as the 'classical period' in which the main foundations of all cardiac cellular models were laid. Since 1990 there has been an explosion of modelling work on the heart (see Hunter et al 2001, and the volume that this article introduces). There are multiple models of all the cell types, and I confidently predict that there will be

many more to come. Why do we have so many? Couldn't we simply 'standardize' the field and choose the 'best'? To some extent, that is happening. None of the historical models described in this article are now used much in their original form. Knowledge *does* advance, and so do the models that represent it! Nevertheless, it would be a mistake to think that there can be one, canonical, model of anything.

One of the major reasons for the multiplicity of models is that there will always be a compromise between complexity and computability. A good example here is the modelling of Ca^{2+} dynamics (discussed in more detail elsewhere in this volume). As we understand these dynamics in ever greater detail, models become more accurate and they encompass more biological detail, but they also become computationally demanding. This was the motivation behind the simplified dyadic space model of Noble et al (1998), which achieves many of the required features of the initiation of Ca^{2+} signalling with only a modest (10%) increase in computation time, an important consideration when importing such models into models of the whole heart. But no one would use that model to study the fine properties of Ca^{2+} dynamics at the subcellular level. That was not its purpose. There will probably therefore be no unique model that does everything at all levels. Any of the boxes at one level could be deepened in complexity at a lower level, or fused with other processes at a higher level. In any case, all models are only partial representations of reality. One of the first questions to ask of a model therefore is what questions does it answer best. It is through the iterative interaction between experiment and simulation that we will gain that understanding.

It is however already clear that incorporation of cell models into tissue and organ models is capable of spectacular insights. The incorporation of cell models into anatomically detailed heart models (recently extensively reviewed by Kohl et al 2000) has been an exciting development. The goal of creating an organ model capable of spanning the whole spectrum of levels from genes (see Clancy & Rudy 1999, Noble & Noble 1999, 2000) to the electrocardiogram (see Muzikant & Penland 2002, Noble 2002) is within sight, and is one of the challenges of the immediate future. The potential of such simulations for teaching, drug discovery, device development and, of course, for pure physiological insight is only beginning to be appreciated.

References

- Beeler GW, Reuter H 1977 Reconstruction of the action potential of ventricular myocardial fibres. *J Physiol* 268:177–210
- Boyd CA, Noble D 1993 *The logic of life*. Oxford University Press, Oxford
- Carmeliet EE 1961 Chloride ions and the membrane potential of Purkinje fibres. *J Physiol* 156:375–388

- Ch'en FF, Vaughan-Jones RD, Clarke K, Noble D 1998 Modelling myocardial ischaemia and reperfusion. *Prog Biophys Mol Biol* 69:515–538
- Clancy CE, Rudy Y 1999 Linking a genetic defect to its cellular phenotype in a cardiac arrhythmia. *Nature* 400:566–569
- Deck KA, Trautwein W 1964 Ionic currents in cardiac excitation. *Pflügers Arch* 280:65–80
- DiFrancesco D 1981 A new interpretation of the pace-maker current in calf Purkinje fibres. *J Physiol* 314:359–376
- DiFrancesco D, Noble D 1982 Implications of the re-interpretation of I_{K2} for the modelling of the electrical activity of pacemaker tissues in the heart. In: Bouman LN, Jongsma HJ (eds) *Cardiac rate and rhythm*. Nijhoff, Dordrecht p 93–128
- DiFrancesco D, Noble D 1985 A model of cardiac electrical activity incorporating ionic pumps and concentration changes. *Philos Trans R Soc B Biol Sci* 307:353–398
- Earm YE, Noble D 1990 A model of the single atrial cell: relation between calcium current and calcium release. *Proc R Soc Lond B Biol Sci* 240:83–96
- Earm YE, Ho WK, So IS 1990 Inward current generated by Na–Ca exchange during the action potential in single atrial cells of the rabbit. *Proc R Soc Lond B Biol Sci* 240:61–81
- Fabiato A 1983 Calcium-induced release of calcium from the cardiac sarcoplasmic reticulum. *Am J Physiol* 245:C1–C14
- Fabiato A 1985 Time and calcium dependence of activation and inactivation of calcium-induced release of calcium from the sarcoplasmic reticulum of a skinned canine cardiac Purkinje cell. *J Gen Physiol* 85:247–298
- FitzHugh R 1960 Thresholds and plateaus in the Hodgkin-Huxley nerve equations. *J Gen Physiol* 43:867–896
- Hall AE, Hutter OF, Noble D 1963 Current-voltage relations of Purkinje fibres in sodium-deficient solutions. *J Physiol* 166:225–240
- Hilgemann DW 1986 Extracellular calcium transients and action potential configuration changes related to post-stimulatory potentiation in rabbit atrium. *J Gen Physiol* 87:675–706
- Hilgemann DW, Noble D 1987 Excitation-contraction coupling and extracellular calcium transients in rabbit atrium: reconstruction of basic cellular mechanisms. *Proc R Soc Lond B Biol Sci* 230:163–205
- Hodgkin AL, Huxley AF 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* 117:500–544
- Hunter PJ, Kohl P, Noble D 2001 Integrative models of the heart: achievements and limitations. *Philos Trans R Soc Lond A Math Phys Sci* 359:1049–1054
- Hutter OF, Noble D 1960 Rectifying properties of heart muscle. *Nature* 188:495
- Jafri MS, Rice JJ, Winslow RL 1998 Cardiac Ca^{2+} dynamics: the roles of ryanodine receptor adaptation and sarcoplasmic reticulum load. *Biophys J* 74:1149–1168
- Kimura J, Noma A, Irisawa H 1986 Na–Ca exchange current in mammalian heart cells. *Nature* 319:596–597
- Kohl P, Noble D, Winslow RL, Hunter P 2000 Computational modelling of biological systems: tools and visions. *Philos Trans R Soc Lond A Math Phys Sci* 358:579–610
- Le Guennec JY, Noble D 1994 Effects of rapid changes of external Na^+ concentration at different moments during the action potential in guinea-pig myocytes. *J Physiol* 478:493–504
- McAllister RE, Noble D, Tsien RW 1975 Reconstruction of the electrical activity of cardiac Purkinje fibres. *J Physiol* 251:1–59
- Mitchell MR, Powell T, Terrar DA, Twist VA 1984 The effects of ryanodine, EGTA and low-sodium on action potentials in rat and guinea-pig ventricular myocytes: evidence for two inward currents during the plateau. *Br J Pharmacol* 81:543–550
- Muzikant AL, Penland RC 2002 Models for profiling the potential QT prolongation risk of drugs. *Cur Opin Drug Discov Dev* 5:127–135

- Noble D 1960 Cardiac action and pacemaker potentials based on the Hodgkin-Huxley equations. *Nature* 188:495–497
- Noble D 1962 A modification of the Hodgkin-Huxley equations applicable to Purkinje fibre action and pacemaker potentials. *J Physiol* 160:317–352
- Noble D 1984 The surprising heart: a review of recent progress in cardiac electrophysiology. *J Physiol* 353:1–50
- Noble D 2002 Modelling the heart: from genes to cells to the whole organ. *Science* 295:1678–1682
- Noble D, Noble PJ 1999 Reconstruction of cellular mechanisms of genetically-based arrhythmias. *J Physiol* 518:2–3P
- Noble D, Tsien RW 1968 The kinetics and rectifier properties of the slow potassium current in cardiac Purkinje fibres. *J Physiol* 195:185–214
- Noble D, Tsien RW 1969 Outward membrane currents activated in the plateau range of potentials in cardiac Purkinje fibres. *J Physiol* 200:205–231
- Noble D, Rudy Y 2001 Models of cardiac ventricular action potentials: iterative interaction between experiment and simulation. *Philos Trans R Soc A Maths Phys Sci* 359:1127–1142
- Noble D, Varghese A, Kohl P, Noble PJ 1998 Improved guinea-pig ventricular cell model incorporating a dyadic space, I_{Kr} and I_{Ks} , and length- and tension-dependent processes. *Can J Cardiol* 14:123–134
- Noble D, Noble SJ, Bett GCL, Earm YE, Ho WK, So IS 1991 The role of sodium–calcium exchange during the cardiac action potential. *Ann NY Acad Sci* 639:334–353
- Noble PJ, Noble D 2000 Reconstruction of the cellular mechanisms of cardiac arrhythmias triggered by early after-depolarizations. *Jpn J Electrocardiol* 20:15–19
- Novartis Foundation 2001 Complexity in biological information processing. Wiley, Chichester (Novartis Found Symp 239)
- Reuter H 1967 The dependence of the slow inward current in Purkinje fibres on the extracellular calcium concentration. *J Physiol* 192:479–492
- Weidmann S 1951 Effect of current flow on the membrane potential of cardiac muscle. *J Physiol* 115:227–236
- Weidmann S 1956 *Elektrophysiologie der herzmuskelfaser*. Huber, Bern
- Winslow RL, Scollan DF, Holmes A, Yung CK, Zhang J, Jafri MS 2000 Electrophysiological modeling of cardiac ventricular function: from cell to organ. *Ann Rev Biomed Eng* 2:119–155

DISCUSSION

Winslow: I think there are some instances where the reverse happens: sometimes the experiments are at fault, and not the models. There's a tendency among biologists to think of the experimental data as being the last word. They don't always appreciate that there are many things that can't be controlled in their particular preparations. Sometimes the model can shed insight into what those uncontrolled variables might be and explain a discrepancy between experiment and model.

Noble: You gave a nice example of this in your work: the failure to realize that Ca^{2+} buffers don't do the job we hoped they would do. This is a good example of this kind of problem in experimental analysis.

Asburner: Sydney Brenner put this very well: we should never throw away a good theory because of bad facts (Brenner 2001).

Crampin: Denis Noble, if you are right in saying that models are most useful when they fail, and that this is a message that needs to be got across to the biology community, could this not lead to a problem if we are also trying to sell these technologies to the pharmaceutical industry? If they think that part of the point of what we are doing is that simulations will also fail, might they be less ready to take them on board?

Noble: I believe that the pharmaceutical industry is vastly more sophisticated than that. Of course, we don't design a model to fail. Let me illustrate a good way in which one could put the point that would be of relevance to the pharmaceutical industry, or indeed any of us with an interest in unravelling the 'logic of life'. Suppose that you find that there is an element missing from your model, or at least what you have got is an aspect of what it is that you are trying to reconstruct but you can't reconstruct it. A good example of this in one of my areas of modelling, pacemaker activity in the heart, would be the recent discovery by Akinori Noma and his colleagues in Japan of yet another pacemaker mechanism (Guo et al 1995). There are two things that modelling has contributed to that, including the models that failed earlier on because they lacked it. The first is an understanding of the robustness of that particular functionality. If you have something like four fail-safe mechanisms involved in the pacemaker mechanism, then it is clearly an important thing for evolution to have developed. It is not surprising that it has developed so many fail-safe mechanisms. What the progressive addition of one mechanism after another involved is revealing is that you have unravelled part of the logic of life, part of the reason for the robustness of that particular physiological function.

Berridge: What might the evolutionary pressure have been? What were the evolutionary changes that would have led to a cell selecting such a fail-safe mechanism? Presumably if it has selected a second mechanism and the first one failed, it would have a selective advantage over an organism with just one. But it is difficult to imagine how a cell would develop three or four different fail-safe mechanisms.

Asburner: Noble's example of four fail-safe heart pacemakers illustrates the 'Boeing 747' theory of evolution. It is very common among naïve molecular biologists.

Noble: The answer I was going to give was to refer to the new mechanism that Akinori Noma has identified, which is a low voltage-activated Na^+ channel. It comes in at a certain phase in the pacemaker depolarization. If you put it into our cell models, there is virtually no change. It is as though this particular mechanism doesn't matter until you start to put on agents such as acetylcholine or adrenaline that change frequency. Then, this mechanism turns out to be a beautiful refining mechanism. I don't know how evolution discovered that, but I can see its function. The previous models show that this refinement is lacking. It is not a case where the

modelling actually identified the need for an extra channel, but it has certainly enabled us to understand this and then in turn to understand if you developed a drug to go for this what use it would be. Now let me give you an example the other way round, where the spotting of a failure helped enormously. One of the gaps that early pacemaker modelling identified was the need to suppose that there had to be a background Na^+ channel. That is, not a Na^+ channel that activates at the threshold for the standard Na^+ current, but that there is a background Na^+ flux. At the time this was introduced there was no experimental evidence for it. It has now been confirmed that there is such a background Na^+ channel (Kiyosue et al 1993): we know its characteristics and selectivity, but we don't know its protein or gene. It leads to a very significant result. The background channel contributes about 80% of the pacemaker depolarization. If this had been wrong, it would have been a huge mistake, but it was necessary and the modelling identified it as necessary. We have no blocker for this channel at the moment, but in the model you can do the 'thought experiment' of blocking it. It produces a counterintuitive result. Since it is carrying 80% of the current, if we block it we'd at least expect to see a large slowing. But what we see is that there is almost no change in pacemaker activity: a fail-safe mechanism kicks in and keeps the pacemaker mechanism going (Noble et al 1992). There is just a slight deceleration. Such a deceleration of cardiac pacemaker activity could be therapeutic in certain circumstances, and so attempts to find cardiac slowers might be worthwhile. If we could find a drug that targets this channel it would be a marvellous cardiac slower, but we don't yet know the protein or gene. Nevertheless, we have a clue. I was at a meeting recently at which David Gadsby gave an account of how the Na^+/K^+ exchange pump can be transformed into a channel by digesting part of it off (Artigas & Gadsby 2001). He and I went through the characteristics of this ' Na^+ pump channel', and it almost exactly matches the properties of the background Na^+ channel. I have a hunch that what nature has done is to use this Na^+ pump protein to make a channel by a little bit of deletion that has this property. The complicated answer to your question, Edmund Crampin, is that it unpacks differently in each case, yet it would amaze me if people working in the pharmaceutical world were not sufficiently sophisticated to appreciate that it is the unpacking that gives the insight, and it is this that gives us the leads to potentially important drugs.

Levin: I was struck by your account of the history of model building. It may be worth reflecting that the Hodgkin–Huxley model was published at around the same time as Crick, Watson and others developed their work on DNA. In my opinion there seems to have occurred a default within the world of biology at that moment between molecular biology and physiology. A large number of scientists saw experimental biology progressing largely down the road of molecular biology, while a smaller number were increasingly restricted to experimental physiology (and the domain of modelling). Over the years this

division has been progressively more emphasized. The work in the 1970s on genetic manipulation enhanced and accelerated this process. With the emerging understanding of biological complexity, this process has now come full circle. We are now seeing a convergence, putting back into perspective the relative role of the reductive and integrative sciences. I don't think the question is so much what will it take biologists to get back into modelling — they will be forced to by biology. But instead it is, what will it take for modellers to actually think about biological problems?

A sbburner: Denis Noble, I think you made a very strong case for the utility of failure, but your models may not be a typical example. The fundamental intellectual basis of the modelling that you have done over the last decades hasn't changed. More knowledge has come, but there are in the history of theoretical biology dramatic examples of fundamental failure of 'models' which have no utility at all, because the whole intellectual basis of that modelling was wrong. These failures have cast theoretical biology in a very poor light.

Noble: Let's remember also that there have been spectacular dead ends in experimental work, too.

A sbburner: I see from what you have presented now that modelling undergoes progressive evolution and you are learning from your mistakes.

References

- Brenner S 2001 My life in science. BioMed Central, London
- Guo J, Ono K, Noma A 1995 A sustained inward current activated at the diastolic potential range in rabbit sino-atrial node cells. *J Physiol* 483:1–13
- Kiyosue T, Spindler AJ, Noble SJ, Noble D 1993 Background inward current in ventricular and atrial cells of the guinea-pig. *Proc R Soc Lond B Biol Sci* 252:65–74
- Noble D, Denyer JC, Brown HF, DiFrancesco D 1992 Reciprocal role of the inward currents $i_{b,Na}$ and i_f in controlling and stabilizing pacemaker frequency of rabbit sino-atrial node cells. *Proc R Soc Lond B Biol Sci* 250:199–207

General discussion IV

Noble: I have identified three somewhat interlocking topics that we ought to address during this general discussion. One is the filling of the gap: what is it that the people who are not here would be able to tell us, and in particular the sort of work that Lee Hood is doing (I am going to ask Jeremy Levin and Shankar Subramaniam to comment on this). Second, there is the issue of the acceptability or otherwise of modelling in the biological community, and connected to that, thirdly, the question of training.

Levin: Although I wouldn't want to attempt to represent what Lee Hood or his institute are doing, I would like to draw out some of the essence of this work. Across the world there are many different modelling groups. In Lee's case, the Institute for Systems Biology has brought together a fairly remarkable group of people from diverse backgrounds, including mathematicians, physicists, biologists and talented engineers who build instruments required for high-throughput biology. These people have been brought together to solve a set of particular problems ranging from bacterial metabolism through to innate immunity. If I were to encapsulate the discussions I have had with Lee and members of his team it would be that they understand the requirements for biological computing to be part of an integrative spectrum that extends from bioinformatics through to simulation, and is an essential step to take for biologists. They also understand the interplay of experimental design as a core component in modelling, such that modelling becomes the basis for experimental design. We have had extensive discussions around the importance of iterating between experimental design, developing a particular instrument to measure the specific data that will then be incorporated in the model and that will in turn then test the experiment, creating a better model.

Subramaniam: The way they think about modelling biological phenomena is that they start with molecular-level processes. Then there is the integration mode, which is already entering the systems-level approaches, dealing with the collections and interactions of molecules. The next level is modelling the network in terms of equations of motions, with standard physical equations. The question is, how do we bridge these different levels? There are four ways of doing it that are generally used, and Lee Hood's is one of these. The first way is to ask the following question. All this molecular level interaction is data modelling, so how do we incorporate this in an effective way into equation modelling? This issue is to

some extent unresolved. Having said this, there are three approaches people take to solve these kinds of things. One, taken by some of the chemotaxis people, is to use control theory level modelling: they create a network, ask a question and carry out sensitivity analysis of this dynamical network. Can we model such a network using simple equations of motion? The second approach is the one Greg Stephanopolous and Bernhard Palsson do. They are chemical engineers and they use flux balance modelling. It is easy to do this in metabolic processes: you start with a metabolite which gets successively degraded in different forms. You can ask the question, can I use simple conservation loss of the overall concentrations to combine coupled concentration equations and solve a matrix? This may give solutions that will narrow the space down and tell us what are the spatial solutions under which a cell can operate. This is what Bernhard Palsson talks about with regard to genotype–phenotype relationships: he can say that one of the spaces is restricted by using this choice of conditions. For that you need to know all the reactions of the cell. It is only good for linear networks. The third level of modelling deals with kinetic modelling, which is that once you know all the reactions you can piece it all together into kinetic schemes and model it in a similar way to what Les Loew does. You can fit this into an overall kinetic equation network pathway model. This is more explanatory at this point than predictive. What Lee Hood’s group wants to do is to combine all these different approaches, but his main focus is the following, and this is illustrated by his one publication which deals with galactose pathway modelling. The galactose pathway modelling idea is very interesting, because it tries to combine the data modelling (experimental data from expression profile analysis) along with a pathway model which is obtained by taking all these different nodes in a pathway and seeing what combinatorics you can get with the constraints of the experiment. This is an element that is very important, and is missing today in biology: how do we take experimental data and use it to constrain the models at a physical level? This is to a large extent what Lee would like to do with mammalian cells. The moment we talk about cell signalling it is no longer possible.

McCulloch: Presumably this is because you can’t invoke conservation of mass to constrain the solution space.

Subramaniam: Let me summarize all of this. We currently have high-throughput data coming from chemical analysis, reaction networks and cellular analysis. How can we use these high-throughput data as constraints in equations in a model? Is each case going to be its own case, in which case it becomes a task for every modeller to do their own thing? Or are there general principles that are emerging? How can we incorporate the use of high-throughput data as constraints in this modelling? Is this also going to be generalizable at some level, or is it going to be specific to each problem? This is a fundamental issue that Jeremy Levin and I wanted to bring up for discussion.

McCulloch: I'd like to add a question here. If you use the analogy with flux balance analysis and/or energy balance analysis approaches to metabolic pathway modelling, there are two features that have been employed. One is the use of physical constraints to narrow the solution space. This still leaves an infinite number of solutions. The way that Bernhard Palsson, for example, has been able to find particular solutions is by invoking an optimization criterion, usually that of maximizing growth. I have a question: is it a worthwhile endeavour to search for equivalent optimality criteria in signalling pathways? Is this a search for a theoretical biology, or does it not exist?

Subramaniam: We have a partial answer to that question which we have not tested extensively. In a metabolic pathway case, you have a starting point and an end point, and these are the constraints. The intermediate constraints are metabolite concentrations. In signalling, you don't have such a thing. What you really have is signal flux, which bifurcates and branches out. It is the enzymes such as kinases which phosphorylate things that are often intermediate constraints: this is a conserved concentration of either phosphorylated or unphosphorylated states. If you talk about protein-protein interactions, there is the interacting state and the non-interacting state. These are local constraints in terms of going through this chain of flux of the signal. We should be able to use these local constraints to do similar types of constrained modelling, and optimize the ultimate phenotype, which in this case would be the end point of the signalling, such as transcription factor initiation.

Winslow: One additional reason that it is important is because the kinds of biophysically based models that we are all constructing are now so complex that our ability to build them from the bottom up is becoming very limited. It is hard to add a new component to a complex model and ensure that all the data that the model was based on originally are still being described well by the new model. It is difficult to know how precisely to adjust parameters to bring that new model into accordance with all of the ever-increasing body of data. What we need (this is an easy thing to say) is an understanding of how nature self-assembles these systems. This may mean that we need to understand the optimality principles: the cost functions that are being minimized by nature.

Cassman: One way of addressing this is to look for functional motifs within models, such as amplifiers and transducers. For example, Jim Ferrall has shown that the phosphorylation cascade is not an amplification mechanism, but actually an on-off switch, and works as a hysteresis module to go up very sharply and then remain on and return to 'off' only very slowly. Perhaps this is one way to put together modules. We could build models by trying to identify the operating components: the switches, transducers, amplifiers and so on. These may be conserved within biological systems.

Levin: It would be surprising if they weren't.

Asburner: I'm very surprised to hear that you can't use optimization in modelling signal transduction.

Subramaniam: I am not saying you cannot do optimization. You can do this, but you do not know the local constraints. There are infinite solutions which will give you the same optimum endpoint.

Asburner: Telecommunications engineers have been working hard for a long time to find out how best to optimize getting the signal from one end of the world to the other.

Subramaniam: Absolutely, but that is easier to do because there are standard components.

Hunter: I want to make the point that in thinking about models generally (and in thinking about people in the modelling community who are not represented here), and relating back to the comment that one of the great attractions of reaction–diffusion models is that they apply in many areas, we don't want to lose sight of the fact that there are many classes of modelling that we haven't considered here, yet are relevant to human physiology. I would list things like soft tissue mechanics, fluid flow, circulation, issues of transport generally, electromagnetic modelling and optics. There is a whole class of models that haven't arisen in our discussions.

Paterson: One way to look at what is optimal for signalling is to talk about what function a particular cellular component is playing in the role of the entire organism. In much of my organization's work we don't know the identity of all the proteins that characterize the input–output relationships for different cells, and that participate in different systems. For example, in the work that we have done in diabetes, the number of constraints that we have by starting and looking at clinical data, simply to make the whole body metabolism stable under different levels of exercise and food intake, are quite powerful. We may not currently understand every protein interaction for every signal transduction cascade, but the only way to make the system stable and reproduce a variety of different clinical data that perturb the system in very orthogonal ways, is for us to characterize the *in vivo* envelope of that part of the system. I think there are some powerful ways to impose those constraints by means of the context. Whether or not anyone has tried to figure out how to do an optimization around that is another question.

Noble: You could call those top–down constraints. Incidentally, you could think of what I described rather colourfully as the pact that evolution has made with the devil in terms of cardiac repolarization as a lovely optimization problem. What has happened there is that it has gone for optimizing energy consumption — you can have as long an action potential as you need with minimal energy consumption — and presumably in the balance someone dropping dead at the age of 55 after a squash game is a small price to pay for the rest of humanity having all of that energy saved!

Hinch: Linking back to what Raimond Winslow was saying about when you add an additional thing to a model worrying about what it does to the previous data, it is useful to consider the work of John Reinitz in Stonybrook on patterning in *Drosophila*. They have a model based on a gene network, with six or seven genes and loads of interactions: the model has some 60 parameters. They really don't have good idea about many of the parameters, but they have a very good, large data set. They use numerical techniques to fit all the parameters to all their data in one go. By doing this they come up with some interesting things about the topology of the network. This is something that I have felt may be worth doing with cardiac cells: getting all the data together and then using one of these numerical techniques to piece all the parameters together in one go.

Shimizu: There must be an upper limit to how many parameters you could use.

Hinch: It would probably depend on the quality and the type of data. They are doing about 60 parameters, and it takes a large cluster of computers quite a long time to do it. In principle, it can be done but you are correct in suggesting that scalability is a potential problem.

Hunter: There are people looking at this issue, such as Socrates Dokos in Sydney, who is looking at parameter optimization for connecting cardiac models to measured current–voltage data and action potential data. It is needed.

Noble: I wonder whether we could now focus on the issues of the acceptability of modelling and the training of people who could operate in this area. I threw up a challenge to the mathematicians and engineers to think a bit about this.

Cassman: Earlier, Jeremy Levin put the onus on the modellers to develop mechanisms to be able to make them accessible to the biologists. I would go the other way round, frankly. You mentioned that 99.9% of the biologists don't do modelling, and I think there are several reasons for this, and these are serious barriers that have to be overcome. One is that for a long time much of biology has been an area for people who want to do science without mathematics. These are not people who are going to readily accept mathematical models, because they are afraid of maths. The second is that a mindset has developed as a consequence of the success of molecular genetics that regards single-gene defects as the primary paradigm for the way one thinks about biology. This means that thinking about networks is going to require some degree of retraining. People just aren't conditioned to do that. Frankly, I think the answer is not that dissimilar from the way it is in much of science: you have to wait for people to die before a paradigm changes! The students are very interested and are anxious to get into programs that will give them both the biology and the mathematics. The real question is, what do you expect as an endpoint? Do you expect people to be able to do both themselves, with intensive training in both biology and maths? Or do you want people who can communicate, and this is good enough? I don't think it's either/or. There will be both, although relatively fewer of the first type. We need to be able to design

programs that will allow people to make that change. I think that most of the more senior current practitioners will have a hard time making that transition.

Levin: This is a good point. At its core, the process requires inducing students who are learning molecular biology to appreciate the role mathematics has as an integral component of their science. We run the training programs for molecular biologists, and it has been striking how extraordinarily responsive these young people are to modelling and using computers to extend their thinking. They are able to enter into an environment they have never seen before, adopt a mode of thinking very rapidly over a period of one or two months, and adopt modelling. I would say that it is incumbent on us as modellers, in as much as we are training biologists, to start thinking about the tools we use to train people. It is absolutely impossible for one of the new students to go back and become a mathematician. What they want to be able to do is to have intuitive tools which by virtue of, for example, drawing a simple diagram (a 'visual model'), automatically generate a mathematical model that they can then populate with their data. From this they can design experiments and create and test hypotheses. I find myself deeply impressed with the rate at which the young people pick this up, given the right tools.

Asburner: Marvin Cassman, part of your diagnosis is wrong. What I would call the aristocratic period of molecular genetics probably ended in 1974 and was succeeded by the demotic period dominated by cloning and sequencing; the elucidation of the life cycle of λ is nothing if it is not a model. Go back to Jacob and Monod: this was done by classical genetic analysis, but the famous 1961 paper was a model (Jacob & Monod 1961).

Cassman: Let's talk about quantitative models as opposed to descriptive models. If you are talking about quantitative models, that of Changeux, Monod and Wyman was a quantitative theory, because Wyman was around, not because of Monod and Changeux. There are other examples, but they are relatively limited in molecular genetics.

Winslow: Marvin Cassman asked the question about how we should educate students in biology. It is critical that biologists should know mathematics and/or engineering principles. It has to be part of their education. Modelling is not going to advance if the process of modelling is turn key. If you go to a system and you don't know the mathematics and numerical methods behind it, and simply use that system, just as Denis Noble has pointed out that models are inherent failures, so the numerical methods embodied in the simulators are also subject to inherent failures. People who use them need to understand this. The problem is that there is so much to teach a student, so how do we balance it? We face this every day in our biomedical engineering programs. The way we approach it is to teach biology in the context of engineering. Biology is presented to students via an engineering and mathematics framework.

Cassman: For what it is worth, the National Academy of Sciences in the USA is completing a study on a revision of curricula in biology at the undergraduate and graduate levels. They are going to recommend a heavy dose of mathematics, starting early. It is a little hard to bring people into the graduate level and to expect them to swallow all of that, but they need to be started at a relatively early stage.

Noble: That's when I learned my mathematics, but that is ancient history!

Cassman: It is not that you can't do it, just that it is much harder then.

Boissel: Don't you think there are two different objectives that should be achieved through training? One is for the biologists to use formal modelling. The other is to have more specialized people able to develop new models. For the first, we need to enable biologists to communicate with engineers; for the second we need people who are able to do both biology and maths.

Subramaniam: The education issue is significant. There are two institutions—Caltech and MIT—which have started a compulsory biology course for undergraduate engineers. Similarly, for biology students there is currently no equivalent requirement for mathematics. Having said this, some of the problems start at the high school level: many biology undergraduates will not have studied maths at a calculus level. They automatically preselect themselves to go into biology because they say there is no maths in biology. I also want to comment on Raimond Winslow's point. It is not sufficient if you just train biologists in terms of learning calculus or differential equations, for the following reason. One of the fundamental principles the engineers learn is how to coarse grain a system: how to model a system to get the required level of sophistication to compare with the experiment. This is not something that comes from just learning maths. It comes from engineering approaches towards systems. The third issue deals with quantitatively orientated people learning biology. One of the problems we have encountered with physicists, chemists and mathematicians is they do not care for the 'devil is in the details' biology. If you don't care for the details, you only contribute superficially. One of the first things we need to do is to make sure that the students who come to us learn gory details in biology. At least then they can do the coarse graining principles at some level. The fourth thing that needs to be addressed is that it is not sufficient to know maths or biology. You also need to know scientific computing and data management. This is very different from just learning maths. We have students who know very sophisticated maths and biology, but who don't know anything about computing. This is a serious problem, because today all the modelling is done with real data. To deal with this we need to understand data structures. It may be worthwhile integrating people who make games for children and students to make games involving cells. Instead of warriors shooting each other's heads off, you could have T cells and B cells fighting infections!

Paterson: It has been our experience that precisely because it takes years of training to do mathematics and biology really well, there is tremendous value in keeping those skill sets in two separate heads. The checks and balances you get by providing a team environment where those people communicate efficiently capitalizes on the strengths.

McCulloch: That's what people said 15 years ago about bioengineering as an undergraduate discipline. Medical device companies started off hiring specialists, but now bioengineering graduates are employed for those positions. It is not that the specialists don't still exist in these companies, but having people with the information in one head has proven extremely valuable.

Levin: Tom Paterson has a good point. There is a distinction between creating an educational environment and creating an organization that is designed to deliver a product every two months. There is an element of what Tom is describing that is temporal: we don't yet have the expertise that provides for organizations such as Tom's and mine the kind of people who can deliver on a model straight out of the graduate programs. We are forced to recruit the experts and create teams. An optimal team in our case is three-headed: a superb mathematician, a superb engineer and a biologist who really understands what they are doing. They can grow on either side of that. One thing that is essential, underlying this, is that they all understand the language.

Loew: Marvin Cassman mentioned the older generation of biologists, and I agree that this is an issue. On the other hand, from a practical point of view, they are also influencing the younger scientists. We shouldn't be giving up on them. From the point of view of the science of biology, there are many of these scientists and they are very good scientists. What I have found is that the good biology investigator is highly focused on his problem and will be motivated to learn a new technique such as modelling if he or she is convinced that it will advance their work. We need to be ambassadors for modelling in terms of reaching out to our colleagues.

Noble: I have had a recent and beautiful experience of that nature, attending a meeting that was addressing the question of whether the $\text{Na}^+/\text{Ca}^{2+}$ exchanger was or was not a good guy in relation to cardiac ischaemia. A set of experimentalists were presenting data that I was able to show, using simulation, were inexplicable. I demonstrated that the level of Ca^{2+} they had recorded during ischaemia was not possible to understand with their data. We retired and sat down at the computer, trying to work what else was going on that could explain this phenomenon. We ran a simulation that got the required result (Noble 2002). This is a team that has absolutely no modelling experience, but what they have now asked for is to have hands on the model. They want to play with it. Once it is demonstrated that not only can you point out that something isn't understood, but you can then interact with a set of experimentalists to determine what would be needed in order to

resolve the issue, then you have people queuing up to get further understanding. This comes back to the point I emphasized earlier on. The 'hands on' is necessary.

References

- Jacob F, Monod J 1961 Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol* 3:318–356
- Noble D 2002 Simulation of Na–Ca exchange activity during ischaemia. *Ann NY Acad Sci*, in press

The IUPS Physiome Project

P. J. Hunter, P.M.F. Nielsen and D. Bullivant

Bioengineering Institute, University of Auckland, Private Bag 92019, Auckland, New Zealand

Abstract. Modern medicine is currently benefiting from the development of new genomic and proteomic techniques, and also from the development of ever more sophisticated clinical imaging devices. This will mean that the clinical assessment of a patient's medical condition could, in the near future, include information from both diagnostic imaging and DNA profile or protein expression data. The Physiome Project of the International Union of Physiological Sciences (IUPS) is attempting to provide a comprehensive framework for modelling the human body using computational methods which can incorporate the biochemistry, biophysics and anatomy of cells, tissues and organs. A major goal of the project is to use computational modelling to analyse integrative biological function in terms of underlying structure and molecular mechanisms. To support that goal the project is establishing web-accessible physiological databases dealing with model-related data, including bibliographic information, at the cell, tissue, organ and organ system levels. This paper discusses the development of comprehensive integrative mathematical models of human physiology based on patient-specific quantitative descriptions of anatomical structures and models of biophysical processes which reach down to the genetic level.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 207–221

Physiology has always been concerned with the integrative function of cells, organs and whole organisms. However, as reductionist biomedical science succeeds in elucidating ever more detail at the molecular level, it is increasingly difficult for physiologists to relate integrated whole organ function to underlying biophysically detailed mechanisms. Understanding a re-entrant arrhythmia in the heart, for example, depends on knowledge of not only numerous cellular ionic current mechanisms and signal transduction pathways, but also larger scale myocardial tissue structure and the spatial distribution of ion channel and gap junction densities.

The only means of coping with this explosion in complexity is mathematical modelling—a situation very familiar to engineers and physicists who have long based their design and analysis of complex systems on computer models. Biological systems, however, are vastly more complex than human engineered systems and understanding them will require specially designed software and instrumentation

and an unprecedented degree of both international and interdisciplinary collaboration.

Furthermore, modern medicine is currently benefiting both from the development of new genomic and proteomic techniques, based on our recently discovered knowledge of protein-encoding sequences in the human genome, and from the development of ever more sophisticated clinical imaging devices (MRI, NMR, micro-CT, ultrasound imaging, electrical field imaging, optical tomography, etc.). This will mean that the clinical assessment of a patient's medical condition could, in the near future, include information from both diagnostic imaging and DNA profile or protein expression data. To relate these two ends of the spectrum, however, will require very comprehensive integrative mathematical models of human physiology based on patient-specific quantitative descriptions of anatomical structures and models of biophysical processes which reach down to the genetic level.

The term 'Physiome Project' means, somewhat loosely, the combination of worldwide efforts to develop databases and models which facilitate the understanding of the integrative function of cells, organs and organisms. It was launched in 1997 by the International Union of Physiological Sciences (see <http://www.physiome.org>). The project aims both to reach down through subcellular modelling to the molecular level and the database generated by the genome project, and to build up through whole organ and whole body modelling to clinical knowledge and applications. The initial goals include both organ specific modelling such as the Cardiome Project (driven partly by a collaboration between Oxford University, UK, the University of Auckland, NZ, the University of California at San Diego and Physiome Sciences Inc, but also involving contributions by many other cardiac research groups around the world) and distributed systems such as the Microcirculation Physiome Project (led by Professor Popel at Johns Hopkins University; <http://www.bme.jhu.edu/news/microphys/>).

The Physiome markup languages

An important aspect of the Physiome Project is the development of standards and tools for handling web-accessible data and models. The goal is to have all relevant models and their parameters available on the web in a way which allows the models to be downloaded and run with easy user-editing of parameters and good visualization of results. By storing models in a machine and application independent form it will become possible to automatically generate computer code implementations of the models and to provide web facilities for validating new code. The most appropriate choice for web based data storage would appear to be the newly approved XML standard (eXtensible Markup Language—see

<http://www.w3c.org/>). XML files contain tags identifying the names, values and other related information of model parameters whose type is declared in associated DTD (Data Type Definition) files. XQL (XML Query Language) is a set of tools designed to issue queries to database search engines to extract relevant information from XML documents (which can reside anywhere on the world wide web). The display of information in web browsers is controlled by XSL (XML Style Language) files. Two groups are currently developing an XML for cell modelling. One group, based at Caltech, is developing SBML (Systems Biology Markup Language) as a language for representing biochemical networks such as cell signalling pathways, metabolic pathways and biochemical reactions (<http://www.cds.caltech.edu/erato/>), and a joint effort by the University of Auckland and Physiome Sciences is developing CellML with an initial focus on models of electrophysiology, mechanics, energetics and signal transduction pathway models (<http://www.cellml.org/>). The CellML and SBML development teams are now working together to achieve a single common standard.

The Auckland group is also developing 'FieldML' to encapsulate the spatial and temporal variation of parameters in continuum (or 'field') models, and 'AnatML' as a markup language for anatomical data (see <http://www.physiome.org.nz/>). When all the pertinent issues for each area have been addressed it may be appropriate to coalesce all three markup languages into one more general Physiome markup language since the need for a standardized description of spatially varying parameters at the organ level is equally important within the cell for models of cellular processes.

The hierarchy of models

A major objective of the Physiome Project is to develop mathematical models which link gene, protein, cell, tissue, organ and whole body systems physiology into one comprehensive framework. Models are currently being developed at many levels in this hierarchy, including

- whole body system models
- whole body continuum models
- tissue and whole organ continuum models
- subcellular ordinary differential equation (ODE) models
- subcellular Markov models
- molecular models
- gene network models

An important issue is how to relate the parameters of a model at one spatial scale to the biophysical detail captured in the model at the level below.

The computational models used in the Physiome Project are largely ‘anatomically based’. That is, they attempt to capture the real geometry and structure of an organ in a mathematical form which can be used together with the cell and tissue properties to solve the physical laws which govern the behaviour of the organ such as the electrical current flow, oxygen transport, mechanical deformation and other physical processes underlying function. Wherever possible the models are also ‘biophysically based’, meaning that the equations used to describe the material properties at both cell and tissue level either directly contain descriptions of the biophysical processes governing those properties or are derived from such descriptions in a computationally tractable form. One important consequence of an anatomically and biophysically based modelling approach is that as more and more detail is added (such as the spatial distribution of ion channel expression) the greater complexity often leads to fewer rather than more free parameters in the models because the number of constraints increases. Another important point is that the governing tissue-level equations represent physical conservation laws that must be obeyed by any material — e.g. conservation of electrical current (Faraday’s law) or conservation of mass and momentum (Newton’s laws). The models are therefore predictive and represent much more than just a summary of experimental data.

The question of how much detail to include in a model is one that all mathematical modellers have to deal with, irrespective of the field of application. If added detail includes more free parameters (model parameters which can be altered to force the model to match observed behaviour at the integrative level) the answer — in keeping with the principle of Occam’s Razor — must be ‘as little as possible’. On the other hand, detail added in the form of anatomical structure and validated biophysical relationships can often constrain possible solutions and therefore enhance physiological relevance. It is surprisingly easy, for example, to create a model of ventricular fibrillation with over-simplified representations of cell electrophysiology. Adding more biophysical detail in the form of membrane ion channels reduces the arrhythmogenic vulnerability to more realistic levels.

A brief summary of the various types of model used in computational physiology is given here in order to highlight the major challenges and the immediate requirements for the Physiome Project.

Tissue mechanics

The equations come from the physical laws of mass conservation and momentum conservation in three dimensions and require a knowledge of the tissue structure and material (constitutive) properties, together with a mathematical characterization of the anatomy and fibrous structure of the organ (or bone, etc.). Solution of the equations gives the deformation, strain and stress distributions

throughout the organ. Examples are the large deformation soft-tissue mechanics of the heart, lungs, skeletal muscles and cartilage, and the small strain mechanics of bones. The mathematical techniques required for these problems are now well established and the main challenge is to define the geometry of all body parts and the spatial variation of tissue structure and material properties. The most urgent requirements are to define the markup language (FieldML) which allows the anatomy and spatial property variations to be captured in a format for storage and exchange, and to develop the visualization tools for viewing the 3D anatomy and computed fields such as stress and strain. Another high priority is to enhance the tools that allow a generic model to be customized to individual patient data from medical imaging devices such as MRI, CAT and ultrasound.

Fluid mechanics

The equations are also based on mass conservation and momentum or energy conservation and the requirement for a mathematical representation of anatomy is similar, but now the constitutive equations come from the rheology of a fluid (e.g. blood or air) and the solution of the equations yields a pressure and flow field. Obvious examples are blood flow in arteries and veins, and gas flow in the lungs. In some cases the equations can be integrated over a vessel or airway cross-section to reduce the problem to the solution of 1D equations, while in others a full 3D solution is required. The top priorities in this area are as above—the markup languages, visualization tools and patient customization tools.

Reaction–diffusion systems

There are many issues of transport by diffusion and advection, coupled to biochemical reactions, in physiological systems. The transport equations are based on well established laws of flux conservation, and the numerical solution strategies are also well developed. Examples are the electrical activation of the heart (equations based on conservation of current) and numerous problems in developmental biology. The need for good anatomical descriptions using FieldML is similar to the above two categories. The main challenges lie in developing good models of the biochemical reactions and capturing these in the CellML format for storage and exchange.

Electrophysiology

All cells make use of ion channels, pumps and exchangers. The mathematical description of the ion channel conductance and voltage (or ion) dependent gating rate parameters is usually based on the Hodgkin–Huxley formalism

(typically using voltage clamp data) or more molecularly-based stochastic models (with patch clamp data). Examples are the Hodgkin–Huxley models of action potential propagation in nerve axons, the Noble and Rudy models for cardiac cell electrophysiology and pancreatic β -cell models of the metabolic dependence of insulin release. The major challenge now is to relate the parameters of these models to our rapidly increasing knowledge of gene sequence and 3D structure for these membrane-bound proteins, together with tissue specific ion channel densities (and isoforms) and known mutations. The CellML markup language is currently being extended to link into FieldML for handling the spatially varying parameters such as channel density. The most urgent requirements are authoring tools, application programming interfaces (APIs) and simulation tools.

Signal transduction and metabolic pathways

The governing equations here are based on mass balance relations. The information content is often based on signal dynamics rather than steady-state properties, so a system dynamics and control theoretical framework is important. An example is the eukaryotic mitogen-activated protein kinase (MAPK) signalling pathway which culminates with activation of extracellular signal-regulated kinases (ERKs). The signal transduction pathway definitions can be encapsulated in CellML and a priority now is the development of tools which will allow the activity of the pathways to be modelled in the context of a 3D cell and linked to ion channel and pumps (e.g. as sites of phosphorylation), and to tissue and organ level models.

Gene networks

This relates to the study of gene regulation, where proteins often regulate their own production or that of other proteins in a complex web of interactions. The biochemistry of the feedback loops in protein–DNA interactions often leads to non-linear equations. Techniques from non-linear dynamics, control theory and molecular biology are used to develop dynamic models of gene regulatory networks.

It should be emphasized that no one model could possibly cover the 10^9 dynamic range of spatial scales (from the 1 nm pore size of an ion channel to the 1 m scale of the human body) or 10^{15} dynamic range of temporal scales (from the $1\mu\text{s}$ typical of Brownian motion to the 70 years or 10^9 s typical of a human lifetime). Rather, it requires a hierarchy of models, such that the parameters of one model in the hierarchy can be understood in terms of the physics or chemistry of the model appropriate to the spatial or temporal scale at the level below. This hierarchy of models must range from gene networks, signal transduction pathways and

stochastic models of single channels at the fine scale, up to systems of ODEs, representing cell level function, and partial differential equations, representing the continuum properties of tissues and organs, at the coarse scale.

Modelling software and databases

There are now a number of cell and organ modelling programs freely available for academic use:

- *PathwayPrism and CardioPrism* (<http://www.physiome.com>) provide access to databases as well as cell modelling and data analysis tools
- *E-Cell* (<http://www.e-cell.org/>) is a modelling and simulation environment for biochemical and genetic processes
- *VCell* (<http://www.nrcam.ucbc.edu/>) is a general framework for the spatial modelling and simulation of cellular physiology
- CMISS is the modelling software package developed by the Bioengineering Research group at the University of Auckland (see <http://www.bioeng.auckland.ac.nz/cmss/cmss.php>)
- CONTINUITY from the Cardiac Bioengineering group at UCSD is a finite element based package targeted primarily at the heart (see <http://cmrg.ucsd.edu>)
- BioPSE from the Scientific and Computing Institute (SCI) deals primarily with bioelectric problems (<http://www.sci.utah.edu>)
- CardioWave from the Biomedical Engineering Department at Duke University is designed for electrical activation of myocardial tissue (<http://bme-www.egr.duke.edu/>).
- XSIM models the transport and exchange of solutes and water in the microvasculature (<http://nsr.bioeng.washington.edu>).

Physiome projects

Several Physiome projects are mentioned briefly here. Figure 1 illustrates the sequence of measuring geometric data for the femur and fitting a finite element model (Fig. 1A,B), incorporating the femur model into a whole skeleton model (Fig. 1C) and then combining with the muscles of the leg (Fig. 1D) for analysis of loads in the knee. Figure 2 illustrates a model of the torso (Bradley et al 1997), including the heart and lungs and the layers of skin, fat and skeletal muscle, which is being used for studying the forward and inverse problems of electrocardiology and for developing the lung physiome. Figure 3 illustrates the fibrous structure, coronary network and epicardial textures in a model of the heart (LeGrice et al 1997, Smith et al 2000, Kohl et al 2000).

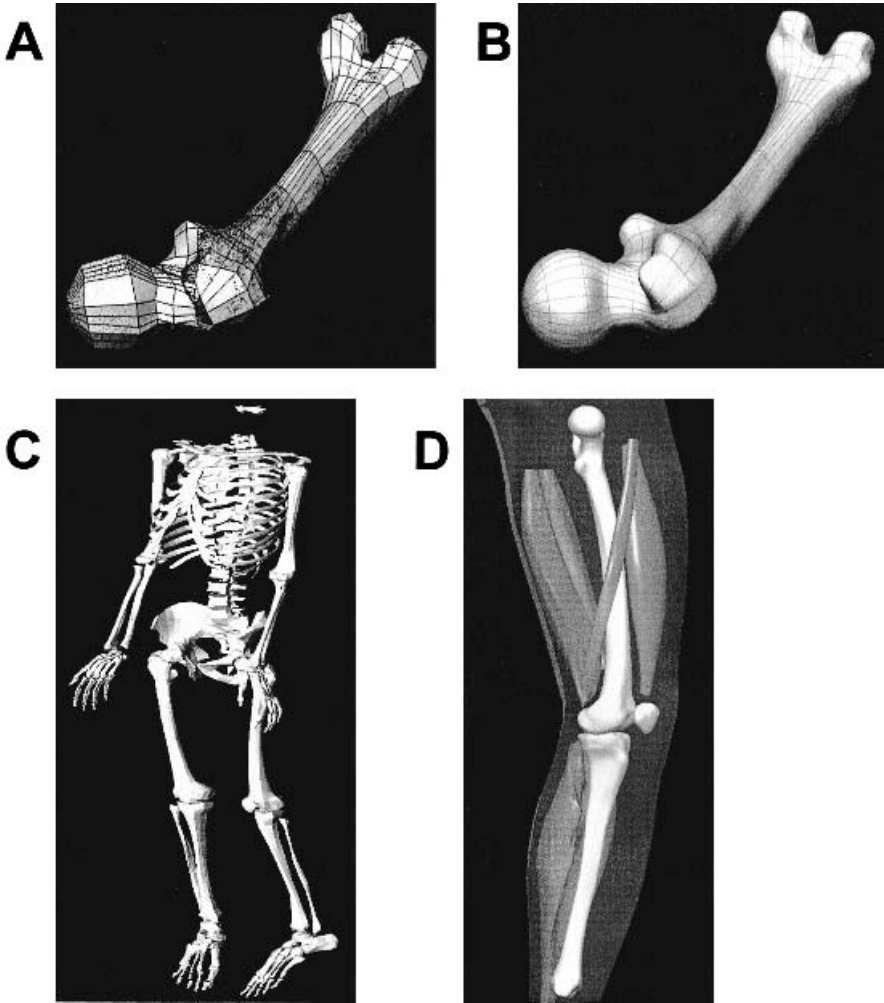


FIG. 1. (A) A finite element mesh of the femur prior to fitting, together with a cloud of data points measured from a bone with a laser scanner, and (B) the same (bicubic Hermite) mesh after fitting the nodal parameters. (C) Anatomically detailed model of the skeleton. (D) Rendered finite element mesh shown for the bones of the leg and a subset of the muscles (sartorius, rectus femoris and biceps femoris in upper leg and gastrocnemius and soleus in lower leg). The musculo-skeletal models contain descriptions of 3D geometry and material properties and are used in computing stress distributions under mechanical loads.

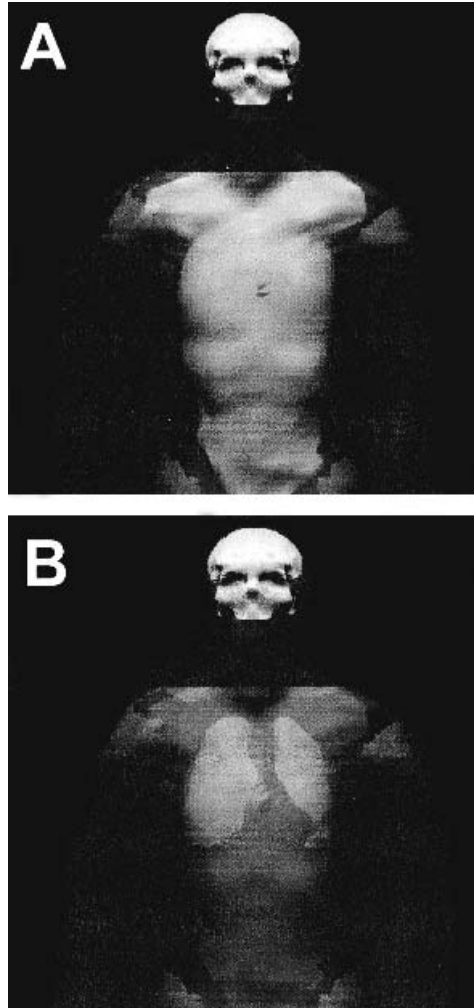


FIG. 2. Computational model of the skull and torso. (A) The layer of skeletal muscle is highlighted. (B) The heart and lungs shown within the torso.

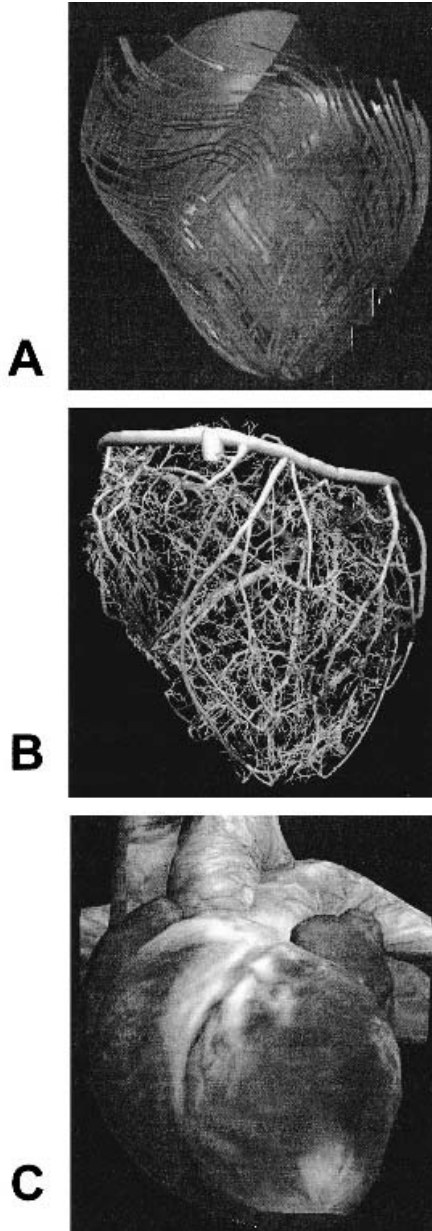


FIG. 3. The heart model. (A) Ribbons showing the fibrous-sheet architecture of the heart are drawn in the plane of the myocardial sheets on the epicardial surface of the heart. (B) Computed flow in the coronary vasculature. (C) The heart model with textured epicardial surface.

Acknowledgements

The Auckland work discussed and illustrated in this paper is the result of the collaborative efforts of many past and present members of the University of Auckland Bioengineering Research group. We are grateful for funding from the University of Auckland and Auckland UniServices Ltd, the NZ Foundation for Research, Science and Technology, the NZ Heart Foundation, the NZ Health Research Council, the Wellcome Trust, Physiome Sciences Inc, Princeton, and LifeFX Inc, Boston. PJH would also like to acknowledge gratefully the support of the Royal Society of NZ for the award of a James Cook Fellowship from July 1999 to July 2001.

References

- Bradley CP, Pullan AJ, Hunter PJ 1997 Geometric modeling of the human torso using cubic hermite elements. *Ann Biomed Eng* 25:96–111
- Kohl P, Noble D, Winslow RL, Hunter PJ 2000 Computational modelling of biological systems: tools and visions. *Philos Trans R Soc Lond A Math Phys Sci* 358:579–610
- LeGrice IJ, Hunter PJ, Smail BH 1997 Laminar structure of the heart: a mathematical model. *Am J Physiol* 272:H2466–H2476
- Smith NP, Pullan AJ, Hunter PJ 2000 Generation of an anatomically based geometric coronary model. *Ann Biomed Eng* 28:14–25

DISCUSSION

Subramaniam: I have a naïve question. In your mechanical model involving the cube you talk about cell walls deforming. What is the frequency at which this happens, and how does it relate to gene expression changes, cell and morphology changes, and what is the feedback mechanism between those things?

Hunter: The time-scale is that of a heartbeat for the deformation that you are looking at.

Subramaniam: So in that time-scale you don't have gene expression, transcription and regulation occurring. I'm curious to know what the long-term consequence is, and how this feeds back into the fibrillation?

Hunter: I'd love to know that. We are still dealing with the time-scales in the order of a heartbeat. We are looking at electrophysiology with Denis Noble and we are looking at cell signalling as this comes out of the Alliance for Cell Signaling, but all of this is on the time-scale of a heartbeat at the moment. It would be very nice to then look at the longer time-scale of minutes to hours to days to see gene expression changes, but this is for the future.

Noble: The way we tackle that particular problem is to run simulations at the cell and tissue level that may go on for many tens of minutes. Then we take snapshots of the states in those simulations. By snapshots, I mean that many of the variables that were parts of the differential equations in the lower-level modelling are frozen, or their vectors are frozen. This is then inserted into 2D or 3D simulations at the tissue or organ level, hoping that we can validly claim that the development of the tissue

states up to this point hasn't been terribly badly perturbed by the fact that the tissue is part of an organ. This is a huge assumption, I agree. One of the people from my group who I feel would have been able to contribute to this meeting enormously is Peter Kohl (see Kohl & Sachs 2001). He deals with the question of the feedback between the whole organ mechanical changes and the electrophysiology. This turns out to be extremely important, particularly for some of the arrhythmias that are known to be mechanically induced. The issues you are highlighting are very important.

McCulloch: There have been a few studies where physiological consequences of signalling events can be seen within the time-scale of a single beat.

Subramaniam: That is at the proteomic level, not gene expression.

Berridge: The same thing applies to the nervous system during memory acquisition. Memory has to be consolidated by gene transcription. It seems that what happens in the brain is that this access to gene transcription occurs during sleep. A temporary modification of the synapses during memory acquisition is then consolidated during slow-wave sleep when gene activation occurs. The brain appears to go offline to carry out all the genetic processes responsible for consolidation. The amazing thing about the heart is that it has to go on pumping, and creating Ca^{2+} pulses while it carries out its genetic changes.

Noble: Could I turn now to the question of cell types. To be provocative, it is possible to take the view of the cardiac conducting system that you were proposing — that there is one cell type, but with different levels of expression for various protein transporters — to an extreme, and say that there is just one cell type. Why not?

Hunter: There are two extremes. That is one, and the other is that there are 10^{15} cell types. The reality is somewhere in between; it is just a question of where we put the demarcation.

Noble: Why does it matter, then? Presumably it matters for the reason that we discussed right at the beginning of this meeting, which is that what you call something does actually matter. Presumably, it will matter from the point of view of the way in which you organize the database of information.

Hunter: I'm thinking of it mattering in terms of modelling, where we want to make sure that we are pulling in all the appropriate functional behaviour of that cell type. It may well be that you go to your CellML file for an electrically active cell from the heart, but then you input the parameter set that is appropriate for the different positions though the conducting system, just as even within myocytes you would need that appropriate difference between M cells and other cells. You have to acknowledge the different expression levels for different types of cells as a function of spatial position. But you certainly don't want to regard each different spatial position as giving rise to a different cell type. There is no one answer; it is simply a pragmatic issue of getting access to information for modelling.

Asburner: I would go further and say that if you are going through this exercise for modelling, it is worth doing in such a way that this classification can be used by others. For example, those who are merely looking at gene expression and protein expression have no interest whatsoever in modelling *per se*.

This may be based on a misconception, but let me put to you the critique that with AnatML and CellML you are confounding the ontology itself and its representation. What you really need is ontology. How you represent that ontology is an independent operation.

Hunter: I accept that. Ontologies are currently being considered in conjunction with the CellML schemas.

Asburner: I am not arguing with that. My problem is that you have wrapped it up in a particular flavour of XML with your own tags.

Subramaniam: I am a consultant for the NIMH database for neuroanatomy-based functional imaging. The neuroanatomy project looks at four brains: mouse, rat, human and primate. There are similar kinds of complications in that one of the first things they are going to do is define clearly the ontology. Once this is defined representation becomes a critical issue here. You cannot say, for example, that a particular region of the brain is going to be exactly the same even across two members of the same species, so you need to map it into a feature space and then use the feature space to define the actual ontology of that object or the element that is being defined. This is exactly what they are proceeding with. On top of that they are having a structure which uses geographical information systems (GIS) to help map this feature space efficiently into the ontology. Doug Bowden has created a beautiful atlas which deals with primate brains (<http://braininfo.rprc.washington.edu/brainatlas.html>) and does exactly the same things that you are talking about. Once you have defined the ontology and have a mapping system within the ontology it is actually a little bit more complicated than just a straightforward database. A flat file system will never do this feature mapping coupled with the definition of an ontology.

Hunter: The reason for a flat file is that you may want to get that information from an entirely different set of relationships. You may want to be looking at a particular cell in the brain across species, or across age. There are all sorts of ways that you may want to access information. If you confine it to a particular tree-like GIS-type structure, you are in danger of limiting access to that information in another way.

Subramaniam: Not really. The caveat here is that some of your representation problems and feature mapping may depend upon relationships between different objects within your ontology. If you use a flat file you lose the flexibility of doing this.

Hunter: I'm suggesting that we have the information in a flat file and we also have the relationships — the ontologies — that allow us to access that.

Subramaniam: But it doesn't scale. When you start scaling to higher levels if you are doing it this way, there are so many microcomputations and calculations in order to do this mapping that pretty soon it becomes an explosively complicated process.

McCulloch: The point that we need the ontology first is key, but with anatomy all the way to cell type there already exists an ontology. This has been done at the University of Washington in a project directed by Cornelius Rosse that expresses anatomic relationships in the form of a directed acyclic graph.

Hunter: Is this in a way that is relevant to modelling?

McCulloch: Certainly in a way that is more relevant to modelling than the index structure of textbooks.

Subramaniam: And it is hierarchical.

Asburner: I have one for *Drosophila* (<http://fly.ebi.ac.uk:7081/docs/lk/bodyparts-cv.txt>).

Noble: There is often discussion, particularly in the media, about the question of whether we are in reach of a virtual human. I usually answer that question in the negative. Yet when I hear your presentation, and watch all the structures that are already in some way or another coded into the mesh, I am left wondering whether I ought not to be more positive. This is a strategic issue, among other things, because it affects the way in which funding agencies see what we are trying to do. This isn't a trivial question, which is why I treat it quite carefully in discussions with the media.

McCulloch: My answer would be that what we see emerging from Peter Hunter's work is a virtual body.

Hunter: I think what will emerge over a relatively short time frame is the description of the anatomy and the material properties relevant to the larger scale continuum problems. But there is a huge gap between gene expression and the tissue or organ-level models. I wouldn't for one moment suggest that we are anywhere near beginning to tackle the complexity of that issue. It is only at the top level that I see things coming together reasonably fast.

Noble: So you are creating the outer mesh.

Hunter: Yes, into which we want to put all the cell types with increasing information about signal transduction systems and so on.

Paterson: One thing that might characterize the transition from having the virtual body to the virtual human is an increased understanding of all the different interacting control systems that allow the 'meat on the bone' to be maintained. As an example, we have worked on epithelial turnover. All the dermatology texts seem to take a standard bricks-and-mortar histological view of the skin. However, when you look at the control systems that are necessary to maintain normal turnover of skin as well as injury repair, there are a huge number of unanswered questions masked by simply giving a picture saying that

it is static when there is actually a large degree of activity from multiple feedback systems that keep this 'static' view stable.

Reference

Kohl P, Sachs F 2001 Mechano-electric feedback in cardiac cells. *Phil Trans Roy Soc A* 359:1173–1185

Using *in silico* biology to facilitate drug development

Jeremy M. Levin, R. Christian Penland, Andrew T. Stamps and Carolyn R. Cho

Physiome Sciences, 307 College Road East, Princeton, NJ 08540-6608, USA

Abstract. G protein-coupled receptor (GPCR) mediation of cardiac excitability is often overlooked in predicting the likelihood that a compound will alter repolarization. While the areas of GPCR signal transduction and electrophysiology are rich in data, experiments combining the two are difficult. *In silico* modelling facilitates the integration of all relevant data in both areas to explore the hypothesis that critical associations may exist between the different GPCR signalling mechanisms and cardiac excitability and repolarization. An example of this linkage is suggested by the observation that a mutation of the gene encoding HERG, the pore-forming subunit of the rapidly activating delayed rectifier K^+ current (I_{Kr}), leads to a form of long QT syndrome in which affected individuals are vulnerable to stress-induced arrhythmia following β -adrenergic stimulation. Using Physiome's In Silico Cell™, we constructed a model integrating the signalling mechanisms of second messengers cAMP and protein kinase A with I_{Kr} in a cardiac myocyte. We analysed the model to identify the second messengers that most strongly influence I_{Kr} behaviour. Our conclusions indicate that the dynamics of regulation are multifactorial, and that Physiome's approach to *in silico* modelling helps elucidate the subtle control mechanisms at play.

2002 'In silico' simulation of biological processes. Wiley, Chichester (Novartis Foundation Symposium 247) p 222–243

Previously in this symposium we have discussed many of the tools of *in silico* biology. For my presentation I will concentrate on one particular aspect of *in silico* biology, building and simulating mathematical models: why model and how to model. I will specifically focus on the role of modelling in the pharmaceutical industry, then dive down to a more granular level and use a case example to examine how we answered a very specific question related to a problem in the pharmaceutical industry. This example will demonstrate why modelling is an advantageous approach. It will also serve to show how a model is constructed — what data are required and how the components are joined. The question that I will try to address throughout the talk, is how can we use modelling and simulation to serve the biological research industry in its goal of identifying control mechanisms that are important for drug discovery?

What is *in silico* modelling in the context of drug discovery? This question is a very different one from those previously discussed in this symposium. *In silico* technologies are complex and interrelated, and they appear everywhere in drug discovery today. They range from molecular structure and docking simulation, mathematical modelling, bioinformatics, high-throughput data gathering and processing, three-dimensional imaging, pathway mapping and network analyses, through to system modelling which includes intelligent decision systems and expert system diagnosis of disease. Importantly, all these technologies complement wet-lab experimentation; we cannot divorce experimentation from modelling. Over the last 20 years we have seen an increased emphasis on the process of data-driven drug discovery. In a philosophical context, this result is a reflection of the complexity of biology and the effort to develop an increasingly deep, but reductionist, understanding of this biology. The result is that we have amassed a body of biological data overwhelming in its complexity and volume. This drives a critical need for new approaches to interpret and extract insight from the data derived from complex biological systems. Many informal modelling methods are designed to interpret data, such as gedanken experiments, drawing cartoon diagrams, developing word or phenomenological models, and so forth. We use mathematics to translate these conceptual models into logically rigorous representations. These models are then used to generate hypotheses that can then be experimentally tested, yielding more data, which in turn are used to refine the original model. Any of the steps in this process may lead to novel biological insight.

We are now moving towards what I believe to be an important change in drug discovery: hypothesis-driven, as compared to data-driven, drug discovery. This is made possible because new technologies for biological modelling enable drug discovery through the exploration of hypotheses *in silico*. This new approach allows integration of diverse types of data as well as re-use of legacy data. Given the large amount of data generated in the industry over the past few decades, the critical issue is how to build and apply the new methodologies of *in silico* biology to address the increasingly complex questions that new high-throughput tools and data sources allow us to pose. The scale of this problem becomes apparent when we examine the choices companies today face with their current programs. For example, companies that may have over 200 pre-clinical drug programmes, yet can only afford to test 40 of those in the clinic, face a very important economic question: which 40 of these drugs are going to work when failures could potentially cost hundreds of millions of dollars for each program? *In silico* biology provides the capability to address this important process of programme selection in a rational and predictive manner by coupling the experiments to hypotheses, efficiently exploring parameter space of experimental variables, and permitting direct comparisons and predicting outcomes.

Physiome technology approach

Before presenting a case example of where *in silico* biology technology can be applied, I would like to talk about the technology itself. I think it is critical for the drug development community to standardize the processes that underlie the technology, such as building, storing and communicating mathematical models, and developing visualization and analysis tools. What is really required? At its core, an open information technology architecture that permits global collaboration is essential. In addition, intuitive software that allows the scientist to organize, view and analyse data, as well as build and simulate models. This software needs to be developed with the plan that it becomes a tool in the hands of the scientists at the bench, not necessarily the modelling specialist, while retaining the functionality required to correctly communicate the details and analysis of the model including annotation, literature references and underlying mathematics. Most importantly, there is a need for the technology to make use of all forms of data, including the reuse of legacy data as well as capturing data from new sources.

New data generation technologies are driving the adoption of *in silico* biological modelling. Biological modelling can be applied to the full spectrum of observable biological phenomena, capable of dealing with data on gene and protein expression all the way through to disease maps and simulations. The approach that we have adopted is to develop a biological simulation environment called In Silico Cell™.

Within this environment we can integrate all the data necessary for modelling of both specific, and broad biological questions. We use this environment to build models, run simulations, and analyse simulation and experimental data. Most importantly, our technology is specifically designed for placement within a pharmaceutical company. The purpose here is to enable the development of *in silico* biological modelling as a core competency within drug discovery groups. In addition, we help companies build models themselves, evaluate their data using our own in-house capabilities, and as a result we are now involved with a number of different companies that are taking the lead in introducing this form of technology among multiple sites around the globe. These companies are either using the completed models developed and customized by us, such as the cellular, tissue and organ cardiac models in our CardioPrism™ program, or the metabolic and signal pathway analysis capability afforded by PathwayPrism™, both of which derive directly from In Silico Cell™. It is not necessary to have all possible data in order to build an effective and utilitarian model, capable of answering important questions for the pharmaceutical scientist. Our process can bring together many different forms of data, all of which are directly applicable to the particular experiment being performed. The data are constrained from the beginning of the modelling process. We do not attempt to integrate all data without a rationale: we

constrain ourselves to the problem and the data that are available. We then look at the available data to see if we have missing pieces, and if so, parameter estimations are performed. On the first build of a model, we have always been able to make use of purely legacy data. Additional data generated from testing the model can then be used to refine the model through iterative steps of experimental and *in silico* hypothesis testing. In order to accommodate the changing model and new data, the modelling environment is developed to be flexible and extensible, so permitting the incorporation of changes with minimal effort.

The process begins by generating a mathematical description of the biological question, and then works systematically through to prediction and hypothesis. The process may suggest new experiments be done, providing new data, which then generate a new biological question and lead to a reiteration of the whole process. There are different ways to address each step in this process, many of which we have discussed in this symposium. No matter what the specific approaches are, the important point is that novel insight may be gained throughout the process, whether it be developing the model, formulating new hypotheses, or analysing the new experimental data that are generated.

What makes our process fundamentally flexible and extensible is our approach to the process of building models (Fig. 1); we identify currently known biological mechanisms beginning with those most commonly and widely observed. We build (or reuse) a model of each of these mechanisms, which we call a motif or module. The categories of these motifs, at the cell and subcellular level, are metabolism, signalling, excitability, transport and cell cycle. Each of these motifs represents mechanisms underlying such fundamental biological functions as glycolysis, translocation and motility. The data supporting each of these motifs may be separated from the model itself and replaced with data relating to another cell type, species and so forth, and modules may be combined so that we can, for example, use clinical parameters to model a variety of diseases such as rheumatoid arthritis, asthma, and osteoporosis. Each module we create can then be reused to build a model in another disease area, so that we minimize the 'reinvention of the wheel'. This concept raises technology implementation issues of how we store our models and data, which I would be happy to discuss after this presentation.

One application of our technology and modelling approach that I would like to highlight is that it can be applied to summarizing and leveraging data within and between research groups of pharmaceutical companies. Our PathwayPrism™ technology illustrates this issue very clearly (Fig. 2). Using such an application, different groups within a pharmaceutical company can create and/or explore different pathways that are internal to their own group and not seen by others. They can then merge these pathways using our technology to form a composite pathway that shares data, annotations, stored simulation data and so forth. The example shown here is the tumour necrosis factor (TNF) pathway, which is a

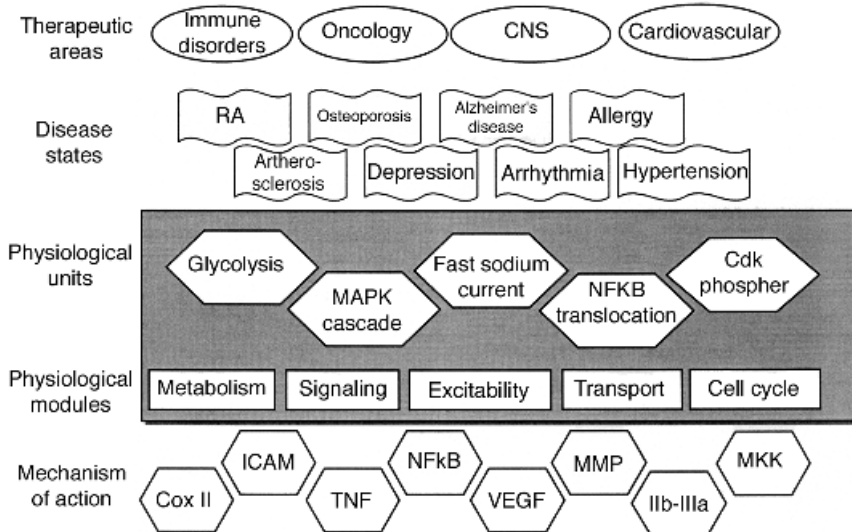


FIG. 1. Modelling motifs. The process of model building reuses mathematical descriptions of individual biological processes. These processes, shown in the figure as 'physiological units', give rise to such fundamental biological motifs as signalling, excitability, and transport, which are indicated as 'physiological units'. Each of these units (e.g. fast sodium current) can be part of a motif (e.g. excitability), which is a widely observed phenomenon in physiological systems. The designation of motifs allows one to describe the critical physiological units of models which can facilitate an understanding between mechanism of action of a drug and the disease state.

merge of many smaller pathways with which we are working internally. This capability provides people with a tool to represent, explore, and understand their combined data in an intuitive graphical format. Moreover, from a drug development point of view, one avenue of exploration (illustrated in Fig. 2) is to compare the behaviours of the many drugs that impact this one pathway. For us, as a modelling company focused on helping pharmaceutical companies find better products, this capability is critical.

In addition to the technology to build pathways, we have developed an analogous technology to build whole cell models. We use this tool to model, for example, cardiac action potentials similar to those of Winslow et al (1999) and others (Luo & Rudy 1994a,b, Noble et al 1998). We have a very different aim than these other groups from a practical point of view. Rather than ever further refining the physiological mechanisms in such myocyte models, we seek to understand the avenues by which pharmaceutical compounds interact with the cells in both beneficial and harmful ways. We accomplish this goal by integrating the modelling with a laboratory equipped to study ion channels and electrophysiology. We also incorporate drug regulatory expertise to understand

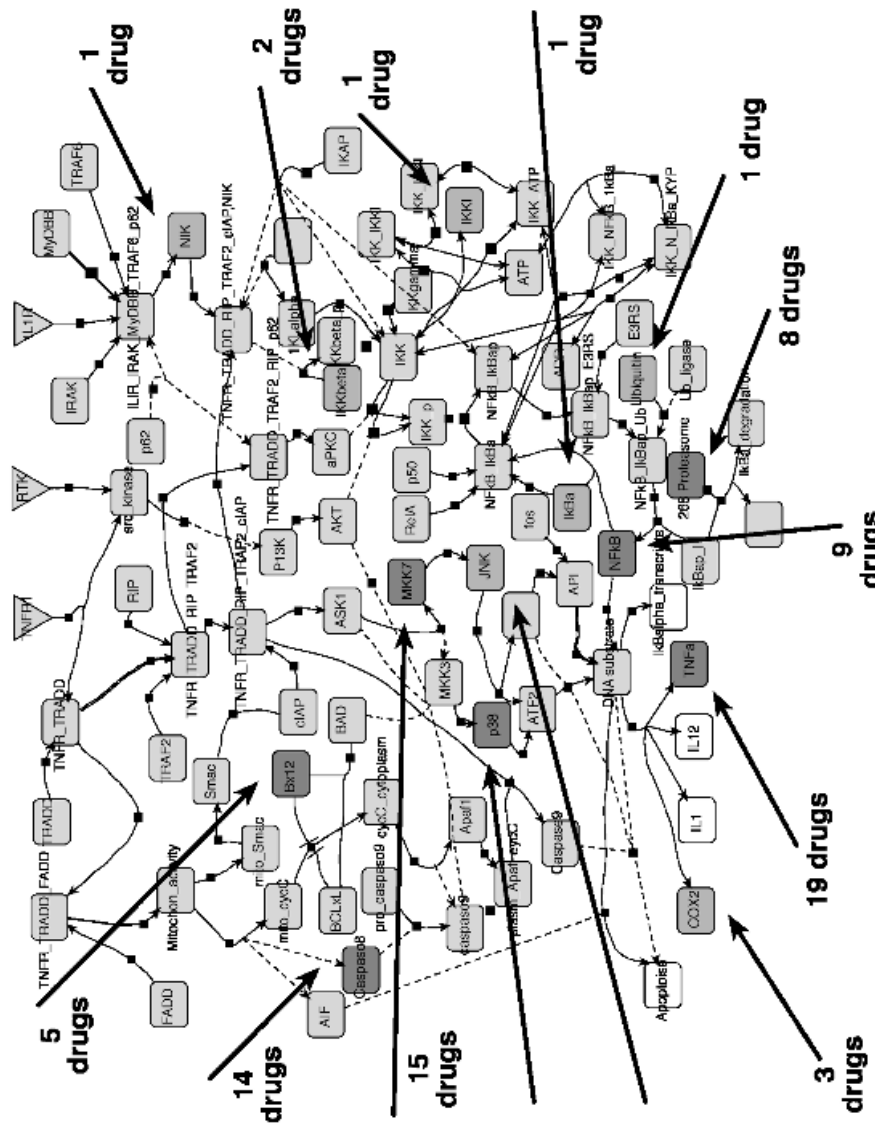


FIG. 2. Signalling pathway in PathwayPrism™. This screen shot from PathwayPrism™ shows some of the pathways involved in TNF signalling, and places where marketed drugs are targeted to intervene. The example shown demonstrates the results of a merge from a number of smaller signalling pathways.

and provide insight where these 10 models fit into the US Food & Drug Administration (FDA) process required to develop a drug.

There are many examples that testify to the value of modelling in the discovery and development process. One area of interest is in preventing unnecessary deaths from cardiac arrhythmias. Though there are many different applications of models in cardiovascular safety, a case study that we often point to is that of the antiarrhythmic d-sotalol, which blocks the rapid component of the delayed rectifier current (I_{Kr}). Tested in 1996 via the SWORD (survivability with oral d-sotalol) trial (Pratt et al 1998), d-sotalol was administered prophylactically to patients surviving myocardial infarctions in the hope that it would reduce their mortality from subsequent arrhythmic episodes. Unfortunately, mortality increased with d-sotalol administration vs. placebo, and surprisingly, women were found to be at much greater risk of death than men. The unanswered question was why?

We constructed a series of canine ventricular myocyte models corresponding to the three different cell types across the ventricular wall (epicardial, endocardial and M cell), and incorporated modifications accounting for data showing ventricular myocytes from female rabbits having 15% less I_{Kr} density and 13% less I_{K1} density compared to those from male rabbits. With no drug onboard, the simulated M cell action potential from the female was only slightly different from that of the male. As drug concentration is increased both male and female action potentials prolong, however only a 50% blockage in I_{Kr} is required to begin to observe early after depolarizations (EADs) in the female action potential, while 80% I_{Kr} block is required to see the same effect in male cells (Fig. 3). This result indicates a threefold differential in the male/female susceptibility to this drug. The reduction in repolarizing currents expressed in females thus makes them more sensitive to action potential abnormalities induced by I_{Kr} block. Though no specific type of arrhythmia was cited in the SWORD trial as leading to mortality, EADs are commonly viewed as a marker for arrhythmic susceptibility. Therefore, our modelling results suggested a possible cause for the gender difference in mortality.

I want now to turn to the issue of integrating data to investigate the significance of individual components in a complex system. The following will illustrate how modelling can make logical inferences from available data to make testable predictions. These predictions provide evidence as to the underlying mechanisms, which is particularly useful when the underlying mechanisms cannot be addressed by current experimental techniques.

Case example: indirect signalling in cardiac excitability

I previously mentioned that leveraging prior efforts is one of the powerful aspects of our approach to modelling. Having discussed two separate Physiome

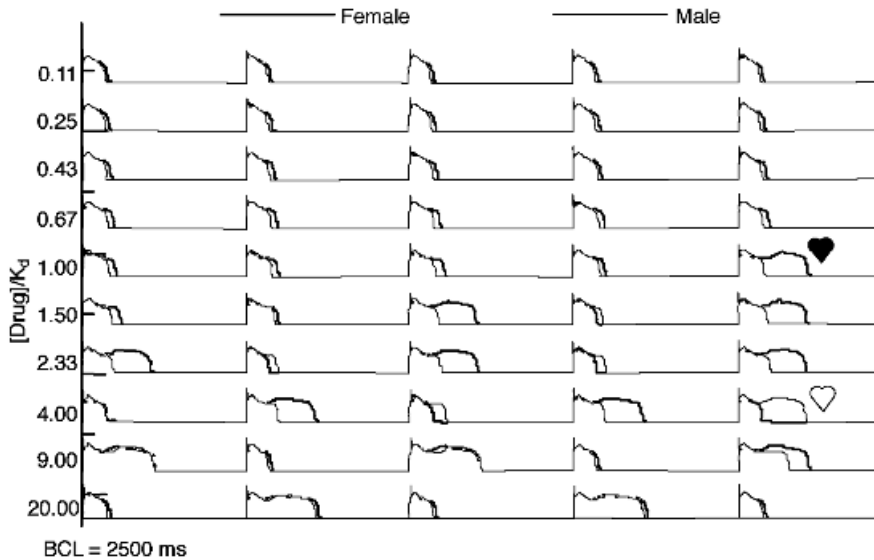


FIG. 3. Simulation of male and female canine M cell action potentials in the presence of a drug that blocks the I_{K_r} channel. As drug concentration increases (top to bottom), an early after depolarization (EAD) occurs at a lower drug concentration for the female than for the male cell, which is indicated by the small heart symbol above the first EAD for each gender. These EADs are thought to be a trigger for drug-induced arrhythmia. The basic cycle length (interval between pacing stimuli) was 2500 ms.

technologies representing two distinct scientific areas, signal transduction and electrophysiology, I want to present a case example that brings together these two diverse areas. This example demonstrates Physiome Sciences' ability to integrate models from both a biological perspective as well as a software implementation perspective. We have joined together two very distinct areas of experimental research using our technology platform to couple separate models into a single simulation of second messenger control of ion channel current. This work was performed by a team of scientists at Physiome, in addition to the authors, including Dr Adam Muzikant, Director of the Modeling Sciences Group, and Ms Neelofur Wasti, in the same group, who provides data and literature support and curation.

Drugs indirectly affect the heart

In the case of d-sotalol, the compound was in fact an antiarrhythmic targeted directly at the I_{K_r} channel to prolong the action potential. A more difficult problem to analyse is that of drugs that affect ion channels of the heart despite

not being targeted specifically to them. More than 60% of all drugs target G protein-coupled receptors (GPCRs). A drug that targets a CNS GPCR, for example, could have severe cardiotoxicity that would not be necessarily be identified in present screening protocols, which are designed to assess direct drug-channel interaction, mostly for I_{Kr} .

Toxicological concerns involving the most common form of drug related cardiac rhythm concern, QT prolongation, are a frequent cause of clinical holds, non-approvals, approval delays, withdrawals and restricted labelling by the FDA. In fact, QT prolongation was a factor in many such actions taken by the FDA since the late 1990s, and continues to form a major hurdle in bringing new drugs to market, regardless of therapeutic class. The regulatory focus on QT prolongation as a toxicological concern derives from its role as a surrogate marker for altered cardiac cell repolarization, and risk of Torsades de Pointes, a life-threatening arrhythmia.

All known drugs that appear to induce cardiac arrhythmia associated with long QT preferentially block I_{Kr} , hence pharmaceutical companies routinely evaluate a compound's QT prolongation risk preclinically by screening for its effect on the HERG channel, the pore-forming subunit of I_{Kr} . Current best practices in preclinical cardiac safety assessment include using voltage clamps in expression systems transfected with HERG; *in vitro* action potential measurements using isolated myocytes, and *in vivo* telemetered electrocardiograms from intact animals. However, these best practices occasionally fail to identify drugs with a high risk of inducing cardiac arrhythmia. For example, grepafloxacin weakly blocks I_{Kr} but has been observed to induce Torsades de Pointes, leading to its withdrawal from the market by Glaxo-Wellcome in 1999. Conversely, these practices may be overly harsh in assessing drugs like verapamil, which despite blocking I_{Kr} and causing QT prolongation is not associated with arrhythmia. To understand this issue better, we must take a closer look at the relationship between arrhythmia and I_{Kr} . According to Shimizu & Antzelevitch (1999), diminished I_{Kr} leads to arrhythmia by preferentially prolonging the action potential in ventricular M cells. This repolarization change leads not only to a cellular substrate with increased dispersion of refractoriness that is vulnerable to arrhythmia, but also to increased incidence of EADs that may trigger such arrhythmias. In contrast blocking I_{Ks} , the slowly activated delayed rectifier K^+ current, more uniformly prolongs the action potential throughout the ventricle, and is not associated with life-threatening arrhythmias.

There are many factors that accentuate the effect of blocking I_{Kr} including decreased heart rate, gender and genetic susceptibility, and though no single factor may greatly alter the action potential their combination may significantly increase the risk of drug-induced arrhythmia. Transmembrane voltage, electrolyte balance, and direct drug-channel binding principally regulate I_{Kr} by

itself. Mutations in channel proteins can dramatically impact the gating of the channel, while drugs that stimulate a second messenger cascade can indirectly regulate the channel. Though poorly understood at present, the second messenger-mediated effects on ion channels like I_{Kr} are gaining increasing attention.

The indirect effects we are concerned about are triggered by cell surface receptors. Specifically, we concentrated on GPCR stimulation because the majority of prescription drugs act via this family. There is a rich literature of experimental data that describes the biochemical pathways that define the second messenger signal transduction pathways. A separate, equally rich literature provides the electrophysiological characterization of HERG, which is often studied in expression systems as a surrogate for the native channel (Trudeau et al 1995). However, experimental approaches to studying the combined second messenger control of ion channel current are difficult. In native cell environments, it is difficult to both control second messenger activation *and* isolate ion channels. In expression systems, it is difficult to ensure that the necessary elements of the native cell signalling system are reconstructed correctly.

These considerations provide an excellent opportunity for modelling. Modelling approaches have been used extensively to study the kinetics of G protein signalling (Bos 2001, Davare et al 2001, Dalhase et al 1999, Destexhe & Sejnowski 1995, Kenakin 2002, Moller et al 2001, Tang & Othmer 1994, 1995); they have also been used extensively to study ion channel currents (Clancy & Rudy 2001, Zeng et al 1995, Winslow et al 1999, Luo & Rudy 1994a,b, Noble et al 1998). Although combining these models does pose a challenge, in a relatively short amount of time we were able to use existing techniques to make predictions about the behaviour of the combined system.

Integrating signalling and electrophysiology motifs

There are a limited amount of data available on direct second messenger regulation of HERG though some investigators have identified cAMP and protein kinase A (PKA) as key players (Cui et al 2000, 2001, Kiehn et al 1998, 1999). From our library of GPCR signalling templates, we selected the cAMP-PKA regulation motif and customized it with available data. Cui et al (2000) showed that PKA phosphorylation of HERG renders the channel less likely to open, but that cAMP also directly binds HERG to counterbalance the PKA effect and lower the activation voltage of the channel ($V_{1/2}$, see Equation 1.3, below). In addition, it is well known that cAMP activates PKA. We therefore described the well-characterized activation kinetics of the second messengers using the standard ordinary differential equation representation of the mass action kinetics.

We formulated the I_{Kr} dependence on voltage and second messengers from previous model-based and experimental studies (Zeng et al 1995, Cui et al 2000). Using a combination of directly applying a membrane-soluble cAMP analogue and mutating the PKA-sensitive phosphorylation sites of HERG, investigators reached three conclusions that were used in our model: (1) channel conductance is regulated by PKA alone; (2) both cAMP and PKA coordinately regulated the strength of channel response to voltage (m , the slope of the voltage-sensitive activation at half-maximal response); and (3) PKA and cAMP independently regulate channel activation in response to voltage ($V_{1/2}$). Based on these observations, we used their reported single-channel current measurements at varying levels of cAMP and PKA to generate the relationship between $V_{1/2}$ and PKA, $V_{1/2}$ and cAMP, m as a function of both PKA and cAMP, and the dependence of conductance on PKA (Equation 1):

$$I_{Kr}(V, cAMP, PKA^*) = [g_{Kr}(PKA^*)][X_{Kr}(V, cAMP, PKA^*)][R(V)][V - E_K] \quad (1)$$

The gating variable X_{Kr} is governed by

$$\frac{dX_{Kr}}{dt} = \frac{X_{\infty} - X_{Kr}}{\tau} \quad (1.1)$$

where

$$X_{\infty}(V, cAMP, PKA^*) = \left\{ 1 + \exp \left[\frac{-V_{1/2} - V}{m} \right] \right\}^{-1} \quad (1.2)$$

and

$$V_{1/2} = \Delta V_{1/2, baseline} + \Delta V_{1/2}(cAMP) + \Delta V_{1/2}(PKA^*). \quad (1.3)$$

We combined our signalling and ion channel models automatically using internally developed software. The environment accepts all the required kinetic and electrophysiological data as well as the mathematical descriptions, and implements fast differential equation solvers to generate predictions from the model.

Predicting ion channel behaviour

Sensitivity analysis. I will briefly present some preliminary predictions from model analysis. The first thing we did was a sensitivity analysis, to predict the relative strengths of the two second messengers as regulators of ion channel current. Of

the several parameters that describe the gating and conductance regulation, we examined the parameters generated from fitting dose-response data to the conductance (g_{K_r} , Equation 1), to the strength of channel response to voltage (m , Equation 1.2), and to the shift parameters describing $V_{1/2}$ (Equation 1.3). Because the system was linear, to a reasonable approximation, a perturbation analysis was performed to compare how the ‘baseline’ behaviour of the model changes in response to changes in parameter values. We used several different baseline behaviours corresponding to the experimental conditions where ‘wild-type’ versus ‘phosphorylation-mutant HERG’ conditions were combined with and without stimulation by cAMP.

We observed that changes in any of the cAMP parameters caused less than a 1% change in ion channel current, while the PKA-dependent strength of channel response to voltage was responsible for more than 75% of the current variation. Thus we predicted that I_{K_r} is most strongly affected by the PKA-controlled gating, independent of cAMP activity. This result suggests that the nucleotide-binding domain of HERG is not as important for its regulation as the PKA-dependent phosphorylation sites.

The implications for a pharmaceutical company are quite significant. First if one were to screen a compound library for new I_{K_r} blockers, these predictions suggest that looking for compounds that control voltage gating would yield more effective candidates than simply screening for compounds that bind the HERG subunit of I_{K_r} . Secondly, in the arena of cardiotoxicology, if you are going to develop a safety screen for a drug, doing a HERG screen may not identify all potentially toxic compounds, and it may in fact eliminate safe compounds. Our results suggest, in fact, that toxicological screens can be developed to assess indirect drug effects by measuring activation of second messengers.

Action potential generation. It may be that second messenger activation is not an available measurement. A common electrophysiological measurement is the action potential from a whole cell. We used a whole cell model of guinea-pig ventricular myocyte (Luo & Rudy 1994b) to report out the predicted action potential, given a predicted I_{K_r} current, to predict the whole cell effects of second messenger regulation of HERG. Figure 4 shows simulated action potentials with no stimulation, PKA stimulation alone, cAMP alone and combined stimulation. The model predicts that cAMP-induced shift in activation potential has only a small effect on the action potential, while activating PKA independently delays repolarization by 5%. The cooperative contribution of cAMP increases this delay slightly.

The experimental difficulty in isolating the effect of PKA stimulation from that of cAMP precludes the possibility that this prediction could be made easily without the use of modelling. This prediction of action potential behaviour illustrates that

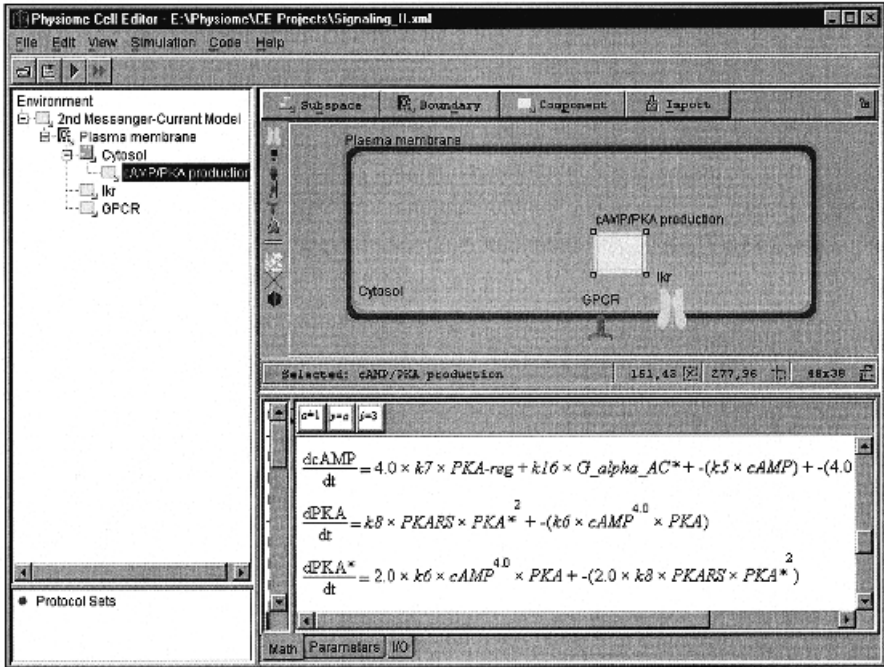
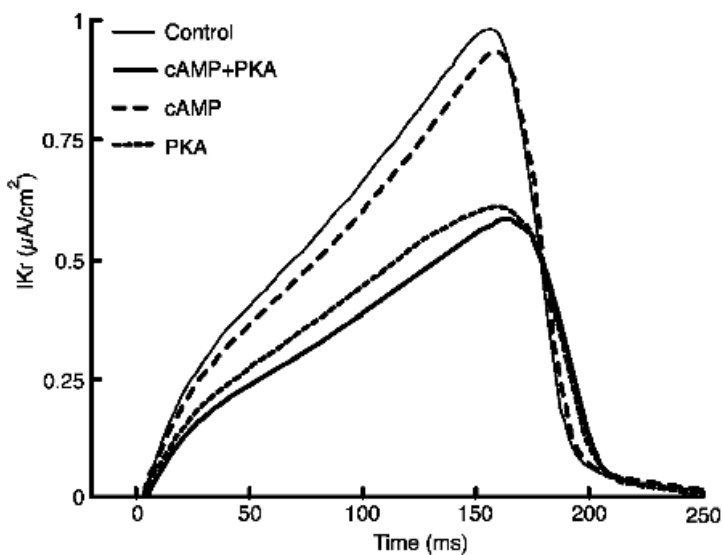
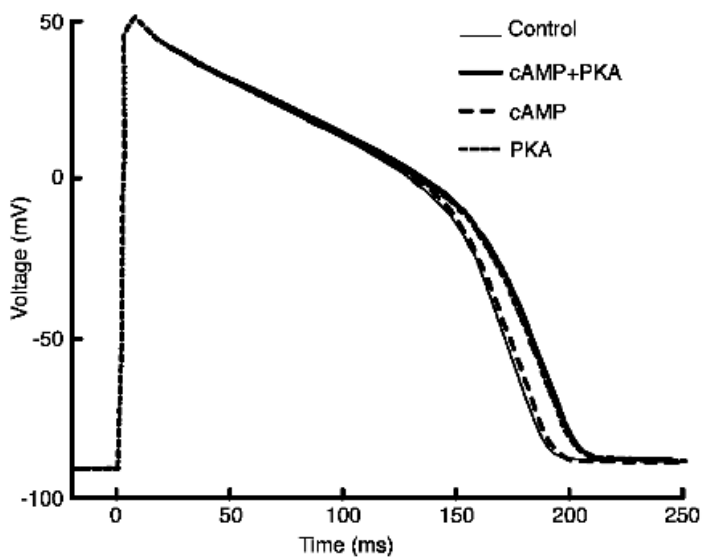


FIG. 4. Merged electrophysiology and signal transduction model in In Silico CellTM software. This screenshot shows how the ion channel and concentrations of second messengers can be represented both graphically (top right pane) and mathematically (lower right pane).

although our model was focused on a single ion channel, we were still able to make some prediction about whole cell behaviour. This finding is important, as stated above, because it provides predictions about a commonly measured indicator of cardiac cell behaviour.

There are a few aspects that I would like to summarize. Although a 5% delay in repolarization is relatively small, it is profoundly important. Firstly, this independent effect of PKA would not otherwise have been predicted, which is quite remarkable. Secondly, this 5% delay is predicted to arise from second

FIG. 5. (*Opposite*) Simulation of second messenger control of the I_{Kr} current and guinea-pig ventricular myocyte action potential. (A) The alteration in simulated I_{Kr} current for the three second-messenger cases described in the text, plus control. This I_{Kr} model was then included into a model of the action potential. (B) The simulated action potentials for the same four cases as in Panel A. The effect of cAMP independent of PKA is small, whereas PKA alone or in combination with cAMP causes up to a 5% delay in repolarization.

a IKr decreased by PKA**b** Extended AP duration with PKA stimulation

messenger regulation *alone*. Yet this kinase is just one of many different factors that impact rectifying current. Our system allows you to then build on this result and consider the additional impact of other effectors, including drugs, different receptors, different G proteins, different second messengers and different ion channels. The key message is this: having created the motif of second messenger control of I_{Kr} , we can now reuse it with new or improved parameters to capture new behaviour, without having to expend extra effort in developing extensions of the model from scratch. It may also be extended to other ion channels, to generate a more complete picture of second messenger regulation of cellular electrophysiology. Previous efforts in developing, parameterizing and optimizing models have paved the way for the work that I have shown you here today. This general approach of motifs is one that we have been using with great success at Physiome. I anticipate that we will be seeing future benefits well beyond what has been demonstrated here. We will be developing motifs to encapsulate regulatory control units in signalling, to tackle the biological scalability problem, and to understand the behaviour of whole systems arising from cellular and subcellular level interactions.

Motif-based modelling

Our modelling approach based on physiological motifs is an application of the concept that cellular behaviour such as signal transduction is comprised of groups of interacting molecules (Hartwell et al 1999, Lauffenburger 2000, Rao & Arkin 2001, Asthagiri & Lauffenburger 2000). The same groups of molecules related by similar interactions are observed from behaviour to behaviour. Indeed, we do not always need to know all the molecules to understand the mechanism by which a motif achieves its function. Additionally, in some cases the identity of the molecules may change while the interactions and function of the motif remain constant. This way is ideal for handling the current state of biological knowledge: there is a wide variation in the amount of available data. Motif-based modelling allows the investigator to use a combination of heuristic and mechanistic descriptions to test a hypothesis.

I have presented work on the regulation of HERG by cAMP and PKA. Within a cardiac myocyte, there are additional protein components of I_{Kr} , such as MiRP1 and minK (Nerbonne 2000, Schledermann et al 2001), other ion channels, other second messengers, and other signalling receptors. The combined signal transduction–electrophysiology model used here is easily extensible to these other biological contexts.

The implications for such an approach go well beyond cardiac electrophysiology. We are working in a number of different areas. One is in CNS diseases, where these excitable cell models are directly applicable, and GPCR drug

effects are known to be important. Bladder cells are also electrically excited, and we have been working in that area as well. Downstream second messenger signalling of NF- κ B, for example, is a motif that is found in such areas as immunological and inflammatory responses, and we have been asked to develop models of these signal transduction pathways. My final illustration, here, is cytokine secretion and recognition in initiating immunological response, which we are modelling in T cells.

This one example motif that I have discussed has very wide-ranging implications. Though it was developed in the extremely specific biological context of the cardiac myocyte K⁺ channel, a straightforward reparameterization will allow this motif to be reused in an incredible range of therapeutic areas, from CNS, to gastrointestinal, to oncology to immune disorders. The challenge for us, as for all modellers, I think, is to understand clearly which are the right motifs to develop. In facilitating drug discovery, I have demonstrated here the role of using mathematical modelling to predict indirect drug effects. Beyond this particular example, the model demonstrates how reusing *in silico* biology motifs can extend hypotheses. These motifs are central to our technology approach, to our thinking about biology, and to our application of our technology for use in the pharmaceutical industry.

Acknowledgements

The authors thank A. L. Muzikant, N. M. Wasti, M. McAlister and V. L. Williams for their valuable contributions to the work presented here.

References

- Asthagiri AR, Lauffenburger DA 2000 Bioengineering models of cell signalling. *Annu Rev Biomed Eng* 2:31–53
- Bos JL 2001 Glowing switches. *Nature* 411:1006–1007
- Clancy CE, Rudy Y 2001 Cellular consequences of HERG mutations in the long QT syndrome: precursors to sudden cardiac death. *Cardiovasc Res* 50:301–313
- Cui J, Melman Y, Palma E, Fishman GI, McDonald TV 2000 Cyclic AMP regulates the HERG K⁺ channel by dual pathways. *Curr Biol* 10:671–674
- Cui J, Kagan A, Qin D, Mathew J, Melman YF, McDonald TV 2001 Analysis of the cyclic nucleotide binding domain of the HERG potassium channel and interactions with KCNE2. *J Biol Chem* 276:17244–17251
- Davare MA, Avdonin V, Hall DD et al 2001 A β 2 adrenergic receptor signaling complex assembled with the Ca²⁺ channel Ca_v1.2. *Science* 293:98–101
- Delhase M, Hayakawa M, Chen Y, Karin M 1999 Positive and negative regulation of I κ B kinase activity through IKK β subunit phosphorylation. *Science* 284:309–313
- Destexhe A, Sejnowski TJ 1995 G protein activation kinetics and spillover of γ -aminobutyric acid may account for differences between inhibitory responses in the hippocampus and thalamus. *Proc Natl Acad Sci USA* 92:9515–9519

- Hartwell LH, Hopfield JJ, Leibler S, Murray AW 1999 From molecular to modular cell biology. *Nature* 402:(Suppl)C47–C52
- Kenakin T 2002 Drug efficacy at G protein-coupled receptors. *Annu Rev Pharmacol Toxicol* 42:349–379
- Kiehn J, Karle C, Thomas D, Yao X, Brachmann J, Kubler W 1998 HERG potassium channel activation is shifted by phorbol esters via protein kinase A-dependent pathways. *J Biol Chem* 273:25285–25291
- Kiehn J, Lacerda AE, Brown AM 1999 Pathways of HERG inactivation. *Am J Physiol* 277: H199–H210
- Lauffenburger DA 2000 Cell signaling pathways as control modules: complexity for simplicity? *Proc Natl Acad Sci USA* 97:5031–5033
- Luo CH, Rudy Y 1994a A dynamic model of the cardiac ventricular action potential: I. Simulations of ionic currents and concentration changes. *Circ Res* 74:1071–1096
- Luo CH, Rudy Y 1994b A dynamic model of the cardiac ventricular action potential. II. Afterdepolarizations, triggered activity, and potentiation. *Circ Res* 74:1097–1113
- Moller S, Vilo J, Croning MD 2001 Prediction of the coupling specificity of G protein coupled receptors to their G proteins. *Bioinformatics* 17:S174–S181
- Nerbonne JM 2000 Molecular basis of functional voltage-gated K⁺ channel diversity in the mammalian myocardium. *J Physiol* 525: 285–298
- Noble D, Varghese A, Kohl P, Noble PJ 1998 Improved guinea-pig ventricular cell model incorporating a diadic space, I_{Kr} and I_{Ks} , and length- and tension-dependent processes. *Can J Cardiol* 14:123–134
- Pratt CM, Camm AJ, Cooper W 1998 Mortality in the Survival With ORal D-sotalol (SWORD) trial: why did patients die? *Am J Cardiol* 81:869–876
- Rao CV, Arkin AP 2001 Control motifs for intracellular regulatory networks. *Annu Rev Biomed Eng* 3:391–419
- Schledermann W, Wulfsen I, Schwarz JR, Bauer CK 2001 Modulation of rat *erg1*, *erg2*, *erg3* and HERG K⁺ currents by thyrotropin-releasing hormone in anterior pituitary cells via the native signal cascade. *J Physiol* 532:143–163
- Shimizu W, Antzelevitch C 1999 Cellular basis for long QT, transmural dispersion of repolarization, and Torsade de Pointes in the long QT syndrome. *J Electrocardiol* 32:177–184
- Tang Y, Othmer HG 1994 A G protein-based model of adaptation in *Dictyostelium discoideum*. *Math Biosci* 120:25–76
- Tang Y, Othmer HG 1995 Excitation, oscillations and wave propagation in a G protein-based model of signal transduction in *Dictyostelium discoideum*. *Philos Trans R Soc Lond B Biol Sci* 349:179–195
- Trudeau MC, Warmke JW, Ganetzky B, Robertson GA 1995 HERG, a human inward rectifier in the voltage-gated potassium channel family. *Science* 269:92–95
- Winslow RL, Rice JJ, Jafri MS, Marban E, O'Rourke B 1999 Mechanisms of altered excitation–contraction coupling in canine tachycardia-induced heart failure. II. Model studies. *Circ Res* 84:571–586
- Zeng J, Laurita KR, Rosenbaum DS, Rudy Y 1995 Two components of the delayed rectifier K⁺ current in ventricular myocytes of the guinea pig type. Theoretical formulation and their role in repolarization. *Circ Res* 77:140–152

DISCUSSION

Winslow: I would like to go back to your opening statement about the company that has 200 compounds that they want to filter down to 40. For the sake of

argument, let's say that they are looking for antiarrhythmic drugs. To model the action of an antiarrhythmic drug requires a great deal of data. Collecting these data is a very labour intensive process. There is the possibility that constructing models of the action of this drug for the 160 that you want to eliminate can take a great deal of time and effort on the part of the company. Have you found that drug companies are willing to follow your guidance in the data that they collect? And are they willing to invest the time and energy in collecting the kind of data that are needed to build models?

Levin: That's an excellent question. There are a number of ways of doing this, but what is required is a standardized technical way of predicting which of these compounds is likely to be successful. Are there standardized data being collected to answer this? The answer is broadly, no. For example, in the case of cardiac toxicity, there is a tremendous effort to collect a standard set of data within one company according to their protocol. We have now evaluated at least 10 different companies' protocols, and they differ quite substantially. As a consequence, we developed a collaboration with Dr Charles Antzelevitch's laboratory to refine the best practices approach to collect standardized data. This can provide a standardized set, or can teach companies the protocols required to generate such data.

Subramaniam: How do you get the kinetic parameters? Do you estimate them, or are they experimentally measured?

Levin: Everything we do is experimentally based. Every model we build has an experimentally based component: if we don't do it ourselves we will find someone to do it. For the kinetic constants it is critical for us to have outside relationships with key scientists who work with us to generate data.

Subramaniam: When you define modules or motifs, do you have any constraints on how you define the modules? What are the ground rules for defining a module?

Levin: There are two ways. Remember that we start with what is important for the pharmaceutical industry. Often, the way we think about modules is with two constraints: what is important for the pharmaceutical industry and what role does it play in the biology? We wrap those two together. In this case we had a specific problem that we had to deal with.

McCulloch: I have a question about compartmentation. In the case of GPCR regulation of the L-type Ca^{2+} channel, if you apply agonist locally to one channel then it will affect just that channel. But if you inject forskolin directly into the cytosol, the other channels will be affected because PKA is partitioned between the membrane and the cytosol. Have you thought about including structural domains as well as functional motifs?

Levin: We have, and we have talked with Les Loew about how some of the work that he has done could be used to create these functional domains, and then fusing them to create a more accurate approach to it. This is essential. It is quite practical

now, but the question is, is it a true representation of biology? I don't think it is meant to be; it is meant to answer quite a specific question.

Loew: I was struck by the semantics: the difference between what we have been calling 'modules' during the course of this symposium and the term 'motif' that you used. It struck me that there really is a difference between the two terms that might be useful. We have been trying to grope for modules that are truly reusable; that can be plugged into different kinds of models with minimum modification. This is certainly a useful goal or concept, and would be enormously beneficial to modelling. But then there is a slightly different approach, which perhaps is encapsulated by the term 'motif'. This is where you can have a particular structure that then can have different components plugged into it as necessary. This is different from a module. Peter Hunter was talking about this in terms of cells that can have various combinations of channels with varying levels of activity, but we can really think about a motif as being the overall structure that can be modified by drawing from the database, and then specialized or customized for a particular kind of cell biological environment or question.

Subramaniam: Then you wouldn't be able to put it into your computational framework, because if you try to take your definition of a structural motif, the time constants are going to be so different that it would not fit very well.

Levin: I don't want to confuse the issue of the general approach. If I have used the word motif, and it is confusing with the concept of the model, let me go back to the original concept: we have adopted basic biological processes that can be adapted from one subcellular level or cell through to another. It is this structured approach that is important to us. This approach to describing components of cells or pathways is a representation of a biological functional unit and also a practical tool. It is economically impractical for us as an organization to constantly have to recreate new entities for each model of a pathway or cell. What we must do is to follow biology. Evolution has been kind to us in that it has offered a way of representing these biological functions in a manner that allows us to encapsulate mathematically the 'module' or 'motif'. I have probably confused the issue; let's put it down to my linguistic slip, but I hope this clarifies the idea.

Loew: I like the idea of expanding the concept of the module, to create a new definition for another more adaptable way of reusing data or model components.

Berridge: The way you have portrayed a module is that it responds to a certain input with a set of outputs and this means that you don't have to worry about what's in the module. However, cells are far more complex because the output signal can vary in both time and space and this then relates to what Shankar Subramaniam says. Therefore, I don't think you can use such a simple definition of a module, because each cell will have a different composition of enzymes, all with different kinetic parameters. Essentially, there is an almost infinite number of modules based on this system. It is a real problem dealing with this because each

cell type has to be treated separately. From what you have said, I understand that you see modules as fixed units with standard output signals.

Levin: Not quite: the data are driven by experimentation. What you have created in the module is the framework for inserting these data. The module is therefore a framework. On the basis of experimental results we can adjust the kinetic parameters for cell type and for species. For example, looking at the example of the myocyte, I showed earlier that in a male the effect of a drug differs from its effect on a female cell. It is important to note that the same framework for the cell exists containing a number of different ionic currents and other components or modules. These frameworks are made sex specific by inserting data into the module that have been developed from experiments on male and female cells. Similarly, if you have a module that has been populated by human data, you can modify the species by inserting data from other species, such as guinea-pig or dog. The output is now species specific.

Paterson: This is an important point in terms of the ‘plug and play’ character of modules. In looking at different cells or across species the structure of the model may be very portable. The parametric configuration of that model is something that will almost certainly have to be fine-tuned and adjusted to accommodate the different cells and tissues. With regard to Les Loew’s point about something that is a little higher-level than a module, there are some lessons to be learned from the software community. A lot of the promise in the early days of object-oriented programming was that it would be possible to build reusable modular programs that had specified inputs and outputs, and from the exterior what went on in the inside didn’t matter. Then these objects could be grafted together to build larger pieces of software without having to work on the details. This promise has not really been fulfilled. However, what has come out of this is the concept of design patterns. That is, for solving a particular class of problem, this is the right approach for dealing with it: you need a class of data structures that looks like this. There needs to be message passing, a graphical user interface (GUI) and at the very least making some parametric changes if not some structural changes to it. I think the idea of plug and play modules in the biology may not be there. There is tremendous leverage to be got from reuse, but we shouldn’t be thinking about modules in terms of plug and play.

Subramaniam: In the Alliance for Cell Signaling we have been struggling with this notion of modules. We have constant discussions about this, mainly because in order for us to quantitatively model once we get frameworks of these signalling proteins, we would need to have some notion of modules. The first definition is that components within a module will not be affected by anything else outside directly, other than the fact that they can have a generic regulation or feedback. The second definition relates to time constants. Within a module, if you don’t have the same set of time constants then the module loses its meaning. Then you

have many processes that are happening elsewhere which will impact that module itself. It doesn't mean they cannot have diverse time constants. The third criterion is that each module has an input–output characteristic that is regulated by just the single feedback-regulation input. Let me give you an example that underlies the complexity of dealing with this. MAP kinase is a good example of a module. If you take MAPK, MAPK K and MAPK KK, in yeast for the same process under low and high osmolarity conditions there are different players with the same module, although the structure of the module itself is preserved. If you go to a mammalian system such as mouse, it becomes very complicated because there are a lot more players. This notion of plug and play will become very difficult. What you are providing is a framework, but a framework should have some constraints that will help you define a module. This brings us back to the markup language (ML) concept.

Levin: This is really important, and I think we are in agreement on this. I don't believe that plug and play *per se* is a realistic approach, unless you can actually define frameworks that have the ability to absorb data of different kinds. The ML concept does this.

Hinch: With these modules you said it takes the results from about 40 papers to deliver the kinetic parameters and the structure of the system. You were saying that if we are going to use this module in a different cell type, the basic structure is transferable, but the experiments will need to be repeated to pull out the kinetic parameters. Do you have a way that, once the structure is defined, of being able to reduce the large number of experiments needed to parameterize the module?

Levin: That's a good question. In certain cases we do. When I said 40 papers, I think I referred specifically to coagulation, which is an extraordinarily well-defined system. This is a system for which we have worked out the kinetics for the last 30 years. What was important in the modelling is that even though these kinetics have already been done for so many years, non-intuitive results emerge all the time when we use the types of modules that we have developed. For example, we were asked to examine the effect of overexpression of factor IX on thrombin production. Intuitively, looking at the coagulation cascade, experts would traditionally say that such over-expression should lead to a more rapid production of thrombin. We modelled this by taking that data in the literature and formatting our model on the published kinetic data. This took us about a week with another day for evaluation of the model using existing compounds. We then analysed the problem and produced a counterintuitive result. Depletion of factor IX leads to a bleeding dyscrasia; interestingly, increasing it also leads to a bleeding dyscrasia, as shown by the model. This has now been demonstrated in animal models. With regard to your main question, can we constrain the data that we require by looking at the model? I think we can in certain cases, but I'm probably not the right person to answer this question in detail. In summary, however, what we do

for the pharmaceutical companies is try to define which are the important points for them to focus experiments on, using a variety of techniques.

Subramaniam: The question you want to ask is if you identify within a module nodes or points in which you can quantitatively measure inputs and outputs, then you can coarse-grain the rest of the structure.

Winslow: I have a comment that stems from something that Shankar Subramaniam said about composing modules that may evolve under different time scales. I think this means that how a module is represented is dependent on the context of the other modules with which it is used. If all the other modules have a slow time scale and you have one that is fast, you have created a stiff system. Somehow you have to recognize that it is composed of these other modules and use a quasi-steady-state approximation to simplify that module. But there are probably many other kinds of interactions like this that will really be necessary in building these modules. This will make it a challenging problem.

Index of contributors

Non-participating co-authors are indicated by asterisks. Entries in bold indicate papers; other entries refer to discussion contributions.

A

Ashburner, M. 34, 36, 37, 38, 39, 41, **66**, 80,
81, 82, 83, 117, 122, 123, 125, 194, 195,
197, 201, 203, 219, 220, 244, 250

B

*Baumgartner Jr., W. **129**
Berridge, M. 38, 64, 65, 82, 83, 103, 148,
149, 160, 195, 218, 240
Boissel, J.-P. 24, 39, 82, 86, 88, 89, 121, 123,
124, 125, 126, 127, 204, 249
*Bray, D. **162**
*Bullivant, D. **207**

C

Cassman, M. 21, 23, 38, 64, 89, 123, 125,
127, 200, 202, 203, 204, 245, 246, 247,
248, 249
*Cho, C. R. **222**
Crampin, E. 61, 63, 65, 102, 180, 195, 247,
250

G

Giles, M. **26**, 35, 36

H

*Helm, P. **129**
Hinch, R. 62, 117, 121, 127, 147, 148, 177,
202, 242
*Huber, G. **4**
Hunter, P. J. 21, 37, 62, 119, 120, 121, 141,
149, 201, 202, **207**, 217, 218, 219, 220,
248, 249, 251

K

Kanehisa, M. 81, **91**, 101, 102, 103
*Krakauer, D. **42**

L

Levin, J. M. 24, 36, 39, 40, 83, 87, 88, 120,
121, 196, 198, 200, 203, 205, **222**, 239,
240, 241, 242, 250
*Lewis, S. **66**
Loew, L. M. 24, 36, 37, 60, 61, 63, 85, 119,
120, 125, 127, **151**, 160, 161, 179, 180,
205, 240

M

Maini, P. K. **53**, 60, 61, 62, 64, 65, 127
McCulloch, A. D. **4**, 20, 21, 22, 23, 24, 38,
39, 81, 86, 90, 103, 117, 120, 123, 125,
141, 142, 149, 179, 199, 200, 205, 218,
220, 239
*McVeigh, E. **129**
*Miller, M. I. **129**

N

*Nielsen, P. M. F. **207**
Noble, D. **1**, 20, 21, 22, 23, 35, 36, 37, 40,
60, 62, 63, 64, 81, 82, 83, 84, 85, 86, 89,
90, 122, 124, 125, 126, 127, 143, 144, 146,
147, 149, 161, 178, 179, **182**, 194, 195,
197, 198, 201, 202, 204, 205, 217, 218,
220, 244, 245, 247, 251

P

Paterson, T. 23, 34, 40, 64, 84, 85, 86, 87,
88, 89, 90, 117, 120, 121, 123, 126, 127,
149, 150, 179, 180, 201, 205, 220, 241,
248, 250

*Peddi, S. **129**

*Penland, R. C. **222**

R

*Ratnanather, T. **129**

Reinhardt, M. 34, 102, 103, 118

S

Shimizu, T. S. 65, 89, 90, 148, **162**, 177,
178, 202

*Stamps, A. T. **222**

Subramaniam, S. 20, 21, 22, 35, 37, 41,
80, 81, 82, 87, 90, 101, 102, 103, **104**,
116, 117, 118, 121, 124, 125, 127, 147,
177, 179, 180, 181, 198, 199, 200, 201,
204, 217, 218, 219, 220, 239, 240, 241,
243, 244, 245, 247, 248, 249, 250, 251

W

Winslow, R. L. 23, 63, 116, 119, 120, **129**,
141, 142, 144, 147, 178, 179, 180, 181,
194, 200, 203, 238, 243

Subject index

A

accessibility 40–41
action potential 7, 226, 228, 233–234, 236
adaptation, molecular brachiation 172–175
adaptive dynamics 48
algorithm development 180
Alliance for Cellular Signaling (AfCS) 105,
117–118
aMAZE 72
amplification 167–168, 200
analytical models 63
AnatML 209
anatomical differences, statistical comparisons
135–138
anatomical ontologies 76
anatomically-based models 210
annotation
 conditionality 80
 Gene Ontology 67–68, 72–74, 81
Anrep effect 22
application service providers (ASPs) 33
archival models 123
arrhythmias 228, 230
ASCI Red 30
ASCI White 30
aspartate signalling 165, 167, 169
ASPs 33
atomic models 16
automatic annotation 72, 73–74
autonomy 124, 125
'avalanche' 187

B

bacterial chemotaxis 162–177
BCT 165
Belousov–Zhabotinskii (BZ) reaction 54–55
Beowulf clusters 30
bifurcation theory 63
Bioelectric Field Modeling Simulation and
 Visualization 181
bioinformatics 5
Biology Workbench 5

BioModel 154, 157
biophysically-based models 210
BioPSE 213
bistability 47
BLAST 72
Borges, Jorge Luis 42
brain 219
BRITE database 92
Brownian random walks 154

C

C++ 31, 39
Ca²⁺ channels 144–150
 heart 185, 187, 190–191, 192
 Virtual Cell 158, 159
cache
 coherency 29
 hierarchy 28
caged thymosin β 159
calcium diffusion 7
cAMP 231–232, 233
CardioPrism™ 213, 224
CardioWave 213
caricaturization 55, 56, 58
cell aggregation 56, 64, 65
cell metabolism, genomic systems models 7
cell signalling 104–116
 amplification 167–168, 200
 Analysis System 107–108
 kinetic models 7
 networks 104–105
 optimization 200, 201
 pathway model construction 108
 pathway reconstruction 114–116
 Signalling Database 107–108
 signalling molecules 105
 state 105
cell types 218
CellML (Cell Markup Language) 111,
119–121, 209
central processing unit, feature size 27
chemical compounds, ontology 74–76

chemotaxis 56, 65, 162–177
 chips 27, 35–36
 classification
 functional 66–67
 list-making 53
 clinical data 64
 CMISS 213
 CNR database 107
 coexistence 47
 collaboration
 electronic publishing 36, 37
 multiple sites 32
 commercial products
 disease-based databases 83
 post-publication pressures 40
 common pool models 145
 communication
 intercommunicability 121–128
 model validation 84, 86–87
 standards 119–121
 complexity 250–251
 Gene Ontology 74–76
 components 124
 computational anatomy 135–138
 computers and computing 26–34, 111
 chips 27, 35–36
 collaboration 32, 36, 37
 distributed-memory systems 30
 grid computing 32
 hard disks 28
 hardware 26–31
 industrial consolidation 27
 memory 27–28, 30, 39
 mobile 27
 networking 28
 operating systems 31
 PC farms 30–31
 power consumption 36
 processors 27
 programming languages 31, 39–40
 remote facility management 32, 33
 research and development costs 27
 shared-memory multiprocessors 29–30
 software 31–32, 33, 213
 support staff 31, 33
 system interconnect 28
 systems 29–31
 vector computing 29
 visualization 28–29, 35
 workstations 30–31
 concepts 67, 69

conditional Gaussian random fields 137
 conformational energy 101
 CONFESSIT 12
 constrained modelling 200, 201, 224–225
 Continuity 12, 213
 continuous approach 153, 165
 continuum models 8, 15
 correlated cluster 96
 counterintuitive results 56, 57, 62–63, 196,
 242
 crossbar switches 28
 curatorial annotation 72, 73

D

d-sotalol trial 228
 data
 amount 34
 collection 238–239
 integration 6
 representation 38
 retrieval 34–35
 storage and access 38, 39
 data-driven drug development 223
 databases 213
 disease-based 83
 integration 68
 metabolic pathways 5
 model linkage 81, 82
 molecular sequence and structure 5
 ontologies 37–38
 see also specific databases
 Dawkins, Richard 245–246
 decision makers 84, 86–87, 88, 89, 90
 description levels 42–44, 45–51
 descriptive models 2, 43, 121–124
 deterministic equation-based modelling 165
 diabetes 201
 DIAN 73, 74
Dictyostelium discoideum 56
 differential equations 55, 56, 153
 diffusion tensor magnetic resonance imaging
 (DTMRI) 131–132
 digital signature 111
 directed acyclic graph (DAG) 69, 220
 cross-product 75
 disease-based databases 83
 distributed-memory systems 30
 distributed queuing systems 31
 Distributed Resource Management (DRM)
 31

documentation 40
 Dolphin Interconnect 28
 drivers 22–23
Drosophila 202, 220, 244
 drug development 222–238
 cardiac excitability, indirect signalling
 228–236
 data-driven 223
 experimental 239
 hypothesis-driven 88, 223
 physiome technology approach 224–228
 programme selection 223
 drug safety screens 233

E

E-Cell 213
 E-Science 32
 early after depolarizations (EADs) 228, 230
 EcoCyc 81
 ecology, levels of description 47–49
 education 37, 202–206
 electrical activity 55
 electronic publishing 36, 37
 electrophysiology 211–212
 elements 124
 embedded computing 27
 emergent property 57
 energy consumption 201
 engineering design 32
 Enterprise Java Beans 111, 113
 Enterprise Java technology 113
 epicardial conduction 130
 epithelium
 Ca²⁺ transport 159
 turnover 220–221
Escherichia coli chemotaxis 162–164, 165,
 167–168, 169
 Ethernet 28
 Euler–Lagrange equations 137
 evolution 195, 245–246, 247
 evolutionary biologists 45
 experimental design 198
 experimentation
 conflicting results 117
 drug development 239
 hypotheses and 23, 25
 uncontrolled variables 194
 explanatory models 2, 121–124
 EXPRESSION database 92

extensible markup language (XML) 111,
 119, 120, 208–209
 external validity 89

F

facial modelling 248
 Fasta files 106
 feature detection 94
 feedback 218
 femur model 213, 214
 FieldML 209
 finite element modelling 180–181
 femur 213, 214
 heart 12, 132–135
 finite volume method 153, 180
 fitness 44, 45–46
 flat file 219
 fluid mechanics 211
 flux balance modelling 199
 Flybase 72, 80
 functional classification 66–67
 functional genomics 5
 functional modules 50–51
 functional states 105, 117
 functionally integrated models 6–7
 heart 8–10

G

G protein-coupled receptors (GPCRs) 116,
 230, 231
 galactose pathway 199
 GenBank 5
 gene
 expression 64
 function 67
 knockouts 161
 networks 212–213
 product annotation 67–68
 regulation 250
 as unit of selection 45
 Gene Ontology 66–80, 81
 annotation 67–68, 72–74, 81
 availability 68–69
 browsers 68
 complexity 74–76
 cross-references 71
 domains 67, 72
 identifiers 69
 isa relationship 70–71

- Gene Ontology (*cont.*)
 modification 82
 partof relationship 70, 71
 redundancy 74–76
 retired terms 70
 structure 69–72
 term changing 69–70
- GENES database 92
- genetic circuits 4
- genetics, levels of description 45–47
- genomics 34–35
- geographical information systems (GIS) 219
- global open biological ontologies (gobo) 76–77
- ‘Go’ 43
- GO Editor 73
- gobo 76–77
- graded release 144
- graph
 comparison 94
 computation 94–96
 feature detection 94
 hierarchy, classification 69
 modules 102–103
 representation 92–94, 101
- graphical user interfaces (GUIs) 35, 108, 114
- grepafloxacin 230
- grid computing 32
- Grid Engine software 31
- H**
- ‘hands on’ use 85, 205–206
- hard disks 28
- hardware development 26–31
- heart 182–194
 action potentials 226, 228, 233–234, 236
 anatomical differences between hearts 135–138
 anatomically detailed models 192
 arrhythmias 228, 230
 Ca²⁺ channels 185, 187, 190–191, 192
 energy conservation in cardiac cycle 183–185
 epicardial conduction mapping 130
 excitability, indirect signalling 228–236
 failure 130, 132, 142, 146
 finite-element modelling 12, 132–135
 hypertrophy 12
 integrative models 8–16, 129–141
- ion concentrations, pumps and exchangers 187–190
- MNT model 185–187
- pacemaker 185–187, 195–196
- Physiome Project 213, 216
- Purkinje fibres 141, 183, 185, 187
- ventricular conduction, three-dimensional modelling 138–139
- ventricular fibres, DTMRI 131–132, 142
- Heartscan 134
- HERG regulation 230, 231, 233
- hierarchical collective motions (HCM) 12, 16, 20–21
- hierarchical graphs 69
- hippocampus, shape variations 138
- Hodgkin–Huxley model 55, 122–123, 182
- holonymy 71
- homogenization 15, 62
- Human Genome Project 5
- human protein annotation 68
- Hutchinson’s epigram 48
- hypernymy/hyponymy 70
- hypotheses
 decision making 84, 86, 88, 89
 drug development 88, 223
 experimentation and 23, 25
- I**
- IBM
 distributed-memory systems 30
 Power4 chip 27, 28
- imaging-based models, heart 129–141
- immunology, levels of description 49–50
- In Silico Cell™ 224
- inference 73
- Infiniband 28
- information science 5
- inositol-1,4,5-trisphosphate (InsP₃) receptors 148–149
 Virtual Cell study 158, 160–161
- instruction scheduler 27
- integration 2, 126–127
- Integrative Biosciences 5
- integrative models 2, 4–19, 121
 heart 8–16, 129–141
- Intel 27
- internal validity 89
- InterProScan 72
- ion channels 187–190, 211–212
 second messenger control 229–236

- isa relationship 70–71
iteration 23–24, 89–90, 198
- J**
Java 111, 113
- K**
KEGG (Kyoto Encyclopaedia of Genes and Genomes) 5, 91–101
categories 91–92
graph computation 94–96
graph representation 92–94, 101
knowledge-based network prediction 96–97
network dynamics 97, 99–101
objects 93, 94
KEYWORD parsing 74
knowledge
gaps 84
models and 24
knowledge-based network prediction 96–97
- L**
L-type Ca²⁺ channels 144, 145, 146
laptops 27
learning, multicellular network models 7
levels of description 42–44, 45–51
levels of selection 44–45, 52
LIGAND database 92
ligand screens 105
linkage disequilibrium 45
Linnean taxonomy 69
Linux 31
Linux PC clusters 30
literature mining 81
local alignment 102
logic of life 62, 125, 188
long-QT 11, 230
long-range inhibition 57
LOVEATFIRSTSIGHT 73
LSF software 31
lumped-parameter models 9, 15
lysine biosynthesis 97, 98
- M**
macromolecular complex models 16
macroscopic description 43
MAGEML 38
MAPK cascades 51, 212, 242
mark-up languages 111, 119–121, 208–209
Markov models 11, 16
mathematical models, complex behaviour 54–59
memory
gene transcription 218
multicellular network models 7
memory (computers) 27–28, 39
distributed-memory 30
shared-memory 29–30
meronymy 71
message passing interface (MPI) 32
metabolic pathways
databases 5
flux balance modelling 199
genetic circuit expression 4
Physiome Project 212
Metabolic Pathways Database 5
methylation 173, 245
MGED 77
microarrays
data retrieval 34–35
data storage and representation 38
microscopic description 43
mitochondria
function and replication 51
morphology, respiratory efficiency 159
mitogen-activated protein kinase (MAPK) cascades 51, 212, 242
mitosis, nuclear envelope breakdown 159
MNT model 185–187
models and modelling
acceptability 202–206
accessibility 40–41
availability 84–85
bottom-up/top-down 1–2
components 124
database linkage 81, 82
detail 210
dissemination 86, 87–88
documentation 40
failure 2, 186, 195–196, 197
history 24–25
levels 2
purposes 121–124
role 2
validation 84–90
modules and modularity 188, 240–242
caricature models 56, 58

modules and modularity (*cont.*)
 definitions 124–125, 147, 239
 from graphs 102–103
 molecular biology 50–51
 motifs and 225, 240
 plug and play 241–242
 time scales 241–242, 243

molecular biology, levels of description
 50–51

molecular brachiation 172–175

molecular models 16

molecular sequence and structure databases
 5

Molecule List, automated data 106–107

Molecule Pages 105, 106, 111
 automated data 106–107
 supporting databases 107

Monte Carlo simulation 168

Moore's law 27, 34

motif-based modelling 225, 236–237, 239,
 240

motor bias 169

MPI (message passing interface) 32

Myricom 28

Myrinet 2000 network 30

MySQL 80–81

N

National Center for Biotechnology
 Information 5

National Simulation Resource 9

natural selection 44

natural system 43

NCBI-NR database 107

nearest-neighbour coupling 168

Nernst–Planck equation 153

nerve impulses 55

networking, computer 28

networks 4, 92
 dynamics 97, 99–101
 genes 212–213
 interactions 94
 knowledge-based prediction 96–97
 optimization 247–249
 prediction 94, 96–97
 reconstruction 105
 robustness 246–247
 signalling 104–105
 topology 102, 116

neuroblastoma cells, Ca²⁺ signalling 158

nuclear envelope, mitosis 159

nuclear medicine 9

nucleic acid sequence comparison 96

O

On Exactitude in Science (Borges) 42

on-line tools 5

ontologies 37–38, 124, 219–220
 gobo 76–77
see also Gene Ontology

OpenMP 31–32

operating systems 31

optimization
 for networks 247–249
 of parameters 202
 signalling pathways 200, 201

Oracle 30, 118

Oracle application server (OAS) 111

organ models 9, 15, 192

Oxsoft 40

P

pacemaker models 185–187, 195–196

PANTHER 73–74

parameters 127
 optimization 202

particle physics 32

part of relationship 70, 71

PATHWAY database 92, 96–97

PathwayPrism™ 213, 224, 225

pattern formation 7
 Turing model 57

PC farms 30–31

personal data assistants (PDAs) 27

perturbations 97, 99, 101, 115–116, 117

pharmaceutical industry
 data collection 238–239
 decision making 84, 86, 88, 90
 model failure 195, 196
see also drug development

phase-plane analysis 55, 60

phenomenological models 43–44, 52

phosphorylation 245

physical principles 179

physiome 5
 drug development 224–228

Physiome Project 207–217
 databases 213
 mark-up languages 208–209

model hierarchy 209–213
 projects 213–216
 software 213
 plastic models 172
 Platform Computing, LSF software 31
 plug and play modules 241–242
 population genetics 45–47
 post-genomics 91
 post-translational modification 245
 power consumption 36
 Power4 chip 27, 28
 predictive models 2, 89–90, 91, 121, 123
 heart 11
 processor performance 27
 programming languages 31, 39–40
 Protein Data Bank 5
 protein folding 8
 protein interaction screens 105
 protein kinase A 231–232, 233, 234, 236
 Protein List 106–107
 protein sequence comparison 96
 pseudo-steady approximation 153
 publication methods 36–37
 Purkinje fibres 141, 183, 185, 187

Q

QT prolongation 11, 230

R

RAID 28
 RanGTPase system, Virtual Cell 159
 reaction–diffusion equations 12, 152–154,
 180, 211
 reduction, mathematical 58, 60–61, 62, 63
 Belousov–Zhabotinskii reaction 55
 cellular aggregation 56
 neural activity 55
 pattern formation 57
 reductionism 1
 redundancy, Gene Ontology 74–76
 redundant array of inexpensive disks (RAID)
 28
 remote facility management 32, 33
 renormalization group theory 44
 representation 124
 respiratory efficiency, mitochondrial
 morphology 159
 RNA granule trafficking, Virtual Cell 154,
 158–159
 robustness 58, 62, 195

definition 64, 127, 248
 evolutionary 247
 ryanodine receptors 148–149

S

safety screens 233
 San Diego Supercomputer Center 5
 SBML 111, 119, 209
 scales, modelling across 10–16, 20–22
 see also time scales
 schema 106, 107
 scientific computing 26–34
 second messenger control 229–236
 selection levels 44–45, 52
 selfish gene 245
 semantics 121–128, 240
 sensitivity 127
 sensitivity analysis 87, 232–233
 sequence comparison 95–96
 SGI 29
 shared-memory multiprocessors 29–30
 short-range activation 57
 signal
 amplification 167–168, 200
 transduction 200–201, 212
 see also cell signalling
 signalling molecules 105
 simplicity 250–251
 simulation models 43–44, 63, 191–192
 single cell models 16
 smooth muscle 149
 snapshots 217
 snoopy bus 29
 software development 31–32, 33, 213
 Sourceforge 41
 species change 47–48
 splice variation 244
 SSDB database 92
 stable coexistence 47
 state dependence 179
 statistical comparisons, heart anatomy
 135–138
 statistical description 43
 stochastic modelling 153–154, 177–178
 STOCHSIM 165–168
 spatially extended 168–171
 stoichiometric ratio 171–172
 structural dynamics 179
 structurally integrated models 5–6, 7
 heart 10–16

Sun Microsystems 29, 31
 supercomputers 26
 SWISS-PROT 68, 74
 SWORD trial 228
 synonymy 71
 synset 71
 systems biology 5, 6–7
 Systems Biology Markup Language (SBML)
 111, 119, 209

T

Tar receptor 165, 166
 Taxman 77
 tethering effect 173–174
 theoretical biology 3, 61, 125, 188, 244–252
 theory validation 89
 thermodynamics 247
 threads, multiple execution 29
 three-dimensional analysis 93–94, 96
 three-dimensional modelling, ventricular
 conduction 138–139
 three-dimensional printers 172
 time scales 13–14, 20–22
 adaptive dynamics 48
 modules 241–242, 243
 structural dynamics 179
 tissue mechanics 210–211
 tissue models, heart 15
 topology 102, 116
 Torsades de Pointes 230
 torso model 213, 215
 toxicology screens 233
 training 37, 202–206
 transduction 200–201, 212
 transforming growth factor β 249
 trivially parallel applications 30–31
 tumour necrosis factor 225–226, 227
 Turing model 57

U

uncertainty 87, 127
 units of selection 44–45
 UNIX 31

V

V Cell 213
 validation
 models 84–90
 pathways 115–116
 variability 127
 VCMDL (Virtual Cell Math Description
 Language) 157
 vector computing 29
 verapamil 230
 Virtual Cell 151–160
 virtual human 220
 virus dynamics 49–50
 visualization 28–29, 35
 vocabularies 67, 106

W

web-based publishing 36, 37
 weighted-ensemble Brownian dynamics 14
 weighted Sobelov norm 133–134
 whole organ models 9, 15, 192
 Windows 31
 WordNet synsets 71
 workstations 30–31

X

XML 111, 119, 120, 208–209
 XQL 209
 XSIM 213
 XSL 209