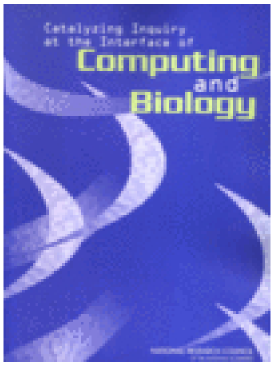


Catalyzing Inquiry at the Interface of Computing and Biology

John C. Wooley and Herbert S. Lin, Editors, Committee on Frontiers at the Interface of Computing and Biology, National Research Council

ISBN: 0-309-54937-X, 468 pages, 8 1/2 x 11, (2005)

Catalyzing Inquiry at the Interface of Computing and Biology

John C. Wooley and Herbert S. Lin, editors

Committee on Frontiers at the Interface of Computing and Biology

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

Support for this project was provided by the Defense Advanced Research Projects Agency under Contract No. MDA972-00-1-0005, the National Science Foundation under Contract No. DBI-0094528, the Department of Health and Human Services/National Institutes of Health (including the National Institute of General Medical Sciences and the National Center for Research Resources) under Contract No. N01-OD-4-2139, the Department of Energy under Contract No. DE-FG02-02ER63336, the Department of Energy's Office of Science (BER) under Interagency Agreement No. DE-FG02-04ER63934, and National Research Council funds. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-09612-X
Library of Congress Control Number: 2005936580

Cover designed by Jennifer M. Bishop.

This report is available from
Computer Science and Telecommunications Board
National Research Council
500 Fifth Street, N.W.
Washington, DC 20001

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

COMMITTEE ON FRONTIERS AT THE INTERFACE OF COMPUTING AND BIOLOGY

JOHN C. WOOLEY, University of California at San Diego, *Chair*
ADAM P. ARKIN, University of California at Berkeley and Lawrence Berkeley National Laboratory
ERIC BRILL, Microsoft Research Labs
ROBERT M. CORN, University of California at Irvine
CHRIS DIORIO, University of Washington
LEAH EDELSTEIN-KESHET, University of British Columbia
MARK H. ELLISMAN, University of California at San Diego
MARCUS W. FELDMAN, Stanford University
DAVID K. GIFFORD, Massachusetts Institute of Technology
TAKEO KANADE, Carnegie Mellon University
STEPHEN S. LADERMAN, Agilent Laboratories
JAMES S. SCHWABER, Thomas Jefferson Medical College

Staff

Herbert Lin, Senior Scientist and Study Director
Geoff Cohen, Consultant to CSTB
Mitchell Waldrop, Consultant to CSTB
Daehee Hwang, Consultant to Board on Biology
Robin Schoen, Senior Staff Officer
Elizabeth Grossman, Senior Staff Officer (through March 2001)
Jennifer Bishop, Program Associate
D.C. Drake, Senior Program Assistant (through March 2003)

COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD

For more information on CSTB, see its Web site at http://www.cstb.org, write to CSTB, National Research Council, 500 Fifth Street, N.W., Washington, DC 20001, or call (202) 334-2605, or e-mail the CSTB at cstb@nas.edu.

Preface

In the last decade of the 20th century, computer science and biology both emerged as fields capable of remarkable and rapid change. Moreover, they evolved as fields of inquiry in ways that draw attention to their areas of intersection. The continuing advancements in technology and the pace of scientific research present the means for computing to help answer fundamental questions in the biological sciences and for biology to demonstrate that new approaches to computing are possible.

Advances in the power and ease of use of computing and communications systems have fueled computational biology (e.g., genomics) and bioinformatics (e.g., database development and analysis). Modeling and simulation of biological entities such as cells have joined biologists and computer scientists (and mathematicians, physicists, and statisticians too) to work together on activities from pharmaceutical design to environmental analysis.

On the other side, computer scientists have pondered the significance of biology for their field. For example, computer scientists have explored the use of DNA as a substrate for new computing hardware and the use of biological approaches in solving hard computing problems. Exploration of biological computation suggests a potential for insight into the nature of and alternative processes for computation, and it also gives rise to questions about hybrid systems that achieve some kind of synergy of biological and computational systems. And there is also the fact that biological systems exhibit characteristics such as adaptability, self-healing, evolution, and learning that would be desirable in the information technologies that humans use.

Making the most of the research opportunities at the interface of computing and biology—what we are calling the BioComp interface—requires illuminating what they are and effectively engaging people from both computing and biology. As in other contexts, the challenges of interdisciplinary education and of collaboration are significant, and each will require attention, together with substantive work from both policy makers and researchers. At the start of the 1990s, attempts were made to stimulate mutual interest and collaboration among young researchers in computing and biology. Those early efforts yielded nontrivial successes, but in retrospect represented a Version 1.0 prototype for the potential in bringing the two fields together. Circumstances today seem much more favorable for progress. New research teams and training programs have been formed as individual investigators from the respective communities, government agencies, and private foundations have become increasingly engaged. Similarly, some larger groups of investigators from different backgrounds have been able to obtain funding to work together to address cross-disciplinary research problems. It is against this background that the committee sees a Version 2.0 of the BioComp interface emerging that will yield unprecedented progress and advance.

The range of possible activities at the BioComp interface is broad, and accordingly so is the range of interested agencies, which include the Defense Advanced Research Projects Agency (DARPA), the National Science Foundation (NSF), the Department of Energy (DOE), and the National Institutes of Health (NIH). These agencies have, to varying degrees, recognized that truly cross-disciplinary work would build on both computing and biology, and they have sought to advance activities at the interface.

This report by the Committee on Frontiers at the Interface of Computing and Biology seeks to establish the intellectual legitimacy of a fundamentally cross-disciplinary collaboration between biologists and computer scientists. That is, while some universities are increasingly favorable to research at the intersection, life science researchers at other universities are strongly impeded in their efforts to collaborate. This report addresses these impediments and describes some strategies for overcoming them.

In addition, this report provides a wealth of well-documented examples. As a rule, these examples have generally been selected to illustrate the breadth of the topic in question, rather than to identify the most important areas of activity. That is, the appropriate spirit in which to view these examples is "let a thousand flowers bloom," rather than one of "finding the prettiest flowers." It is hoped that these examples will encourage students in the life sciences to start or to continue study in computer science that will enable them to be more effective users of computing in their future biological studies. In the opposite direction, the report seeks to describe a rich and diverse domain—biology—within which computer scientists can find worthy problems that challenge current knowledge in computing. It is hoped that this awareness will motivate interested computer scientists to learn about biological phenomena, data, experimentation, and the like—so that they can engage biologists more effectively.

To gather information on such a broad area, the committee took input from a wide variety of sources. The committee convened two workshops in March 2001 and May 2001, and committee members or staff attended relevant workshops sponsored by other groups. The committee mined the published literature extensively. It solicited input from other scientists known to be active in BioComp research. An early draft of the report was examined by a number of reviewers far larger than usual for National Research Council (NRC) reports, and the draft was modified in accordance with their extensive input, which helped the committee to sharpen its message and strengthen its presentation.

The result of these efforts is the first comprehensive NRC study that suggests a high-level intellectual structure for federal agencies for supporting work at the BioComp interface. Although workshop reports have been supported by individual agencies on the subject of computing applied to various aspects of biological inquiry, the NRC has not until now undertaken a study whose intent was to be inclusive.

Within the NRC, the lead unit on this project was the Computer Science and Telecommunications Board (CSTB), and Marjory Blumenthal and Elizabeth Grossman launched the project. The committee also acknowledges with gratitude the contribution of the Board on Biology—Robin Schoen continued work on the project after Elizabeth Grossman's departure. Geoff Cohen and Mitch Waldrop, consultants to CSTB, made major substantive contributions to this report. A variety of project assistants, including D.C. Drake, Jennifer Bishop, Gloria Westbrook, and Margaret Huynh, provided research and administrative support. Finally, grateful thanks are offered to DARPA, NIH, NSF, and DOE for their financial support for this project as well as their patience in awaiting the final report. No single agency can respond to the challenges and opportunities at the interface, and the committee hopes that its analysis will facilitate agency efforts to define their own priorities, set their own path, and participate in what will be a continuing adventure along the frontier at this exciting and promising interface, which will continue to develop throughout the 21st century.

A Personal Note from the Chair

The committee found the scope of the study and the need to achieve an adequate level of balance in both directions around the BioComp interface to be a challenge. This challenge, I hope, has been met, but this was only possible due to the recruitment of an outstanding physicist turned computer science policy expert from the NRC. Specifically, after the original series of meetings, Herb Lin from the CSTB side of the NRC joined the effort, and most notably, followed up on the committee's earlier analyses by interviewing numerous individuals engaged in both biocomputing (applications of biology to computing) and computational biology (applications of computing to biology). This was invaluable, as was Herb's never ending enthusiasm, insight into the nature of the interdisciplinary discussions that are growing, and his willingness to engage in learning a lot about biology. The report could never have been completed without his persistence. His expertise in editing and analytical treatment of policy and technical material allowed us to sustain a broad vision. (Even with the length and breadth of this study, we were able to cover only selected areas at the interface.) The committee's efforts were sustained and accelerated by Herb's determination that we stay the course despite the size of the task, and by his insightful comments, criticisms, and suggestions on every aspect of the study and the report.

John Wooley, *Chair*
Committee on Frontiers at the Interface
of Computing and Biology

Acknowledgment of Reviewers

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

Harold Abelson, Massachusetts Institute of Technology,
Eric Benhamou, Benhamou Global Ventures, LLC,
Mina Bissell, Lawrence Berkeley National Laboratory,
Gaetano Borriello, University of Washington,
Dennis Bray, University of Cambridge,
Steve Burbeck, IBM,
Andrea Califano, Columbia University,
Charles Cantor, Boston University,
David D. Clark, Massachusetts Institute of Technology,
G. Bard Ermentrout, University of Pittsburgh,
Lisa Fauci, Tulane University,
David Galas, Keck Graduate Institute,
Leon Glass, McGill University,
Mark D. Hill, University of Wisconsin-Madison,
Tony Hunter, The Salk Institute for Biological Studies,
Sara Kiesler, Carnegie Mellon University,
Isaac Kohane, Children's Hospital,
Nancy Kopell, Boston University,
Bud Mishra, New York University,
William Noble, University of Washington,

Alan S. Perelson, Los Alamos National Laboratory,
Robert J. Robbins, Fred Hutchinson Cancer Research Center,
Lee Segel, The Weizmann Institute of Science,
Larry L. Smarr, University of California, San Diego,
Sylvia Spengler, National Science Foundation,
William Stead, Vanderbilt University,
Suresh Subramani, University of California, San Diego,
Charles Taylor, University of California, Los Angeles, and
Andrew J. Viterbi, Viterbi Group, LLC.

Although the reviewers listed above have provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Russ Altman, Stanford University. Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring committee and the institution.

Contents

APPENDIXES

Catalyzing Inquiry at the Interface of Computing and Biology

Executive Summary

Despite some apparent differences, biology and information technology (IT) have much in common. They are two of the most rapidly changing fields today—the former because of enormous influxes of new, highly heterogeneous data, and the latter because of exponentially decreasing price-performance ratios. They both deal with entities of astounding complexity (organisms in the case of biology, networks and computer systems in the case of information technology), although in the IT context, the significance of the constituent connections and components is much better understood than in the biological context. Also, they both have profound and revolutionary implications for science and society. Biological science and technology have the potential to contribute strongly to society in improving human health and well-being. The potential impacts include earlier diagnoses and more powerful treatments for diseases, rapid environmental cleanup, and more robust food production. Computing and information technology enable human beings to acquire, store, process, and interpret enormous amounts of information that continue to underpin much of modern society.

Against that backdrop, this report considers potential interactions between biology and computing—the "BioComp" interface. To understand better the potential synergies at the BioComp interface and to facilitate the development of new collaborations between the scientific communities in both fields that can better exploit these synergies, the National Research Council established the Committee on Frontiers at the Interface of Computing and Biology. For simplicity, this report uses "computing" to refer to the broad domain encompassed collectively by terms such as computing, computation, modeling and simulation, computer science, computer engineering, informatics, information technology, scientific computing, and computational science. (Analytical techniques without a strong machine-assisted computational dimension are generally excluded from this study, although they are mentioned from time to time when there is an interesting relationship to computing.) Similarly, the report uses the term "21st century biology" to refer to all fields of endeavor in the biological, biochemical, and biomedical sciences.

Obviously, the union of computing with biology results in an extraordinarily broad area of interest. Thus, this report is not intended to be comprehensive in the sense of seeing how every subfield of biology might connect to every topic in computing. Instead, it seeks to sample the intellectual terrain in enough places so as to give the reader a sense of the kinds of activities under way, and its spirit should be understood as "letting a thousand flowers bloom" rather than "identifying the prettiest flowers in the landscape."

COMPUTING'S IMPACT ON BIOLOGY

Twenty-first century biology will integrate a number of diverse intellectual notions. One integration is that of the reductionist and systems approaches—a focus on components of biological systems combined with a focus on interactions among these components. A second integration is that of many distinct strands of biological research: taxonomic studies of many species, the enormous progress in molecular genetics, steps toward understanding the molecular mechanisms of life, and a consideration of biological entities in relationship to their larger environment. A third integration is that computing will become highly relevant to both hypothesis testing and hypothesis generation in empirical work in biology. Finally, 21st century biology will also encompass what is often called discovery science—the enumeration and identification of the components of a biological system independently of any specific hypothesis about how that system functions (a canonical example being the genomic sequencing of various organisms). Twenty-first century biology will embrace the study of an inclusive set of biological entities, their constituent components, the interactions among components, and the consequences of those interactions, from molecules, genes, cells, and organisms to populations and even ecosystems.

How will computing play in 21st century biology? Life scientists have exploited computing for many years in some form or another. Yet what is different today—and will increasingly be so in the future—is that the knowledge of computing needed to address many interesting biological problems can no longer be learned and exploited simply by "hacking" and reading the manuals. Indeed, the kinds and levels of expertise needed to address the most challenging problems of 21st century biology stretch the current state of knowledge of the field—a point that illuminates the importance of real computing research in a biological context.

This report identifies four distinct but interrelated roles of computing for biology.

1. *Computational tools* are artifacts—usually implemented as software but sometimes hardware—that enable biologists to solve very specific and precisely defined problems. Such biologically oriented tools acquire, store, manage, query, and analyze biological data in a myriad of forms and in enormous volume for its complexity. These tools allow biologists to move from the study of individual phenomena to the study of phenomena in a biological context; to move across vast scales of time, space, and organizational complexity; and to utilize properties such as evolutionary conservation to ascertain functional details.
2. *Computational models* are abstractions of biological phenomena implemented as artifacts that can be used to test insights, to make quantitative predictions, and to help interpret experimental data. These models enable biological scientists to understand many types of biological data in context, even in very large volume, and to make model-based predictions that can then be tested empirically. Such models allow biological scientists to tackle difficult problems that could not readily be posed without visualization, rich databases, and new methods for making quantitative predictions. Biological modeling itself has become possible because data are available in unprecedented richness and because computing itself has matured enough to support the analysis of such complexity.
3. A *computational perspective on or metaphor for biology* applies the intellectual constructs of computer science and information technology as ways of coming to grips with the complexity of biological phenomena that can be regarded as performing information processing in different ways. This perspective is a source of information and computing abstractions that can be used to interpret and understand biological mechanisms and function. Because both computing and biology are concerned with function, information and computing abstractions can provide well-understood constructs that can be used to characterize the biological function of interest. Further,

such abstractions may well provide an alternative and more appropriate language and set of abstractions for representing biological interactions, describing biological phenomena, or conceptualizing some characteristics of biological systems.

4. *Cyberinfrastructure and data acquisition* are enabling support technologies for 21st century biology. Cyberinfrastructure—high-end general-purpose computing centers that provide supercomputing capabilities to the community at large; well-curated data repositories that store and make available to all researchers large volumes and many types of biological data; digital libraries that contain the intellectual legacy of biological researchers and provide mechanisms for sharing, annotating, reviewing, and disseminating knowledge in a collaborative context; and high-speed networks that connect geographically distributed computing resources—will become an enabling mechanism for large-scale, data-intensive biological research that is distributed over multiple laboratories and investigators around the world. New data acquisition technologies such as genome sequencers will enable researchers to obtain larger amounts of data of different types and at different scales, and advances in information technology and computing will play key roles in the development of these technologies.

Why is computing in all of these roles needed for 21st century biology? The answer, in a word, is data. The data relevant to 21st century biology are highly heterogeneous in content and format, multimodal in method of collection, multidimensional in time and space, multidisciplinary in creation and analysis, multiscale in organization, international in relevance, and the product of collaborations and sharing. Consider, for example, that biological data may consist of sequences, graphs, geometric information, scalar and vector fields, patterns of organization, constraints, images, scientific prose, and even biological hypotheses and evidence. These data may well be of very high dimension, since data points that might be associated with the behavior of an individual unit must be collected for thousands or tens of thousands of comparable units.

These data are windows into structures of immense complexity. Biological entities (and systems consisting of multiple entities) are sufficiently complex that it may well be impossible for any human being to keep all of the essential elements in his or her head at once; if so, it is likely that computers will be the vessel in which biological theories are held, formed, and evaluated. Furthermore, because of evolution and a long history of environmental accidents that have driven processes of natural selection, biological systems are more properly regarded as engineered entities than as objects whose existence might be predicted on the basis of the first principles of physics, although the evolutionary context means that an artifact is never "finished" and rather has to be evaluated on a continuous basis. The task of understanding thus becomes one of "reverse engineering"—attempting to understand the construction of a device about whose design little is known but from which much indicative empirical data can be extracted.

Twenty-first century biology will be an information science, and it will use computing and information technology as a language and a medium in which to manage the discrete, nonsymmetric, largely nonreducible, unique nature of biological systems and observations. In some ways, computing and information will have a relationship to the language of 21st century biology that is similar to the relationship of calculus to the language of the physical sciences. Computing itself can provide biologists with an alternative, and possibly more appropriate, language and sets of intellectual abstractions for creating models and data representations of higher-order interactions, describing biological phenomena, and conceptualizing some characteristics of biological systems.

BIOLOGY'S IMPACT ON COMPUTING

From the computing side (i.e., for the computer scientist), there is an as-yet-unfulfilled promise that biology may have significant potential to influence computer design, component fabrication, and software. The essential premise is that biological systems possess many qualities that would be desirable in the information technology that humans use. For example, computer and information scientists are looking for ways to make computers more adaptive, reliable, "smarter," faster, and resilient. Biological systems excel at finding and learning good—but not necessarily optimal—solutions to ill-posed problems on time scales short enough to be useful to them. They efficiently store "data," integrate "hardware" and "software," self-correct, and have many other properties that computing and information science might capture in order to achieve its future goals. Especially for areas in which computer science lacks a well-developed theory or analysis (e.g., the behavior of complex systems or robustness), biology may have the most to contribute.

The impact of biology and biological sciences on advances in computing is, however, more speculative than the reverse, because such considerations are, with only a few exceptions, relevant to future outcomes and not to what has been or is already being delivered. Humans understand computing artifacts much better than they do biological organisms, largely because humans have been responsible for the design of computing artifacts. Absent a comparable base of understanding of biological organisms, the historical and contemporary contributions from biology to computing have been largely metaphorical and can be characterized more readily as inspiration, rather than advances having a straightforward or linear impact.

This difference may be one of time scale. Because today's computing already contributes directly in an essential way to advancing biological knowledge, a path for the near-term future can be readily described. Contemporary advances in computing provide new opportunities for understanding biology, and this will continue to be true for the foreseeable future. Advances in biological understanding may yet have enormous value for changing computing paradigms (e.g., as may be the case if neural information processing is understood more fully)—but these advances are themselves contingent on work done over a considerably longer time scale.

ILLUSTRATIVE PROBLEM DOMAINS AT THE BIOCOMP INTERFACE

Both life scientists and computer scientists will draw inspiration and derive utility from other fields—including each other's—as they see fit. Nevertheless, one way of making progress is to address problems that emerge naturally at the BioComp interface. Problem-focused research carries the major advantage that problems offered by nature do not respect disciplinary boundaries; hence, in making progress against challenging problems, practitioners of different disciplines must learn to work on problems that are shared.

The BioComp interface drives many problem domains in which the expenditure of serious intellectual effort can reasonably be expected to generate significant new knowledge in biology and/or computing. Compared to many of grand challenges in computational biology outlined over the past two decades, making significant progress in these problem domains will call for a longer time scale, greater resources, and more extensive basic progress in computing and in biology.

Biological insight could take different forms—the ability to make new predictions, the understanding of some biological mechanism, the construction of a synthetic biological mechanism. The same is true for computing—insight might take the form of a new biologically inspired approach to some computing problem, different hardware, or novel architecture.

This report discusses a number of interesting problem domains at the BioComp interface, but given the breadth of the cognizant scientific arenas, no attempt is made to be exhaustive. Rather, topics have been selected to span a space of possible problem domains, and no inferences should be made concerning the omission of any problem from this list. The problem domains discussed in this report include high-fidelity cellular modeling and simulation, the development of a synthetic cell, neural information processing and neural prosthetics, evolutionary biology, computational ecology, models that facilitate individualized medicine, a digital human on which a surgeon can operate virtually, computational theories of self-assembly and self-modification, and a theory of biological information and complexity.

THE ROLE OF ORGANIZATION AND INFRASTRUCTURE IN CREATING OPPORTUNITIES AT THE INTERFACE

The committee believes that over time, computing will assume an increasing role in the working lives of nearly all biologists. But given the societal benefits that accompany a fuller and more systematic understanding of biological phenomena, it is better if the computing-enabled 21st century biology arrives sooner rather than later.

This point suggests that cultural and organizational issues have at least as much to do with the nature and scope of the biological embrace of computing as do intellectual ones. The report discusses barriers to cooperation arising from differences in organizational culture and differences in intellectual style.

Consider organizational cultures. In many universities, for example, it is difficult for scholars working at the interface between two fields to gain recognition (e.g., tenure, promotion) from either—a fact that tends to drive such individuals toward one discipline or another. The short-term goals in industrial settings also inhibit partnerships along the interface because of the longer time frame for payoff. Nonetheless, the committee believes that a synergistic cooperation between practitioners in each field, in both basic and applied settings, will have enormous payoffs despite the real differences in intellectual style.

Coordination costs are another issue, because they increase with interdisciplinary work. Computer scientists and biologists are likely to belong to different departments or universities, and when they try to work together, the lack of physical proximity makes it harder for collaborators to meet, to coordinate student training, and to share physical resources. In addition, bigger projects increase coordination costs, and interdisciplinary projects are often larger than unidisciplinary projects. Such costs are reflected in delays in project schedules, poor monitoring of progress, and an uneven distribution of information and awareness of what others in the project are doing. They also reduce people's willingness to tolerate logistical problems that might be more tolerable in their home contexts, increase the difficulty of developing mutual regard and common ground, and can lead to more misunderstandings.

Differences of intellectual style occur because the individuals involved are first and foremost intellectuals. For example, for the computer scientist, the notions of modeling systems and using abstractions are central to his or her work. Using these abstractions and models, computer scientists are able to build some of the most complex artifacts known. But many—perhaps most—biologists today have a deep skepticism about theory and models, at least as represented by mathematics-based theory and computational models. And many computer scientists, mathematicians, and other theoretically inclined researchers fail to recognize the complexity inherent in biological systems. As a result, there is often an intellectual tension between simplification in service of understanding and capturing details in service of fidelity—and such a tension has both positive and negative consequences.

Cooperation will require that practitioners in each field learn enough about the other to engage in substantive conversations about hard biological problems. To take one of the most obvious examples, the different fields place different emphases on the role of empirical data vis-à-vis theory. Accurate data from biological organisms impose "hard" constraints on the biologist in much the same way that results from theoretical computer science impose hard constraints on the computer scientist. A second example is that whereas computer scientists are trained to develop general solutions that give guarantees about events in terms of their worst-case performance, biologists are interested in specific solutions that relate to very particular (though voluminous) datasets.

Finally, institutional difficulties often arise in academic settings for work that is not traditional or not easily identified with existing departments. These differences derive from the structure and culture of departments and disciplines, and they lead to scientists in different disciplines having different intellectual and professional goals and experiencing different conditions for their career success. Collaborators from different disciplines must find and maintain common ground, such as agreeing on goals for a joint project, but must also respect one another's separate priorities, such as having to publish in primary journals, present at particular conferences, or obtain tenure in their respective departments according to departmental criteria. Such cross-pressures and expectations from home departments and disciplinary colleagues remain even if the participants in a collaboration develop similar goals for a project.

FINDINGS AND RECOMMENDATIONS

At the outset, the committee had hoped to identify a deep symmetry between computing and biology. That is, it is clear that the impact of computing on biology is increasingly profound, and the symmetrical notion would be that biology would have a comparable effect on computing. However, this proved not to be the case. The impact of computing on biology will be deep and profound, and indeed will span virtually all areas of life sciences research, and in this direction a focus on interesting problem domains (some of which are illustrated above) is a reasonable way to proceed. By contrast, research that explores the impact of biology on computing falls much more into the "high-risk, high-payoff" category. That is, the ultimate value of biology for changing computing paradigms in deep and fundamental ways is as yet unproven. Nevertheless, various biological attributes—robustness, adaptation, damage recovery, and so on—are so desirable from a computing point of view that any intellectual inquiry is valuable if it can contribute to human-engineered artifacts with these attributes.

It is also clear that a number of other areas of inquiry are associated with the BioComp interface; in addition to biology and computing, the interface also draws from chemistry, materials science, bioengineering, and biochemistry. Three of the most important efforts, which can be loosely characterized as different flavors of biotechnology, are (1) analytical biotechnology (which involves the application of biotechnological tools for the creation of chemical measurement systems); (2) materials biotechnology (which entails the use of biotechnological methods for the fabrication of novel materials with unique optical, electronic, rheological, and selective transport properties); and (3) computational biotechnology (which focuses on the potential replacement of silicon devices with nanoscale biomolecular-based computational systems).

The committee underscores the importance of building human capital and, within that enterprise, the special significance of educational innovation at the BioComp interface. The committee endorses the call from other reports that recommend greater training in quantitative sciences (e.g., mathematics, computer sciences) for biologists, but it also believes that students of the new biology would benefit greatly from some study of engineering. Just as engineers must construct physical systems to operate in the real world, so also must nature operate under these same constraints—physical laws—to "design" successful organisms. Despite this fundamental similarity, biology students rarely learn the important analysis, modeling, and design skills common in engineering curricula. The committee believes that the particular area of engineering (electrical, mechanical, computer, etc.) is probably much less relevant than exposure to essential principles of engineering design: the notion of trade-offs in managing competing objectives, control systems theory, feedback, redundancy, signal processing, interface design, abstraction, and the like.

Of course, more than education will have to change. Fifty years ago, academic biology had to choose between altering the then-dominant styles of research to embrace molecular biology or risking obsolescence. The committee believes that a new dawn is visible—and just as molecular biology has become simply part of the biological sciences as a whole, so also will computational biology ultimately become simply a part of the biological sciences. In the interim, however, considerable effort will be required to build and sustain the infrastructure and to train a generation of biologists and computer scientists who can choose the right collaborators to thrive at the BioComp interface.

The committee believes that 21st century biology will be based on a synergistic mix of reductionist and systems biologies. For systems biology researchers, the committee emphasizes that empirical and experimental hypothesis-testing research will continue to be central in providing experimental verification of putative discoveries—and indeed, relevant as much to studies of how components interact as to studies of components themselves. Thus, disparaging rhetoric about the inadequacies and failures of reductionist biology and overheated zeal in promoting systems biology should be avoided. For researchers more oriented toward experimental or empirical work, the committee emphasizes that systems biology will be central in formulating novel, interesting, and in some cases counterintuitive hypotheses to test. The point suggests that agencies that have traditionally supported hypothesis-testing research would do well to cast a wide "discovery" net that supports the development of alternative hypotheses as well as research that supports traditional hypothesis testing.

Twenty-first century biology will require leadership from both biology and computing that links together first-class research efforts in their respective domains. These efforts will necessarily cross traditional institutional boundaries. For example, research efforts in scientific computing will have to exist in both clinical and biological environments if they are to couple effectively to problem domains in the life sciences. Establishment of a pervasive national infrastructure for life sciences research (including the construction of interdisciplinary teams) and development of the requisite IT-enabled tools for the larger community will require both sustained funding and rigorous oversight. Likewise, the departmental imperatives that characterize much of academe will have to be modified if work at the BioComp interface is to flourish.

In general, the committee believes that the most important change in funding policy for the supporters of this area would be to broaden the kinds of work for which they offer support to include the development of technology for data acquisition and analysis and exploratory research that results in the generation of interesting hypotheses to be tested. That said, there is a direct relationship between the speed with which research frontiers advance and the levels of funding allocated to them. Although it understands the realities of a budget-constrained environment, the committee would gladly endorse an increased flow of funding to the furtherance of a truly integrated 21st century biology.

As for the support of biologically inspired computing, the committee believes that its high-risk, high-payoff nature means that supporting agencies should take a broad view of what "biological inspiration" means and should support the field on a level-of-effort basis, recognizing the long-term nature of such work and taking into account the number of researchers doing and likely to do good work in this area and the potential availability of other avenues to improved computing.

From the committee's perspective, the high-level goals articulated by the agencies and programs that support work related to biology's potential contribution to computing seem generally sensible. This is not to say that every proposal supported under the auspices of these agencies' programs would necessarily have garnered the support of the committee—but that would be true of any research portfolio associated with any program.

One important consequence of supporting high-risk research is that it is unlikely to be successful in the short term. Research—particularly of the high-risk variety—is often more "messy" and takes longer to succeed than managers would like. Managers understandably wish to terminate unproductive lines of inquiry, especially when budgets are constrained. But short-term success cannot be the only metric of the value of research, because when it is, funding managers invite hyperbole and exaggeration on the part of proposal submitters, and unrealistic expectations begin to characterize the field. Those believing the hyperbole (and those contributing to it as well) thus overstate the importance of the research and its centrality to the broader goal of improving computing. When unrealistic expectations are not met (and they will not be met, almost by definition), disillusionment sets in, and the field becomes disfavored from both a funding and an intellectual standpoint.

From this perspective, it is easy to see why support for certain fields rises rapidly and then drops precipitously. Wild budget fluctuations and an unpredictable funding environment that changes goals rapidly can damage the long-term prospects of a field to produce useful and substantive knowledge. Funding levels do matter, but programs that provide steady funding in the context of broadly stated but consistent intellectual goals are more likely to yield useful results than those that do not.

Thus, the committee believes that in the area of biologically inspired computing, funding agencies should have realistic expectations, and these expectations should be relatively modest in the near term. Intellectually, their programs should continue to take a broad view of what "biological inspiration" means. Funding levels in these areas ought to be established on a level-of-effort basis (i.e., what the agency believes is a reasonable level of effort to be expended in this area), by taking into account the number of researchers doing and likely to do good work in an area and the potential availability of other avenues to improved computing. In addition, programmatic continuity for biologically inspired computing should be the rule, with playing rules and priorities remaining more or less constant in the absence of profound scientific discovery or technology advances in the area.

CLOSING THOUGHTS

The impact of computing on biology can fairly be considered a paradigm change as biology enters the 21st century. Twenty-five years ago, biology saw the integration of multiple disciplines from the physical and biological sciences and the application of new approaches to understand the mechanisms by which simple bacteria and viruses function. The impact of the early efforts was so significant that a new discipline, molecular biology, emerged, and many biologists, including those working at the level of tissues or systems and whole organisms, came to adopt the approaches and even often the techniques. Molecular biology has had such success that it is no longer a discipline but simply part of life sciences research itself.

Today, the revolution lies in the application of a new set of interdisciplinary tools: computational approaches will provide the underpinning for the integration of broad disciplines in developing a quantitative systems approach, an integrative or synthetic approach to understanding the interplay of biological complexes as biological research moves up in scale. Bioinformatics provides the glue for systems biology, and computational biology provides new insights into key experimental approaches and how to tackle the challenges of nature. In short, computing and information technology applied to biological problems is likely to play a role for 21st century biology that is in many ways analogous to the role that molecular biology has played across all fields of biological research for the last quarter-century—and computing and information technology will become embedded within biological research itself.

1

Introduction

1.1 EXCITEMENT AT THE INTERFACE OF COMPUTING AND BIOLOGY

Sustained progress across all areas of science and technology over the last half-century has transformed the expectations of society in many ways. Yet, even in this context of extraordinary advances, both the biological sciences and the computer and information sciences share a number of characteristics that are compelling.

First, both fields have been characterized by exponential growth, with doubling times on the order of 1-2 years. In information technology (IT), both the component density of microprocessors and the information storage density on hard disk drives have increased exponentially with doubling times from 9 to 18 months. In biology, the rate of growth of the biological literature is characterized by exponential growth as well (e.g., the growth in GenBank is on the order of 60 percent per year, a rate comparable to Moore's law for microprocessors). While these growth rates cannot continue indefinitely, exponential growth is likely at least in the short term.

Second, both fields deal with organisms and phenomena or artifacts of astounding complexity. Both biological organisms and sophisticated computer systems involve very large numbers of components and interconnections between them, and out of these assemblages of components and connections emerges interesting and useful functionality. In the information technology context, the significance of these connections and components is much better understood than in the biological context, not least because human beings have been responsible for the design of information technology systems such as operating systems and computer systems. Still, the capabilities of existing computing methodologies to design or characterize large-scale information systems and networks are being stretched, and in the biological domain, a systems-level understanding of biological or computer networks is both highly important and difficult to achieve. In addition, information technology is a necessary and enabling technology for the study of complex objects. Computers are the scientific instruments that let us see genomes just as electron microscopes let us see viruses, or radio telescopes let us see quasars.

Third, both biology and information technology have profound and revolutionary implications for science and society. From an intellectual standpoint, biology offers at least partial answers to eternal questions such as, What is life? Also, biological science and technology have the potential for great impact on human health and well-being, including improved disease treatments, rapid environmental cleanup, and more robust food production. Computing and information technology enable human beings to acquire, store, process, and interpret enormous amounts of information, and continue to underpin much of modern society.

Finally, several important areas of interaction between the two fields have already emerged, and there is every expectation that more will emerge in the future. Indeed, the belief of the committee that there are many more synergies at the interface between these two fields than have been exploited to date is the motivation for this report. Against this backdrop, it makes good sense to consider potential interactions between the two fields—what this report calls the "BioComp" interface.

As for the nature of computing that can usefully be exploited by life scientists, there is a range of possibilities. For some problems encountered by biology researchers, a very rudimentary knowledge of computing and information technology is quite sufficient. However, as problems become bigger and/or more complex, what one may pick up by hacking and reading manuals is no longer sufficient. To address such problems, the kinds and levels of expertise needed are more likely to require significant formal study of computer science (e.g., as an undergraduate major in the field). And for still more difficult, larger, or more complex problems, the kinds and levels of expertise needed stretch the current state of knowledge of the field—a point that illuminates the importance of real computer science research in a biological context.

Nor is the utility of computing limited to providing tools or models—no matter how sophisticated—for biologists to use. As discussed in Chapter 6, computing can also provide intellectual abstractions that may provide insight into biological phenomena and a useful language for describing such phenomena. As one example, notions of circuit and network and modularity—originally conceptualized in the world of engineering and computer science—have much applicability to understanding biological phenomena.

On the other side, biology refers to the scientific study of the activities, processes, mechanisms, and other attributes of living organisms. For the purposes of this report, biology, biomedicine, life sciences, and other descriptions of research into how living systems work should be regarded as synonymous. In this context, for the past decade, researchers have spoken increasingly of a new biology, a biology of the 21st century, one that is driven by new technologies, that is more automated with tools and methods provided by industrial models, and that often entails high-throughput data acquisition.[1] This report examines the BioComp interface from the perspective of 21st century biology, as a science that integrates traditional empirical and experimental biology with a systems-level biology that considers the multiscale, hierarchical, highly interwoven, or interactive aspects intrinsic to living systems.

1.2 PERSPECTIVES ON THE BIOCOMP INTERFACE

This report addresses computationally inspired ways of understanding biology and biologically inspired ways of understanding computing. Although the committee started its work with the idea that it would discover a single community and intellectual synthesis of biology and computing, closer examination showed that the appropriate metaphor is one of an interface between the two fields rather than a common, shared area of inquiry. Thus, the adventures along the frontier cannot be treated as coming from a single community, and the different objectives have to be recognized.

[1] For example, see National Research Council, *Opportunities in Biology*, National Academy Press, Washington, DC, 1989. High-throughput data acquisition is an approach that relies on the large-scale parallel interrogation of many similar biological entities. Such an approach is essential for the conduct of global biological analyses, and it is often the approach of choice for rapid and comprehensive assessment of biological system properties and dynamics. See, for example, T. Ideker, T. Galitski, and L. Hood, "A New Approach to Decoding Life: Systems Biology," *Annual Review of Genomics and Human Genetics* 2:343-372, 2001. A number of the high-throughput data acquisition technologies mentioned in that article are discussed in Chapter 7 of his report.

1.2.1 From the Biology Side

Biologists have a long history of applying tools from other disciplines to provide more powerful methods to address or even solve their research problems. For example, Anton Van Leeuwenhoek's invention of the optical microscope in the late 1600s opened up a previously unknown world and ultimately brought an entirely new vista to biology—namely, the existence of cells and cellular structures. This remarkable revolutionary discovery would have been impossible without the study of optics—and Leeuwenhoek was a clockmaker.

The biological sciences have drawn heavily from chemistry, physics, and more recently, mathematical modeling. Indeed, the reductionist revolution in biological sciences—which led to the current state of understanding of biological function and mechanism at the molecular level or of specific areas such as neurophysiology—in the past five decades began as chemists, physicists, microbiologists, and others interacted and created what is now known as molecular biology. The applications from the physical sciences are already so well established that it is unnecessary to discuss them at length.

Mathematics and statistics have at times played important roles in designing and optimizing biological experiments. For example, statistical analysis of preliminary data can lead to improved data collection and interpretation in subsequent experiments. In many cases, simple mathematical or physical ideas, accompanied by calculations or models, can suggest experiments or lead to new ideas that are not easily identified with biological reasoning alone. An example of this category of contribution is William Harvey's estimation of the volume of the blood and his finding that a closed circulatory system would explain the anomaly in such calculations. Traditionally, biologists have resisted mathematical approaches for various reasons discussed at length in Chapter 10. To some extent, this history is being changed in modern biology, and it is the premise of this report that an acceleration of this change is highly worthwhile.

Approaches borrowed from another discipline may provide perspectives that are unavailable from inside the disciplinary research program itself. In some cases, these lead to a new integrative explanation or to new ways of studying and appreciating the intricacies of biology. In other cases, this borrowing opens an entirely new subfield of biology. The discovery of the helical structure and the "code" of DNA, impossible without crystallography and innovative biological thinking, is one example. The understanding of electrical signaling in neurons by voltage-gated channels, and the Hodgkin-Huxley equations (based on the theory of electrical circuits), constitute another example. Both of these approaches revolutionized the way biology was conducted and required significant, skilled input from other fields.

The most dramatic scenarios arise when major subfields emerge. An example dating back some decades, and described above in another context, is molecular biology, whose tools and techniques (using advanced chemistry, physics, and equipment based on the above) changed the face of biology. A more recent, current example is genomics with its indelible mark on the way that biology as a discipline is conducted and will be conducted for years to come.

The committee believes that from the perspective of the biology researcher, there is both substantial legacy and future promise regarding the application of computing to biological problems. Some of this legacy is manifested in a several-decade development of private-sector databases (mostly those of pharmaceutical companies) and software for data analysis, in public-sector genetic databases, in the use of computer-generated visualization, and in the use of computation to determine the crystal structures of increasingly complex biomolecules.[2]

Several life sciences research fields have begun to take computational approaches. For example, ecology and evolution were among the first subfields of biology to develop advanced computational simulations based on theory and models of ecosystems and evolutionary pathways. Cardiovascular

[2] See, for example, T. Head-Gordon and J.C. Wooley, "Computational Challenges in Structural and Functional Genomics," *IBM Systems Journal* 40(2):265-296, 2001, available at http://www.research.ibm.com/journal/sj/402/headgordon.pdf.

physiology and studies of the structure and function of heart muscle have involved bioengineering models and combined experimental and computational approaches. All of these computational approaches would have been impossible without solid preexisting mathematical models that led to the intuition and formed the basis for the emerging computational aspects.

Nevertheless, genomics research is simply not possible without information technology. It is not an exaggeration to say that it was the sequencing of complete genomes, more than any other research activity, that brought computational and informatics approaches to the forefront of life sciences research, as well as identifying the need for basic underlying algorithms to tackle biological problems. Only through computational analysis have researchers begun to uncover the implications of genomic-scale sequence data. Apart from specific results thereby obtained, such analysis, coupled with the availability of complete genomic sequences, has changed profoundly how many biologists think, conduct research, and plan strategically to address central research problems.

Today, computing is essential to every aspect of molecular and cell biology, as researchers expand their scope of inquiry from gene sequence analysis to broader investigations of biological complexity. This scope includes the structure and function of proteins in the context of metabolic, genetic, and signaling networks, the sheer complexity of which is overwhelming. Future challenges include the integration of organ physiology, catalogs of species-wide phenotypic variations, and understanding of differences in gene expression in various states of health and disease.

1.2.2 From the Computing Side

From the viewpoint of the computer scientist, there is an as-yet-unfulfilled promise that biology may have significant potential to influence computer design, component fabrication, and software. Today, the impact of biology and biological sciences on advances in computing is more speculative than the reverse (as described in Section 1.2.1), because such considerations are, with only a few exceptions, relevant to future outcomes and not to what has been or is already being delivered.

In one sense, this should not be very surprising. Computing is a "science of the artificial,"[3] whereas biology is a science of the natural, and in general, it is much easier for humans to understand both the function and the behavior of a system that they have designed to fulfill a specific purpose than to understand the internal machinery of a biological black box that evolved as a result of forms and pressures that we can only sketchily guess.[4] Thus, paths along which biology may influence computing are less clear than the reverse, and work in this area should be expected to have longer time horizons and to take the form of many largely independent threads, rather than a hierarchy of interrelated or intellectual thrusts.

Nevertheless, exploring why the biological sciences might be relevant to computing is worthwhile in particular because biological systems possess many qualities that would be desirable in the information technology that humans use. For example, computer and information scientists are looking for ways to make computers more adaptive, reliable, "smarter," faster, and resilient. Biological systems excel at finding and learning adequate—but not necessarily optimal—solutions to ill-posed problems on time scales short enough to be useful to them. They efficiently store "data," integrate "hardware" and "software," self-correct, and have many other properties that computing and information science

[3] "We speak of engineering as concerned with 'synthesis,' while science is concerned with 'analysis.' Synthetic or artificial objects—and more specifically prospective artificial objects having desired properties—are the central objective of engineering activity and skill. The engineer, and more generally the designer, is concerned with how things *ought* to be—how they ought to be in order to *attain goals*, and to *function*." H.A. Simon, *Sciences of the Artificial*, 3rd ed., MIT Press, Cambridge, MA, 1996, pp. 4-5.

[4] This is what neuroscientist Valentino Braitenberg called his law of uphill analysis and downhill synthesis, in *Vehicles: Experiments in Synthetic Psychology*, MIT Press/A Bradford Book, Cambridge, MA, 1984. Cited in Daniel C. Dennett, "Cognitive Science as Reverse Engineering: Several Meanings of 'Top-down' and 'Bottom-up'," *Proceedings of the Ninth International Congress of Logic, Methodology and Philosophy of Science*, D. Prawitz, B. Skyrms, and D. Westerstahl, eds., Elsevier Science North-Holland, 1994.

might capture in order to achieve its future goals. Especially for areas in which computer science lacks a well-developed theory or analysis (e.g., the behavior of complex systems or robustness), biology may have the most to contribute.

To hint at some current threads of inquiry, some researchers envision a hybrid device—a biological computer—essentially, an organic tool for accomplishing what is now carried out in silicon. As an information storage and processing medium, DNA itself may someday be the substance of a massively dense memory storage device, although today the difficulties confronting the work in this area are significant. DNA may also be the basis of nanofabrication technologies.

Biomimetic devices are mechanical, electrical, or chemical systems in which an attempt has been made to mimic the way that a biological system solves a particular problem. Successes include robotic locomotion (based on legged movements of arthropods), artificial blood or skin, and others. Approaches with general-purpose applicability are less clearly successes, though they are still intriguing. These include attempts to develop approaches to computer security that are modeled on the mammalian immune system and approaches to programming based on evolutionary concepts.

Hybrid systems are a promising new technology for measurement of or interaction with small biological systems. In this case, hybrid systems refer to silicon chips or other devices designed to interact directly with a biological sample (e.g., record electrical activity in the flight muscles of a moth) or analyze a small biological sample under field conditions. Here the applications of the technology both to basic scientific problems and to industrial and commercially viable products are exciting.

In the domain of algorithms, swarm intelligence (a property of certain systems of nonintelligent, independently acting agents that collectively exhibit intelligent behavior) and neural nets offer approaches to programming that are radically different from many of today's models. Such applications of biological principles to nonbiological computing could have much value, and Chapter 8 addresses in greater detail some possible biological inspirations for computing. Yet it is also possible that a better understanding of information-processing principles in biological systems will lead as well to greater biological insight; so the dividing line between "applying biological principles to information processing" and "understanding biological information processing" is not as clear as it might appear at first glance. Moreover, even if biology ultimately proves unhelpful in providing insight into potential computing solutions, it is still a problem domain par excellence—one that offers interesting intellectual challenges in which progress will require that the state of computing research be stretched immeasurably.

1.2.3 The Role of Organization and Culture

The possibility—or even the fact—that one field may be well positioned to make or facilitate significant intellectual contributions to the other does not, by itself, lead to harmonious interchange between practitioners in the two fields. Cultural and organizational issues are also very much relevant to the success or failure of collaborations across different fields. For example, one important issue is the fact that much of today's biological research is done in individual laboratories, whereas many interesting problems of 21st century biology will require interdisciplinary teams and physical or virtual centers with capable scientists, distributed wherever they work, involved in addressing difficult problems.

Twenty-first century biology will also see the increasing importance of research programs that have a more industrial flavor and involve greater standardization of instruments and procedures. A small example is that reagent kits are becoming more and more popular, as labs realize that the small advantages that might accrue through the use of a set of customized reagents are far outweighed by the savings in effort associated with the use of such kits. A larger example might be shared devices and equipment of larger-scale and assembly-line-like processes that replace the craft work of individual technicians.

As biologists recognize the inherent difficulties posed by the data-intensive nature of these new research strategies, they will require different—and additional—training in quantitative methods and science. Computing is likely to be central, but since the nature and scope of the computing required will go far beyond what is typically taught in an introductory computing course, real advancement of the frontier will require that computer scientists and biologists recognize and engage each other as intellectual coequals. At the same time, computer scientists will have to learn enough about biology to understand the nature of problems interesting to biologists and must refrain from regarding the problem domain as a "mere" application of computing.

The committee believes that such peer-level engagement happens naturally, if slowly. But accelerating the cultural and organizational changes needed remains one of the key challenges facing the communities today and is one that this report addresses. Such considerations are the subject of Chapter 10.

1.3 Imagine What's Next

In the long term, achievements in understanding and harnessing the power of biological systems will open the door to the development of new, potentially far-reaching applications of computing and biology—for example, the capability to use a blood or tissue sample to predict an individual's susceptibility to a large number of afflictions and the ability to monitor disease susceptibility from birth, factoring in genetics and aging, diet, and other environmental factors that influence the body's functions over time and ultimately to treat such ailments.

Likewise, 21st century biology will advance the abilities of scientists to model, before a treatment is prescribed, the likely biological response of an individual with cancer to a proposed chemotherapy regime, including the likelihood of the effectiveness of the treatment and the side effects of the drugs. Indeed, the promise of 21st century biology is nothing less than a system-wide understanding of biological systems both in the aggregate and for individuals. Such understanding could have dramatic effects on health and medicine. For example, detailed computational models of cellular dynamics could lead to mechanism-based target identification and drug discovery for certain diseases such as cancer,[5] to predictions of drug effects in humans that will speed clinical trials,[6] and to a greater understanding of the functional interactions between the key components of cells, organs, and systems, as well as how these interactions change in disease states.[7]

On another scale of knowledge, it may be possible to trace the genetic variability in the world's human populations to a common ancestral set of genes—to discover the origins of the earliest humans, while learning, along the way, about the earliest diseases that arose in humans, and about the biological forces that shape the world's populations. Work toward all of these capabilities has already begun, as biologists and computer scientists compile and consider vast amounts of information about the genetic variability of humans and the role of that variability in relation to evolution, physiological functions, and the onset of disease.

At the frontiers of the interface, remarkable new devices can be pictured that draw on biology for inspiration and insight. It is possible to imagine, for example, a walking machine—an independent set of legs as agile, stable, and energy-efficient as those of humans or animals—able to negotiate unknown terrain and recover from falls, capable of exploring and retrieving materials. Such a machine would overcome the limitations of present-day rovers that cannot do such things. Biologists and computer scientists have begun to examine the locomotion of living creatures from an engineering and biological perspective simultaneously, to understand the physical and biological controls on balance, gait, speed, and energy expended and to translate this information into mechanical prototypes.

[5] J.B. Gibbs, "Mechanism-Based Target Identification and Drug Discovery in Cancer Research," *Science* 287:1969, 2000.

[6] C. Sander, "Genomic Medicine and the Future of Health Care," *Science* 287:1977, 2000.

[7] D. Noble, "Modeling the Heart—From Genes to Cells to the Whole Organ," *Science* 295:1678, 2002.

We can further imagine an extension of present-day bioengineering from mechanical hearts and titanium hip joints to an entirely new level of devices, such as an implantable neural prosthetic that could assist stroke patients in restoring speech or motor control or could enhance an individual's capability to see more clearly in the dark or process complex information quickly under pressure. Such a prosthetic would marry the speed of computing with the brain's capacity for intelligence and would be a powerful tool with many applications.

With the advancement of computational power and other capabilities, there is a great opportunity and challenge in whether human functions can be represented in digital computational forms. One form of representation of a human being is how it is constructed, starting with genes and proteins. Another form of representation is how a human being functions. Human functions can be viewed at many different levels—physioanatomical, motion-mechanical, and psychocognitive, for example. If it were possible to represent a human being at any or all of these functional levels, then a "digital human" could be created inside the computer, to be used for many applications such as medical surgical training, human-centered design of products, and societal simulation. (There are already such simulations at varying levels of fidelity for particular organs such as the heart.)

The potential breadth and depth of the interface of computing and biology are vast. Box 1.1 is a representative list of research areas already being pursued at the interface; Appendix B at the end of this report provides references to more detailed discussions of these efforts. The excitement and challenge of all of these possibilities drive the increasing interest in and enthusiasm for research at the interface.

Box 1.1
Illustrative Research Areas at the Interface of Computer Science and Biology

- Structure determination of biological molecules and complexes
- Simulation of protein folding
- Whole genome sequence assembly
- Whole genome modeling and annotation
- Full genome-genome comparison
- Rapid assessment of polymorphic genetic variations
- Complete construction of orthologous and paralogous groups of genes
- Relating gene sequence to protein structure
- Relating protein structure to function
- In silico drug design
- Mechanistic enzymology
- Cell network analysis-simulation of genetic networks and the sensitivity of these pathways to component stoichiometry and kinetics
- Dynamic simulation of realistic oligomeric systems
- Modeling of cellular processes
- Modeling of physiological systems in health and disease
- Modeling behavior of schools, swarms, and their emergent behavior
- Simulation of membrane structure and dynamic function
- Integration of observations across scales of vastly different dimension and organization for model creation purposes
- Development of bio-inspired autonomous locomotive devices
- Development of biomimetic devices
- Bioengineering prosthetics

1.4 SOME RELEVANT HISTORY IN BUILDING THE INTERFACE

1.4.1 The Human Genome Project

According to Cook-Deegan,[8] the Human Genome Project resulted from the collective impact of three independent public airings of the idea that the human genome should be sequenced. In 1985, Robert Sinsheimer and others convened a group of scientists to discuss the idea.[9] In 1986, Renato Dulbecco noted that sequencing the genome would be an important tool in probing the genetic origins of cancer.[10] Then in 1988, Charles DeLisi developed the idea of sequencing the genome in the context of understanding the biological and genetic effects of ionizing radiation on survivors of the Hiroshima and Nagasaki atomic bombs.[11]

In 1990, the International Human Genome Consortium was launched with the intent to map and sequence the totality of human DNA (the genome).[12] On April 14, 2003, not quite 50 years to the day after James Watson and Francis Crick first published the structure of the DNA double helix,[13] officials announced that the Human Genome Project was finished.[14] After 13 years and $2.7 billion, the international effort had yielded a virtually complete listing of the human genetic code: a sequence some 3 billion base pairs long.[15]

1.4.2 The Computing-to-Biology Interface

For most of the electronic computing age, biological computing applications have been secondary compared to those associated with the physical sciences and the military. However, over the last two decades, use by the biological sciences—in the form of applications related to protein modeling and folding—went from virtually nonexistent to being the largest user of cycles at the National Science Foundation Centers for High Performance Computing by FY 1998. Nor has biological use of computing capability been limited to supercomputing applications—a plethora of biological computing applications have emerged that run on smaller machines.

During the last two decades, federal agencies also held a number of workshops on computational biology and bioinformatics, but until relatively recently, there was no prospect for significant support

[8] Cook-Deegan's perspective on the history of the Human Genome Project can be found in R.M. Cook-Deegan, *The Gene Wars: Science, Politics, and the Human Genome,* W.W. Norton and Company, New York, 1995.

[9] R. Sinsheimer, "The Santa Cruz Workshop," *Genomics* 5(4):954-956, 1989.

[10] R. Dulbecco, "A Turning Point in Cancer Research: Sequencing the Human Genome," *Science* 231(4742):1055-1056, 1986.

[11] C. DeLisi, "The Human Genome Project," *American Scientist* 76:488-493, 1988.

[12] Cook-Deegan identifies three independent public airings of the idea that the human genome should be sequenced, airings that collectively led to the establishment of the HGP. In 1985, Robert Sinsheimer and others convened a group of scientists to discuss the idea. (See R. Sinsheimer, "The Santa Cruz Workshop," *Genomics* 5(4):954-956, 1989.) In 1986, Renato Dulbecco noted that sequencing the genome would be an important tool in probing the genetic origins of cancer. (See R. Dulbecco, "A Turning Point in Cancer Research: Sequencing the Human Genome," *Science* 231(4742):1055-1056, 1986.) In 1988, Charles DeLisi developed the idea of sequencing the genome in the context of understanding the biological and genetic effects of ionizing radiation on survivors of the Hiroshima and Nagasaki atomic bombs. (See C. DeLisi, "The Human Genome Project," *American Scientist* 76:488-493, 1988.) Cook-Deegan's perspective on the history of the Human Genome Project can be found in R. Cook-Deegan, T*he Gene Wars: Science, Politics, and the Human Genome*, W.W. Norton and Company, New York, 1995.

[13] J.D. Watson and F.H. Crick, "Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid," *Nature* 171(4356):737-738, 1953.

[14] The "completion" of the project had actually been announced once before, on June 26, 2000, when U.S. President Bill Clinton and British Prime Minister Tony Blair jointly hailed the release of a preliminary, draft version of the sequence with loud media fanfare. However, while that draft sequence was undoubtedly useful, it contained multiple gaps and had an error rate of one mistaken base pair in every 10,000. The much-revised sequence released in 2003 has an error rate of only 1 in 100,000, and gaps in only those very rare segments of the genome that cannot reliably be sequenced with current technology. See http://www.genome.gov/11006929.

[15] Various histories of the Human Genome Project can be found at http://www.ornl.gov/sci/techresources/Human_Genome/project/hgp.shtml.

for academic work at the interface. The Keck Foundation and the Sloan Foundation supported training, and numerous database activities have been supported by federal agencies. As the impact of the Human Genome Project and comparative genomics began to reach the community as a whole, the situation changed. An important step came from the Howard Hughes Medical Institute, which in 1999 held a special competition to select professors in bioinformatics and thus provided a strong endorsement of the role of computing in biology.

In 1999, the National Institutes of Health (NIH) also took a first step toward integrating ad hoc support by requesting an analysis of the opportunities, requirements, and challenges from computing for biomedicine. In June 1999, the Botstein-Smarr Working Group on Biomedical Computing presented a report to the NIH entitled *The Biomedical Information Science and Technology Initiative*.[16] Specifically tasked with investigating the needs of NIH-supported investigators for computing resources, including hardware, software, networking, algorithms, and training, the working group made recommendations for NIH actions to support the needs of NIH-funded investigators for biomedical computing.

That report embraces a vision of computing as the hallmark of tomorrow's biomedicine. To accelerate the transition to this new world of biomedicine, the working group sought to find ways "to discover, encourage, train, and support the new kinds of scientists needed for tomorrow's science." Much of the report focuses on national programs to create "the best opportunities that can be created for doing and learning at the interfaces among biology, mathematics, and computation," and argues that "with such new and innovative programs in place, scientists [would] absorb biomedical computing in due course, while supporting the mission of the NIH." The report also identifies a variety of barriers to the full exploitation of computation for biological needs.

In the intervening 4 years, the validity of the Botstein-Smarr Working Group report vision has not been in question; if anything, the expectations, opportunities, and requirements have grown. Computation in various forms is rapidly penetrating all aspects of life sciences research and practice.

- State-of-the-art radiology (and along with it other fields dependent on imaging—neurology, for example) is highly dependent on information technology: the images are filtered, processed reconstructions that are acquired, stored, and analyzed computationally.
- Genomics and proteomics are completely dependent on computation.
- Integrative biology aimed at predictive modeling is not just computationally enabled—it literally cannot occur in a noncomputational environment.

Biomedical scientists of all stripes are increasingly using public resources and computational tools at high levels of intensity such that very significant fractions of the overall effort are in this domain, and it is highly likely that these trends will continue. Yet many of the barriers to full exploitation of computation in the biological sciences that were identified in the Botstein-Smarr report still remain. One primary focus of the present report is accordingly to consider the intellectual, organizational, and cultural barriers that impede or even prevent the full benefits of computation from being realized for biomedical research.

1.4.3 The Biology-to-Computing Interface

The application of biological ideas to the design of computing systems appears through much of the history of electronic computers, in most cases as an outgrowth of attempts to model or simulate a biological system. In the early 1970s, John H. Holland (the first person in the United States to be awarded a Ph.D. in computer science) pioneered the idea of *genetic algorithms*, which use simulated genetic processes (crossover, mutation, and inversion) to search a large solution space of algorithms.[17]

[16] Available at http://www.nih.gov/about/director/060399.htm.

[17] J.H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan Press, Ann Arbor, 1975.

This work grew out of research in the 1950s and 1960s to simulate just such processes in the natural world. A second wave of popularity of this technique came after John Koza described genetic programming, which used similar techniques to modify symbolic expressions that comprised entire programs.[18] Both of these approaches are in use today, especially in research and academic settings.

The history of artificial neural networks also shows a strong relationship between attempts to simulate biology and attempts to construct a new software tool. This research predates even the modern electronic digital computers, since Warren McCulloch and Walter Pitts published a model of a neuron that incorporated analog weights into a binary logic scheme in 1943.[19] This was meant to be used as a model of biological neurons, not merely as an abstract computational processing approach. Research on neural nets continued throughout the next decades, focusing on network architectures (particularly random and layered), mechanisms of self-assembly, and pattern recognition and classification. Significant among this research was Rosenblatt's work on perceptrons.[20] However, lack of progress caused a loss of interest in neural networks in the late 1970s and early 1980s. Hopfield revived interest in the field in 1982,[21] and progress throughout the 1980s and 1990s established neural networks as a standard tool for learning and classifying patterns.

A similar pattern characterizes research into cellular automata. John von Neumann's attempts to provide a theory of biological self-assembly inspired him to apply traditional automata theory to a two-dimensional grid;[22] similar work was being done at the same time by Stanislaw Ulam (who may have suggested the approach to von Neumann). Von Neumann also showed that cellular automata could simulate a Turing machine, meaning that they were a system that could provide universal computation. A boom of popularity for cellular automata followed the publication of the details of John Conway's Game of Life.[23] In the early 1980s, Stephen Wolfram made important contributions to formalizing cellular automata, especially in their role in computational theory,[24] and Toffoli and Margolus stressed the general applicability of automata as systems for modeling.[25]

At a more metaphorical level, IBM has taken initiatives in biologically inspired computing. Specifically, IBM launched its Autonomic Computing initiative in 2001. Autonomic computing is inspired by biology in the sense that biological systems—and in particular the autonomic nervous system—are capable of doing many things that would be desirable in complex computing systems. Autonomic computing is conceived as a way to manage increasingly complex and distributed computing environments as traditional approaches to system management reach their limits. IBM takes special note of the fact that "the autonomic nervous system frees our conscious brain from the burden of having to deal with vital but lower-level functions."[26] Autonomic computing, by IBM's definition, requires that a system be able to configure and reconfigure itself under varying and unpredictable conditions, to continually optimize its workings, to recover from routine and extraordinary events that might cause

[18] J.R. Koza, "Genetically Breeding Populations of Computer Programs to Solve Problems in Artificial Intelligence," pp. 819-827 in *Proceedings of the Second International Conference on Tools for Artificial Intelligence*, IEEE Computer Society Press, Los Alamitos, CA, 1990.

[19] W.S. McCulloch and W.H. Pitts, "A Logical Calculus of the Ideas Immanent in Nervous Activity," *Bulletin of Mathematical Biophysics* 5:115-137, 1943.

[20] R. Rosenblatt, *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*, Spartan Books, Washington, DC, 1962.

[21] J.J. Hopfield, "Neural Networks and Physical Systems with Emergent Collective Computational Abilities," *Proceedings of the National Academy of Sciences* (USA) 79(8):2554-2558, 1982.

[22] J. von Neumann, *Theory of Self-reproducing Automata* (edited and completed by A. W. Burks), University of Illinois Press, 1966.

[23] M. Gardner, "MATHEMATICAL GAMES: The Fantastic Combinations of John Conway's New Solitaire Game 'Life'," *Scientific American* 223(October):120-123, 1970.

[24] S. Wolfram, "Computation Theory of Cellular Automata," *Communications in Mathematical Physics* 96:15-57, 1984.

[25] T. Toffoli and N. Margolus, *Cellular Automata Machines: A New Environment for Modeling*, MIT Press, Cambridge, MA, 1987.

[26] G. Ganek and T.A. Corbi, "The Dawning of the Autonomic Computing Era," *IBM Systems Journal* 42(1):5-18, 2003.

some parts to malfunction in a manner analogous to the healing of a biological system, and to protect itself against dangers in its (open) environment.

1.5 BACKGROUND, ORGANIZATION, AND APPROACH OF THIS REPORT

To better understand potential synergies at the BioComp interface and to facilitate the development of collaborations between scientific communities in both fields that can better exploit these synergies, the National Research Council established the Committee on Frontiers at the Interface of Computing and Biology. The committee hopes that this report will be valuable and important to a variety of interested parties and constituencies and that scientists who read it will be attracted by the excitement of research at the interface. To researchers in computer science, the committee hopes to demonstrate that biology represents an enormously rich problem domain in which their skills and talents can be of enormous value in ways that go far beyond their value as technical consultants and also that they may in turn be able to derive inspiration for solving computing problems from biological phenomena and insights. To researchers in the biological sciences, the committee hopes to show that computing and information technology have enormous value in changing the traditional intellectual paradigms of biology and allowing interesting new questions to be posed and answered. To academic administrators, the committee hopes to provide guidance and principles that facilitate the conduct of research and education at the BioComp interface. Finally, to funding agencies and organizations, the committee hopes to provide both a rationale for broadening the kinds of work they support at the BioComp interface and practices that can enhance and create links between computing and biology.

A note on terminology and scope is required for this report. Within the technology domain are a number of interconnecting aspects implied by terms such as computing, computation, modeling, computer science, computer engineering, informatics, information technology, scientific computing, and computational science. Today, there is no one term that defines the breadth of the science and technology within the computing and information sciences and technologies. The intent is to use any of these terms with a broad rather than narrow construction and connotation and to consider the entire domain of inquiry in terms of an interface to life science. For simplicity, this report uses the term "computing" to refer to intellectual domains characterized by roots in the union of the terms above.

Although the words "computing" and "computation" are used throughout this report, biology in the new millennium connects with a number of facets of the exact sciences in a way that cannot be separated from computer science per se. In particular, biology has a synergistic relationship with mathematics, statistics, physics, chemistry, engineering, and theoretical methods—including modeling and analysis as well as computation and simulation. In this relationship, blind computation is no surrogate for insight and understanding. In many cases, the fruits of computation are reaped only after careful and deliberate theoretical analysis, in which the physics, biology, and mathematics underlying a given system are carefully considered. Although much of the focus of this report is on the exchange between biology and computing, the reader should consider how the same ideas may be extended to encompass these other aspects.

Consider, for example, the fact that mathematics plays an essential role in the interpretation of experimental data and in developing algorithms for machine-assisted computing. Computing is implicitly mathematical, and as techniques for mathematical analysis evolve and develop, so will new opportunities for computing.

These points suggest that any specific limits on the range of coverage of this report are artificial and somewhat forced. Yet practicality dictates that some limits be set, and thus the committee leaves systematic coverage of certain important dimensions of the biology-computing interface to other reports. For example, a 2005 report of the Board on Mathematical Sciences (BMS) of the National Research Council (NRC) recommends a mathematical sciences research program that allows biological scientists to make the most effective use of the large amount of existing genomic information and the much larger and more diverse collections of structural and functional genomic information that are being created, covering both current research needs and some higher-risk research that might lead to innovative approaches for the future.[27] The BMS study takes a very broad look at what will be required for bioinformatics, biophysics, pattern matching, and almost anything related to the mathematical foundations of computational biology; thus, it is that BMS report, rather than the present report, that addresses analytical techniques.

Similar comments apply to the present report's coverage of medical devices based on embedded information technologies and medical informatics. Medical devices such as implanted defibrillators rely on real-time analysis of biological data to decide when to deliver a potentially lifesaving shock. Medical informatics can be regarded as computer science applied directly to problems of medicine and health care, focusing on the management of medical information, data, and knowledge for medical problem solving and decision making. Medical devices and medical informatics have many links and similarities to the subject matter of this report, but they, too, are largely outside its scope, although from time to time issues and challenges from the medical area are mentioned. Comprehensive studies describing future needs in medical informatics and medical devices must await future NRC work.

Yet another area of concern unaddressed in this report is the area of ethics associated with the issues discussed here. To ask just a few questions: Who will own DNA data? What individual biomedical data will be collected and retained? What are the ethics involved in using this data? What should individuals be told about their genetic futures? What are the ethical implications of creating new biological organisms or of changing the genetics of already living individuals? All of these questions are important, and philosophers and ethicists have begun to address some of them, but they are outside the scope of this report or the expertise of the committee.

In developing this report, the committee chose to characterize the overarching opportunities at the interface of biology and the computer and information sciences, and to highlight several diverse examples of activities at the interface. These points of intersection broadly represent and illustrate characteristics of research along the interface and include promising areas of exploration, some exciting from a basic science perspective and others from the point of view of novel applications.

Chapter 2 presents perspectives on 21st century biology, a synthesis among a variety of different intellectual approaches to biological research. Chapter 3 is a discussion of the nature of biological data and the requirements that biologists put on data.

Chapter 4 discusses computational tools for biology that help to solve specific and precisely defined problems. Chapter 5 focuses on models and simulations in biology as approaches for exploring and predicting biological phenomena.

Chapter 6 describes the value of a computational and engineering perspective in characterizing biological functionality of interest. Chapter 7 addresses roles in biological research for cyberinfrastructure and technologies for data acquisition.

Chapter 8 describes the potential of computer science applications and processes to utilize biological systems—to emulate, mimic, or otherwise draw inspiration from the organization, behavior, and structure of living things or to make use of the physical substrate of biological material in hybrid systems or other information-processing applications.

Chapter 9 presents a number of illustrative problem domains. These are technical challenges, potential future applications, and specific research questions that exemplify points along the interface of computing and biology. They illustrate the two overarching themes described in Chapter 2, and describe in detail the specific technological goals that must be met in order to successfully meet the challenge.

Chapter 10 is a discussion of the research infrastructure—people and resources need to vitalize the interface. The chapter examines the requisite scientific expertise, the false starts of the past, cultural and other barriers that must be addressed, and the coordinated effort needed to move research at the interface forward.

[27] National Research Council, *Mathematics and 21st Century Biology*, The National Academies Press, Washington, DC, 2005.

Finally, Chapter 11 summarizes key findings about opportunities and barriers to progress at the interface and provides recommendations for priority areas of research, tools, education, and resources that will propel progress at the interface.

Appendix A is a reprint of a chapter from a 1995 NRC report entitled *Calculating the Secrets of Life*. The chapter, "The Secrets of Life: A Mathematician's Introduction to Molecular Biology," is essentially a short primer on the fundamentals of molecular biology for nonbiologists. Appendix B lists some of the research challenges in computational biology discussed in other reports. Short biographies of committee members, staff, and the review coordinator are given in Appendix C.

Throughout this report, examples of relevant work are provided quite liberally where they are relevant to the topic at hand. The reader should note that these examples have generally been selected to illustrate the breadth of the topic in question, rather than to identify the most important areas of activity. That is, the appropriate spirit in which to view these examples is "letting a thousand flowers bloom," rather than one of "finding the prettiest flowers."

2

21st Century Biology

Biology, like any science, changes when technology introduces new tools that extend the scope and type of inquiry. Some changes, such as the use of the microscope, are embraced quickly and easily, because they are consonant with existing values and practices. Others, such as the introduction of multivariate statistics as performed by computers in the 1960s, are resisted, because they go against traditions of intuition, visualization, and conceptions of biology that separate it clearly from mathematics.

This chapter attempts to frame the challenges and opportunities created by the introduction of computation to the biological sciences. It does so by first briefly describing the existing threads of biological culture and practice, and then by showing how different aspects of computational science and technology can support, extend, or challenge the existing framework of biology.

Computing is only one of a large number of fields playing a role in the transformation of biology, from advanced chemistry to new fields of mathematics. And yet, in many ways, computers have proven the most challenging and the most transformative, rooted as they are in a tradition of design and abstraction so different from biology. Just as computers continue to radically change society at large, however, there is no doubt that they will change biology as well. As it has done so many times before, biology will change with this new technology, adopting new techniques, redefining what makes good science and good training, and changing which inquiries are important, valued, or even possible.

2.1 WHAT KIND OF SCIENCE?

2.1.1 The Roots of Biological Culture

Biology is a science with a deep history that can be linked to the invention of agriculture at the very dawn of civilization and, even earlier, to the first glimmerings of oral culture: “Is that safe to eat?” As such, it is a broad field, rich with culture and tradition, that encompasses many threads of observational, empirical, and theoretical research and spans scales from single molecules to continents. Such a broad field is impossible to describe simply; nevertheless, this section attempts to identify a number of the main threads of the activity and philosophy of biology.

First, biology is an *empirical* and a *descriptive* science. It is rooted in a tradition of qualitative observation and description dating back at least to Aristotle. Biological researchers have long sought to catalog the characteristics, behaviors, and variations of individual biological organisms or populations through the direct observation of organisms in their environments, rather than trying to identify general principles through mathematical or abstract modeling. For this reason, the culture of biology is both strongly visual and specific. Identifying a new species, and adequately describing its physical appearance, environment, and life cycle, remains a highly considered contribution to biological knowledge.

It is revealing to contrast this philosophy with that of modern physics, where the menagerie of new subatomic particles discovered in the 1960s and 1970s was a source of faint embarrassment and discomfort for physicists. Only with the introduction of quarks, and the subsequent reduction in the number of fundamental particles, did physicists again feel comfortable with the state of their field. Biology, in strong contrast, not only prizes and embraces the enormous diversity of life, but also considers such diversity a prime focus of study.

Second, biology is an *ontological* science, concerned with taxonomy and classification. From the time of Linnaeus, biologists have attempted to place their observations into a larger framework of knowledge, relating individual species to the identified span of life. The methodology and basis for this catalog is itself a matter of study and controversy, and so research activity of this type occurs at two levels: specific species are placed into the tree of life (or larger taxa are relocated), still a publishable event, and the science of taxonomy itself is refined.

Biology is a *historical* science. Life on Earth apparently arose just once, and all life today is derived from that single instance. A complete history of life on Earth—which lineage arose from which, and when—is one of the great, albeit possibly unachievable, goals of biology. Coupled to this inquiry, but separate, are the questions, How? and Why? What are the forces that cause species to evolve in certain ways? Are there secular trends in evolution, for example, as is often claimed, toward increasing complexity? Does evolution proceed smoothly or in bursts? If we were to "replay the tape" of evolution, would similar forms arise? Just as with taxonomy (and closely related to it), there are two levels here: what precisely happened and what the forces are that cause things to happen.

These three strands—empirical observations of a multitude of life forms, the historical facts of evolution, and the ordering of biological knowledge into an overarching taxonomy of life—served to define the central practices of biology until the 1950s and still in many ways affect the attitudes, training, philosophy, and values of the biological sciences. Although biology has expanded considerably with the advent of molecular biology, these three strands continue as vital areas of biological research and interest.

These three intellectual strands have been reflected in biological research that has been qualitative and descriptive throughout much of its early history. For example, empirical and ontological researchers have sought to catalog the characteristics, behaviors, and variations of individual biological organisms or populations through the direct observation of organisms in their environments.

Yet as important and valuable as these approaches have been for biology, they have not provided—and cannot provide—very much detail about underlying mechanisms. However, in the last half-century, an intellectual perspective provided by molecular biology and biochemistry has served as the basis for enormous leaps forward.

2.1.2 Molecular Biology and the Biochemical Basis of Life

In the past 50 years, biochemical approaches to analyzing biological questions and the overall approaches now known as molecular biology have led to the increased awareness, identification, and knowledge of the central role of certain mechanisms, such as the digital code of DNA as the mechanism underlying heredity, the use of adenosine triphospate (ATP) for energy storage, common protein signaling protocols, and many conserved genetic sequences, some shared by species as distinct as humans, sponges, and even single-cell organisms such as yeast.

This new knowledge both shaped and was shaped by changes in the practice of biology. Two important threads of biological inquiry, both existing long before the advent of molecular biology, came to the forefront in the second half of the 20th century. These threads were biological experimentation and the search for the underlying mechanics of life.

Biological experimentation and the collection of data are not new, but they acquired a new importance and centrality in the late 20th century. The identification of genes and mutations exemplified by experiments on *Drosophila* became an icon of modern biological science, and with this a new focus emerged on collecting larger amounts of quantitative data.

Biologists have always been interested in how organisms live, a question that ultimately comes down to the very definition of life. A great deal of knowledge regarding anatomy, circulation, respiration, and metabolism was gathered in the 18th and 19th centuries, but without access to the instruments and knowledge of biochemistry and molecular biology, there was a limit to what could be discovered. With molecular biology, some of the underlying mechanisms of life have been identified and analyzed quantitatively.

The effort to uncover the basic chemical features of biological processes and to ascertain all aspects of the components by way of experimental design will continue to be a major aspect of basic biological research, and much of modern biology has sought to reduce biological phenomena to the behavior of molecules.

However, biological researchers are also increasingly interested in a systems-level view in which completely novel relationships among system components and processes can be ascertained. That is, a detailed understanding of the components of a biological organism or phenomenon inevitably leads to the question of how these components interact with each other and with the environment in which the organism or phenomenon is embedded.

2.1.3 Biological Components and Processes in Context, and Biological Complexity

There is a long tradition of studying certain biological systems in context. For example, ecology has always focused on ecosystems. Physiology is another example of a life science that has generally considered biological systems as whole entities. Animal behavior and systematics science also considers biological phenomena in context. However, data acquisition technologies, computational tools, and even new intellectual paradigms are available today that enable a significantly greater degree of in-context understanding of many more biological components and processes than was previously possible, and the goal today is to span the space of biological entities from genes and proteins to networks and pathways, from organelles to cells, and from individual organisms to populations and ecosystems.

Following Kitano,[1] a systems understanding of a biological entity is based on insights regarding four dimensions: (1) system structures (e.g., networks of gene interactions and biochemical pathways and their relationship to the physical properties of intracellular and multicellular structures), (2) system dynamics (e.g., how a system behaves over time under various conditions and the mechanisms underlying specific behaviors), (3) control mechanisms (e.g., mechanisms that systematically control the state of the cell), and (4) design principles (e.g., principles underlying the construction and evolution of biological systems that have certain desirable properties).[2]

As an example, consider advances in genomic sequencing. Sequence genomics has created a path for establishing the "parts list" for living cells, but to move from isolated molecular details to a comprehensive understanding of phenomena from cell growth up to the level of homeostasis is widely recog-

[1] H. Kitano, "Systems Biology: A Brief Overview," *Science* 295(5560):1662-1664, 2002.

[2] For example, such principles might occur as the result of convergent evolution, that is, the evolution of species with different origins toward similar forms or characteristics, and an understanding of the likely ways that evolution can take to solve certain problems. Alternatively, principles might be identified that can explain the functional behavior of some specific biological system under a wide set of circumstances without necessarily being an accurate reflection of what is going on inside the system. Such principles may prove useful from the standpoint of being able to manipulate the behavior of a larger system in which the smaller system is embedded, though they may not be useful in providing a genuine understanding of the system with which they are associated.

nized as requiring a very different approach. In the highly interactive systems of living organisms, the macromolecular, cellular, and physiological processes, themselves at different levels of organizational complexity, have both temporal and spatial components. Interactions occur between sets of similar objects, such as two genes, and between dissimilar objects, such as genes and their environment.

A key aspect of biological complexity is the role of chance. One of the most salient instances of chance in biology is evolution, in which chance events affect the fidelity of genetic transmission from one generation to the next. The hand of chance is also seen in the development of an organism—chance events affect many of the details of development, though generally not the broad picture or trends. But perhaps the most striking manifestation is that individual biological organisms—even as closely related as sibling cells—are unlikely to be identical because of stochastic events from environmental input to thermal noise that affect molecular-level processes. If so, no two cells will have identical macromolecular content, and the dynamic structure and function of the macromolecules in one cell will never be the same as even a sibling cell. This fact is one of the largest distinctions between living systems and most silicon devices or almost any other manufactured or human-engineered artifact.

Put differently, the digital "code of life" embedded in DNA is far from simple. For example, the biological "parts list" that the genomic sequence makes available in principle may be unavailable in practice if all of the parts cannot be identified from the sequence. Segments of the genome once assumed to be evolutionary "junk" are increasingly recognized as the source of novel types of RNA molecules that are turning out to be major actors in cellular behavior. Furthermore, even a complete parts list provides a lot less insight into a biological system than into an engineered artifact, because human conventions for assembly are generally well understood, whereas nature's conventions for assembly are not.

A second example of the complexity is that a single gene can sometimes produce *many* proteins. In eukaryotes, for example, mRNA cannot be used as a blueprint until special enzymes first cut out the introns, or noncoding regions, and splice together the exons, the fragments that contain useful code.[3] In some cases, however, the cell can splice the exons in different ways, producing a series of proteins with various pieces added or subtracted but with the same linear ordering (these are known as splice variants). A process known as RNA editing can alter the sequence of nucleotides in the RNA after transcription from DNA but before translation into a protein, resulting in different proteins. An individual nucleotide can be changed into a different one ("substitution editing"), or nucleotides can be inserted or deleted from the RNA ("insertion-deletion editing"). In some cases (however rare), the cell's translation machinery might introduce an even more radical change by shifting its "reading frame," meaning that it starts to read the three-base-pair genetic code at a point displaced by one or two base pairs from the original. The result will be a very different sequence of amino acids and, thus, a very different protein.

Furthermore, even after the proteins are manufactured at the ribosome, they undergo quite a lot of postprocessing as they enter the various regulatory networks. Some might have their shapes and activity levels altered by the attachment, for example, of a phosphate group, a sugar molecule, or any of a variety of other appendages, while others might come together to form a multiprotein structure. In short, knowing the complete sequence of base pairs in a genome is like knowing the complete sequence of *1*s and *0*s that make up a computer program: by itself, that information does not necessarily yield insight into what the program does or how it may be organized into functional units such as subroutines.[4]

A third illustration of biological complexity is that few, if any, biological functions can be assigned to a single gene or a single protein. Indeed, the unique association between the hemoglobin molecule and the function of oxygen transport in the bloodstream is by far the exception rather than the rule.

[3] Virtually all introns are discarded by the cell, but in a few cases, an intron has been found to code—by itself—for another protein.

[4] A meaningful analogy can be drawn to the difference between object code and source code in a computer. Object code, consisting of binary digits, is what runs on the computer. Source code, usually written in a high-level programming language, is compiled into object code so that a program will run, but source code—and therefore program structure and logic—is much more comprehensible to human beings. Source code is also much more readily changed.

Much more common is the situation in which biological function depends on interactions among many biological components. A cell's metabolism, its response to chemical and biological signals from the outside, its cycle of growth and cell division—all of these functions and more are generally carried out and controlled by elaborate webs of interacting molecules.

François Jacob and Jacque Monod won the 1965 Nobel Prize in medicine for the discovery that DNA contained regulatory regions that governed the expression of individual genes.[5] (They further emphasized the importance of regulatory feedback and discussed these regulatory processes using the language of circuits, a point of relevance in Section 5.4.3.3.) Since then, it has become understood that proteins and other products of the genome interact with the DNA itself (and with each other) in a regulatory web.

For example, RNA molecules have a wide range of capabilities beyond their roles as messengers from DNA to protein. Some RNA molecules can selectively silence or repress gene transcription; others operate as a combination chemoreceptor-gene transcript ("riboswitch") that gives rise to a protein at one end of the molecule when the opposite end comes in contact with the appropriate chemical target. Indeed, it may even be that a significant increase in the number of regulatory RNAs on an evolutionary time scale is largely responsible for the increase in eukaryotic complexity without a large increase in the number of protein-coding genes. Understanding the role of RNA and other epigenetic phenomena that result in alternative states of gene expression, molecular function, or organization—"systems [that] are far more complex than any problem that molecular biology, genetics or genomics has yet approached,"[6] is critical to realizing genomics' promise.

A fourth example of biological complexity is illustrated by the fact that levels of biological complexity extend beyond the intricacies of the genome and protein structures through supramolecular complexes and organelles to cellular subsystems and assemblies of these to form often functionally polarized cells that together contribute to tissue form and function and, thereby to an organism's properties. Although the revolution of the last half of the last century in biochemistry and molecular biology has contributed significantly to our knowledge of the building blocks of life, we have only begun to scratch the surface of a data-dense and Gordian knot-like puzzle of complex and dynamic molecular interactions that give rise to the complex behaviors of organisms. In short, little is known about how the complexities of physiological processes are governed by molecular, cellular, and transcellular signaling systems and networks. Available information is deep only in limited spatial or temporal domains, and scarce in other key domains, such the middle spatial scales (e.g., 10 Å-10 μm), and there are no tools that make intelligent links between relatable pieces of scientific knowledge across these scales.

Complexity, then, appears to be an essential aspect of biological phenomena. Accordingly, the development of a coherent intellectual approach to biological complexity is required to understand systems-level interactions—of molecules, genes, cells, organisms, populations, and even ecosystems. In this intellectual universe, both "genome syntax" (the letters, words, and grammar associated with the DNA code) and "genome semantics" (what the DNA code can express and do) are central foci for investigation. Box 2.1 describes some of the questions that will arise in cell biology.

2.2 TOWARD A BIOLOGY OF THE 21ST CENTURY

A biology of the 21st century will integrate a number of diverse intellectual themes.[7] One integration is that of the reductionist and systems approaches. Where the component-centered reductionist

[5] F. Jacob and J. Monod, "Genetic Regulatory Mechanisms in the Synthesis of Proteins," *Journal of Molecular Biology* 3:318-356, 1961.

[6] F.S. Collins et al., "A Vision for the Future of Genomic Research," *Nature* 422:835-847, 2003.

[7] What this report calls 21st century biology has also been called "bringing the genome to life," an intentional biology, an integrative biology, synthetic biology, the new biology or even the next new biology, Biology 21, beyond the genome, postgenomic biology, genome-enabled science, and industrialized biology.

Box 2.1
Some Questions for Cell Biology in the 21st Century

In the Human Genome Institute's recently published agenda for research in the postgenome era, Francis Collins and his coauthors repeatedly emphasized how little biologists understand about the data already in hand. Collins et al. argue that biologists are a very long way from knowing everything there is to know about how genes are structured and regulated, for example, and they are virtually without a clue as to what's going on in the other 95 percent of the genome that does not code for genes. This is why the agenda's very first grand challenge was to systematically endow those data with meaning—that is, to "comprehensively identify the structural and functional components encoded in the human genome."[1]

The challenge, in a nutshell, is to understand the cellular information processing system—all of it—from the genome on up. Weng et al. suggest that the essential defining feature of a cell, which makes the system as a whole extremely difficult to analyze, is the following:[2]

> [The cell] is not a machine (however complex) drawn to a well-defined design, but a machine that can and does constantly rebuild itself within a range of variable parameters. For a systematic approach, what is needed is a relatively clear definition of the boundary of this variability. In principle, these boundaries are determined by an as-yet-unknown combination of intrinsic capability and external inputs. The balance between intrinsic capability and the response to external signals is likely to be a central issue in understanding gene expression. . . . A large body of emerging data indicates that early development occurs through signaling interactions that are genetically programmed, whereas at the later stages, the development of complex traits is dependent on external inputs as well. A quantitative description of this entire process would be a culmination and synthesis of much of biology.

Some of the questions raised by this perspective include the following:

- What is the proteome of any given cell? How do these individual protein molecules organize themselves into functional subnetworks—and how do these subnetworks then organize themselves into higher- and higher-level networks?[3] What are the functional design principles of these systems? And how, precisely, do the products of the genome react *back* on the genome to control their own creation?
- To what extent are active elements (such as RNA) present in the noncoding portions of the genome? What is the inventory of epigenetic mechanisms (e.g., RNA silencing, DNA methylation, histone hypoacetylation, chromatin modifications, imprinting) that cells use to control gene expression? These mechanisms play important roles in controlling an organism's development and, in some lower organisms, are defense responses against viruses and transposable elements. However, epigenetic phenomena have also been implicated in several human diseases, particularly cancer development due to the repression of tumor suppressor genes. What activates these mechanisms?
- How do these dynamically self-organizing networks vary over the course of the cell cycle (even though most cells in an organism are not proliferating and have exited from the cell cycle)? How do they change as the cell responds to its surroundings? How do they encode and process information? Also, what accounts for life's *robustness*—the ability of these networks to adapt, maintain themselves, and recover from a wide variety of environmental insults?

[1] F.S. Collins, E.D. Green, A.E. Guttmacher, and M.S. Guyer, "A Vision for the Future of Genomic Research," *Nature* 422(6934):835-847, 2003. To help achieve this grand challenge, the institute has launched the ENCODE project, a public research consortium dedicated to building an annotated encyclopedia of all known functional DNA elements. See http://www.genome.gov/10005107.

[2] G. Weng, U.S. Bhalla, and R. Iyengar, "Complexity in Biological Signaling Systems," *Science* 284(5411):92-96, 1999.

[3] The hierarchy of levels obviously doesn't stop at the cell membrane. Although deciphering the various cellular regulatory networks is a huge challenge in itself, systems biology ultimately has to deal as well with how cells organize themselves into tissues, organs, and the whole organism. One group that is trying to lay the groundwork for such an effort is the Physiome Project at the University of Auckland in New Zealand. See http://www.webopedia.com/TERM/W/Web_services.html.

- How do cells develop spatial structure? The cytoplasm is far from a uniform mixture of all of the biomolecules that exist in a cell; proteins and other macromolecules are often bound to membranes or isolated inside various cellular compartments (especially eukaryotes). A full account of the regulatory networks has to take this compartmentalization into account, along with such spatial factors as diffusion and the transport of various species through the cytoplasm and across membranes.
- How do the networks organize and reorganize themselves over the course of embryonic development, as each cell decides whether its progeny are going to become skin, muscle, brain, or whatever?[4] Then, once the cells are through differentiating, how do the networks actually vary from one cell type to the next? What constitutes the difference, and what happens to the networks as cells age or are damaged? How do flaws in the networks manifest themselves as maladies such as cancer?
- How do the networks vary between individuals? How do those variations account for differences in morphology and behavior? Also—especially in humans—how do those variations account for individual differences in the response to drugs and other therapies?
- How do multicellular organisms operate? A full account of multicellular organisms will have to include an account of signaling (in all its varieties, including cell-cell; cell-substratum; autocrine, paracrine, and exocrine signaling), cellular differentiation, cell motility, tissue architecture, and many other "community" issues.
- How do the networks vary between species? To put it another way, how have they changed over the course of evolution? Since the "blueprint" genes for proteins and RNA seem to be quite highly conserved from one species to the next, is it possible that most of evolution is the result of rearrangements in the genetic regulatory system?[5]

[4] Physiological processes such as metabolism, signal transduction, and the cell cycle take place on a time scale that ranges from milliseconds to days and are reversible in the sense that an activity flickers on, gene expression is adjusted as needed, and then everything returns to some kind of equilibrium. But the commitments that the cell makes during development are effectively *ir*reversible. Becoming a particular cell line means that the genetic regulatory networks in each successive generation of cells have to go through a cascade of decisions that end up turning genes on and off by the thousands. Unless there is some drastic intervention, as in the cloning experiments that created Dolly the Sheep, those genes are locked in place for the life span of the organism. Of course, the developmental program does not proceed in an isolated, "open-loop" fashion, as a computer scientist might say. Very early in the process, for example, the growing embryo lays out its basic body plan—front versus back, top versus bottom, and so on—by establishing embryo-wide chemical gradients, so that the concentration of the appropriate compound tells each cell what to do. Similar tricks are used at every stage thereafter: each cell is always receiving copious feedback from its neighbors, with chemical signals providing a constant stream of instructions and course corrections.

[5] After all, even very small changes in the timing of events during development, and in the rates at which various tissues grow, can have a profound impact on the final outcome.

approach is based on identifying the constituent parts of an organism and understanding the behavior of the organism in terms of the behavior of those parts (in the limit, a complete molecular-level characterization of the biological phenomena in question), systems biology aims to understand the mechanisms of a living organism across all relevant levels of hierarchy.[8] These different foci—a focus on components of biological systems versus a focus on interactions among these components—are complementary, and both will be essential for intellectual progress in the future.

Twenty-first century biology will bring together many distinct strands of biological research: taxonomic studies of many species, the enormous progress in molecular genetics, steps towards understanding the molecular mechanisms of life, and an emerging systems biology that will consider biological entities in relationship to their larger environment. Twenty-first century biology aims to understand fully the mechanisms of a living cell and the increasingly complex hierarchy of cells in metazoans, up to

[8] As a philosophical matter, the notion of reductionist explanation has had a long history in the philosophy of science. Life is composed of matter, and matter is governed by the laws of physics. So, the ultimate in reductionist explanation would suggest that life can be explained by the properties of Schrödinger's equation.

processes operating at the level of the organism and even populations and ecosystems. However, this kind of understanding is fundamentally dependent on synergies between a systems understanding as described above and the reductionist tradition.

Twenty-first century biology also brings together empirical work in biology with computational work. Empirical work is undertaken in laboratory experiments or field observations and has led to both hypothesis testing and hypothesis generation. Hypothesis testing relies on the data provided by empirical work to accept or reject a candidate hypothesis. However, data collected in empirical work can also suggest new hypotheses, leading to work that is exploratory in nature. In 21st century biology, computational work provides a variety of tools that support empirical work, but also enables much of systems biology through techniques such as simulation, data mining, and microarray analysis—and thus underlies the generation of plausible candidate hypotheses that will have to be tested. Note also that hypothesis testing is relevant to both reductionist and systems biology, in the sense that both types of biology are formulated around hypotheses (about components or about relationships between components) that may—or may not—be consistent with empirical or experimental results.

In this regard, a view expressed by Walter Gilbert in 1991 seems prescient. Gilbert noted that "in the current paradigm [i.e., that of 1991], the attack on the problems of biology is viewed as being solely experimental. The 'correct' approach is to identify a gene by some direct experimental procedure—determined by some property of its product or otherwise related to its phenotype—to clone it, to sequence it, to make its product and to continue to work experimentally so as to seek an understanding of its function." He then argued that "the new paradigm [for biological research], now emerging [i.e., in 1991], is that all the genes will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis. The actual biology will continue to be done as 'small science'—depending on individual insight and inspiration to produce new knowledge but the reagents that the scientist uses will include a knowledge of the primary sequence of the organism, together with a list of all previous deductions from that sequence."[9]

Finally, 21st century biology encompasses what is often called discovery science. Discovery science has been described as "enumerat[ing] the elements of a system irrespective of any hypotheses on how the system functions" and is exemplified by genome sequencing projects for various organisms.[10] A second example of discovery science is the effort to determine the transcriptomes and proteomes of individual cell types (e.g., quantitative measurements of all of the mRNAs and protein species).[11] Such efforts could be characterized as providing the building blocks or raw materials out of which hypotheses can be formulated—metaphorically, words of a biological "language" for expressing hypotheses. Yet even here, the Human Genome Project, while unprecedented in its scope, is comfortably part of a long tradition of increasingly fine description and cataloging of biological data.

All told, 21st century biology will entail a broad spectrum of research, from laboratory work directed by individual principal investigators, to projects on the scale of the human genome that generate large amounts of primary data, to the "mesoscience" in between that involves analytical or synthetic work conducted by multiple collaborating laboratories. For the most part, these newer research strategies involving discovery science and analytical work will complement rather than replace the traditional, relatively small laboratory focusing on complementary empirical and experimental methods.

[9] W. Gilbert, "Towards a Paradigm Shift in Biology," *Nature* 349(6305):99, 1991.

[10] R. Aebersold, L.E. Hood, and J.D. Watts, "Equipping Scientists for the New Biology," *Nature Biotechnology* 18:359, 2000.

[11] These examples are taken from T. Ideker, T. Galitski, and L. Hood, "A New Approach to Decoding Life: Systems Biology," *Annual Review of Genomics and Human Genetics* 2:343-372, 2001. The transcriptome is the complete collection of transcribed elements of the genome, including all of the genetic elements that code for proteins, all of the mRNAs, and all noncoding RNAs that are used for structural and regulatory purposes. The proteome is the complete collection of all proteins involved in a particular pathway, organelle, cell, tissue, organ, or organism that can be studied in concert to provide accurate and comprehensive data about that system.

Grand questions, such as those concerning origins of life, the story of evolution, the architecture of the brain, and the interactions of living things with each other in populations and ecosystems, are up for grabs in 21st century biology, and the applications to health, agriculture, and industry are no less ambitious. For example, 21st century biology may enable the identification of individuals who are likely to develop cancer, Alzheimer's, or other diseases, or who will respond to or have a side effect from a particular disease treatment. Pharmaceutical companies are making major investments in transcriptomics to screen for plausible drug targets. Forward-thinking companies want to develop more nutritious plants and animals, commandeer the machinery of cells to produce materials and drugs, and build interfaces to the brain to correct impaired capabilities or produce enhanced abilities. Agencies interested in fighting bioterrorism want to be able to rapidly identify the origins and ancestry of pathogen outbreaks, and stewards of natural systems would like to make better predictions about the impacts of introduced species or global change.

2.3 ROLES FOR COMPUTING AND INFORMATION TECHNOLOGY IN BIOLOGY

To manage biological data, 21st century biology will integrate discovery science, systems biology, and the empirical tradition of biological science and provide a quantitative framework within which the results of efforts in each of these areas may be placed. The availability of large amounts of biological data is expected to enable biological questions to be addressed globally, for example, examining the behavior of all of the genes in a genome, all of the proteins produced in a cell type, or all of the metabolites created under particular environmental conditions. However, enabling the answering of biological questions by uncovering the raw data is not the same as answering those questions—the data must be analyzed and used in intellectually meaningful and significant ways.

2.3.1 Biology as an Information Science

The data-intensive nature of 21st century biology underlies the dependence of biology on information technology (IT). For example, even in 1990 it was recognized that IT would play a central role in the International Human Genome Consortium for the storage and retrieval of biological gene sequence data—recording the signals, storing the sequence data, processing images of fluorescent traces specific to each base, and so on. Also, as biology unfolds in the 21st century, it is clear that the rate of production of biological data will not abate. Data acquisition opportunities will emerge in most or all life science subdisciplines and fields, and life scientists will have to cope with the coming deluge of highly multivariate, largely nonreducible data, including high-resolution imaging and time series data of complex dynamic processes.

Yet beyond data management issues, important and challenging though they are, it has also become clear that computing and information technology will play crucial roles in identifying meaningful structures and patterns in the genome (e.g., genes, genetic regulatory elements), in understanding the interconnections between various genomic elements, and in uncovering functional biological information about genes, proteins, and their interactions. This focus on information—on acquiring, processing, structuring, and representing information—places genomic studies squarely in the domain of computing and information science.

Of course, genomic studies are not the whole of modern biology. For life sciences ranging from ecology, botany, zoology, and developmental biology to cellular and molecular biology—all of which can be characterized as science with diverse data types and high degrees of data heterogeneity and hierarchy—IT is essential to collect key information and organize biological data in methodical ways in order to draw meaningful observations. Massive computing power, novel modeling approaches, new algorithms and mathematical or statistical techniques, and systematic engineering approaches will provide biologists with vital and essential tools for managing the heterogeneity and volume of the data and for extracting meaning from those data.

Ultimately, what calculus is to the language of the physical sciences, computing and information will be to the language of 21st century biology, or at least to its systems biology thread.[12] The processes of biology, the activities of living organisms, involve the usage, maintenance, dissemination, transformation or transduction, replication, and transmittal of information across generations. Biological systems are characterized by individuality, contingency, historicity, and high digital information content—every living thing is unique. Furthermore, the uniqueness and historical contingency of life means that for population-scale problems, the potential state space that the population actually inhabits is huge.[13] As an information science, the life sciences use computing and information technology as a language and a medium in which to manage the discrete, asymmetric, largely irreducible, unique nature of biological systems and observations.

In the words above, those even marginally familiar with the history of biology will recognize hints of what was once called theoretical biology or mathematical biology, which in earlier days meant models and computer simulations based on such then-fashionable ideas as cybernetics and general systems theory.[14] The initial burst of enthusiasm waned fairly quickly, as it became clear that the available experimental data were not sufficient to keep the mathematical abstractions tethered to reality. Indeed, reliable models are impossible when many or most of the quantitative values are missing. Moreover, experience since then has indicated that biological systems are much more complex and internally interlinked than had been imagined—a fact that goes a long way towards explaining why the models of that era were not very successful in driving productive hypothesis generation and research.

The story is radically different today. High-throughput data acquisition technologies (themselves enabled and made practical by today's information technologies), change a paucity of data into a deluge of it, as illustrated by the use of these technologies for sequencing of many eukaryotic organisms. This is not to say that more data are not needed, merely that the acquisition of necessary data now seems to be possible in reasonable amounts of time.

The same is true for the information technologies underpinning 21st century biology. In the past, even if data had been available, the IT then available would have been inadequate to make sense out of those data. But today's information technologies are vastly more powerful and hold considerable promise for enabling the kinds of data management and analytical capabilities that are necessary for a systems-level approach. Moreover, information technology as an underlying medium has the advantage of growing ever more capable over time at exponential rates. As information technology becomes more capable, biological applications will have an increasingly powerful technology substrate on which to draw.

[12] Biological Sciences Advisory Committee on Cyberinfrastructure for the Biological Sciences, *Building a Cyberinfrastructure for the Biological Sciences (CIBIO): 2005 and Beyond: A Roadmap for Consolidation and Exponentiation*, July 2003. Available from http://research.calit2.net/cibio/archived/CIBIO_FINAL.pdf. This is not to deny that calculus also has application in systems biology (mostly through its relevance to biochemistry and thermodynamics), but calculus is not nearly as central to systems biology as it is to the physical sciences nor as central as computing and information technology are to systems biology.

[13] The number of possible different 3-billion-base-pair genomes, assuming only simple base substitution mutations, is 4 to the 3-billionth power. That's a big number. In fact, it is so big that the ratio of that number (big) to the number of particles in the known universe (small) is much greater than the ratio of the diameter of the universe to the diameter of a carbon atom. Thus, exhaustive computer modeling of that state space is effectively precluded. Even more tractable state spaces, such as the number of different possible human haploid genotypes, still produce gigantic numbers. For example, if we assume that the entire human population is heterozygous at just 500 locations throughout the genome (a profound underestimate of existing diversity), with each site having only two states, then the number of possible human haplotypes is 2 to the 500th power, which also exceeds the number of electrons in the known universe. These back-of-the-envelope calculations also show that it is impossible for the state space of existing human genotypes to exist in anything approaching linkage equilibrium.

[14] N. Wiener, *Cybernetics, or Control and Communication in the Animal and the Machine*, 2nd ed., MIT Press, Cambridge, MA, 1961; L. von Bertalanffy, *General Systems Theory: Foundations, Development, Applications*, George Braziller, New York, 1968. This history was recently summarized in O. Wolkenhauer, "Systems Biology: The Reincarnation of Systems Theory Applied in Biology?" *Briefings in Bioinformatics* 2(3):258-270, 2001.

In short, the introduction of computing into biology has transformed, and continues to transform, the practice of biology. The most straightforward, although often intellectually challenging, way involves computing tools with which to acquire, store, process, and interpret enormous amounts of biological data. But computing (when used wisely and in combination with the tools of mathematics and physics) will also provide biologists with an alternative and possibly more appropriate language and set of abstractions for creating models and data representations of higher-order interactions, describing biological phenomena, and conceptualizing some characteristics of biological systems.

Finally, it should be noted that although computing and information technology will become an increasingly important part of life science research, researchers in different subfields of biology are likely to understand the role of computing differently. For example, researchers in molecular biology or biophysics may focus on the ability of computing to make more accurate quantitative predictions about enzyme behavior, while researchers in ecology may be more interested in the use of computing to explore relationships between ecosystem behavior and perturbations in the ambient environment. These perspectives will become especially apparent in the chapters of this report dealing with the impact of computing and IT on biology (see Chapter 4 on tools and Chapter 5 on models).

This report distinguishes between computational tools, computational models, information abstractions and a computational perspective on biology, and cyberinfrastructure and data acquisition technologies. Each of these is discussed in Chapters 4 through 7, respectively, preceded by a short chapter on the nature of biological data (Chapter 3).

2.3.2 Computational Tools

In the lexicon of this report, computational tools are artifacts—usually implemented as software, but sometimes as hardware—that enable biologists to solve very specific and precisely defined problems. For example, an algorithm for gene finding or a database of genomic sequences is a computational tool. As a rule, these tools reinforce and strengthen biological research activities, such as recording, managing, analyzing, and presenting highly heterogeneous biological data in enormous quantity. Chapter 4 focuses on computational tools.

2.3.3 Computational Models

Computational models apply to specific biological phenomena (e.g., organisms, processes) and are used for several purposes. They are used to test insight; to provide a structural framework into which observations and experimental data can be coherently inserted; to make hypotheses more rigorous, quantifiable, and testable; to help identify key or missing elements or important relationships; to help interpret experimental data; to teach or present system behavior; and to predict dynamical behavior of complex systems. Predictive models provide some confidence that certain aspects of a given biological system or phenomenon are understood, when their predictions are validated empirically. Chapter 5 focuses on computational models and simulations.

2.3.4 A Computational Perspective on Biology

Coming to grips with the complexity of biological phenomena demands an array of intellectual tools to help manage complexity and facilitate understanding in the face of such complexity. In recent years, it has become increasingly clear that many biological phenomena can be understood as performing information processing in varying degrees; thus, a computational perspective that focuses on information abstractions and functional behavior has potentially large benefit for this endeavor. Chapter 6 focuses on viewing biological phenomena through a computational lens.

2.3.5 Cyberinfrastructure and Data Acquisition

Cyberinfrastructure for science and engineering is a term coined by the National Science Foundation to refer to distributed computer, information, and communication technologies and the associated organizational facilities to support modern scientific and engineering research conducted on a global scale. Cyberinfrastructure for the life sciences is increasingly an enabling mechanism for a large-scale, data-intensive biological research effort, inherently distributed over multiple laboratories and investigators around the world, that facilitates the integration of experimental data, enables collaboration, and promotes communication among the various actors involved.

Obtaining primary biological data is a separate question. As noted earlier, 21st century biology is increasingly a data-intensive enterprise. As such, tools that facilitate acquisition of the requisite data types in the requisite amounts will become ever more important in the future. Although they are not by any means the whole story, advances in IT and computing will play key roles in the development of new data acquisition technologies that can be used in novel ways.

Chapter 7 focuses on the roles of cyberinfrastructure and data acquisition for 21st century biology.

2.4 CHALLENGES TO BIOLOGICAL EPISTEMOLOGY

The forthcoming integration of computing into biological research raises deep epistemological questions about the nature of biology itself. For many thousands of years, a doctrine known as vitalism held that the stuff of life was qualitatively different from that of nonlife and, consequently, that living organisms were made of a separate substance than nonliving things or that some separate life force existed to animate the materials that composed life.

While this belief no longer holds sway today (except perhaps in bad science fiction movies), the question of how biological phenomena can be understood has not been fully settled. One stance is based on the notion that the behavior of a given system is explained wholly by the behaviors of the components that make up that system—a view known as reductionism in the philosophy of science. A contrasting stance, known as autonomy in the philosophy of science, holds that in addition to understanding its individual components, understanding of a biological system must also include an understanding of the specific architecture and arrangement of the system's components and the interactions among them.

If autonomy is accepted as a guiding worldview, introducing the warp of computing into the weft of biology creates additional possibilities for intellectual inquiry. Just as the invention of the microscope extended biological inquiry into new arenas and enlarged the scope of questions that were reasonable to ask in the conduct of biological research, so will the computer. Computing and information technology will enable biological researchers to consider heretofore inaccessible questions, and as the capabilities of the underlying information technologies increase, such opportunities will continue to open up.

New epistemological questions will also arise. For example, as simulation becomes more pervasive and common in biology, one may ask, Are the results from a simulation equivalent to the data output of an experiment? Can biological knowledge ever arise from a computer simulation? (A practical example is the following: As large-scale clinical trials of drugs become more and more expensive, under what circumstances and to what extent might a simulation based on detailed genomic and pharmacological knowledge substitute for a large-scale trial in the drug approval process?) As simulations become more and more sophisticated, pre-loaded with more and more biological data, these questions will become both more pressing and more difficult to answer definitively.

3

On the Nature of Biological Data

Twenty-first century biology will be a data-intensive enterprise. Laboratory data will continue to underpin biology's tradition of being empirical and descriptive. In addition, they will provide confirming or disconfirming evidence for the various theories and models of biological phenomena that researchers build. Also, because 21st century biology will be a collective effort, it is critical that data be widely shareable and interoperable among diverse laboratories and computer systems. This chapter describes the nature of biological data and the requirements that scientists place on data so that they are useful.

3.1 DATA HETEROGENEITY

An immense challenge—one of the most central facing 21st century biology—is that of managing the variety and complexity of data types, the hierarchy of biology, and the inevitable need to acquire data by a wide variety of modalities. Biological data come in many types. For instance, biological data may consist of the following:[1]

- *Sequences.* Sequence data, such as those associated with the DNA of various species, have grown enormously with the development of automated sequencing technology. In addition to the human genome, a variety of other genomes have been collected, covering organisms including bacteria, yeast, chicken, fruit flies, and mice.[2] Other projects seek to characterize the genomes of all of the organisms living in a given ecosystem even without knowing all of them beforehand.[3] Sequence data generally

[1] This discussion of data types draws heavily on H.V. Jagadish and F. Olken, eds., *Data Management for the Biosciences, Report of the NSF/NLM Workshop of Data Management for Molecular and Cell Biology*, February 2-3, 2003, Available at http://www.eecs.umich.edu/~jag/wdmbio/wdmb_rpt.pdf. A summary of this report is published as H.V. Jagadish and F. Olken, "Database Management for Life Science Research," *OMICS: A Journal of Integrative Biology* 7(1):131-137, 2003.

[2] See http://www.genome.gov/11006946.

[3] See, for example, J.C. Venter, K. Remington, J.F. Heidleberg, A.L. Halpern, D. Rusch, J.A. Eisen, D. Wu, et al., "Environmental Genome Shotgun Sequencing of the Sargasso Sea," *Science* 304(5667):66-74, 2004. Venter's team collected microbial populations en masse from seawater samples originating in the Sargasso Sea near Bermuda. The team subsequently identified 1.045 billion base pairs of nonredundant sequence, which they estimated to derive from at least 1,800 genomic species based on sequence relatedness, including 148 previously unknown bacterial phylotypes. They also claimed to have identified more than 1.2 million previously unknown genes represented in these samples.

consist of text strings indicating appropriate bases, but when there are gaps in sequence data, gap lengths (or bounds on gap lengths) must be specified as well.

- *Graphs.* Biological data indicating relationships can be captured as graphs, as in the cases of pathway data (e.g., metabolic pathways, signaling pathways, gene regulatory networks), genetic maps, and structured taxonomies. Even laboratory processes can be represented as workflow process model graphs and can be used to support formal representation for use in laboratory information management systems.
- *High-dimensional data.* Because systems biology is highly dependent on comparing the behavior of various biological units, data points that might be associated with the behavior of an individual unit must be collected for thousands or tens of thousands of comparable units. For example, gene expression experiments can compare expression profiles of tens of thousands of genes, and since researchers are interested in how expression profiles vary as a function of different experimental conditions (perhaps hundreds or thousands of such conditions), what was one data point associated with the expression of one gene under one set of conditions now becomes 10^6 to 10^7 data points to be analyzed.
- *Geometric information.* Because a great deal of biological function depends on relative shape (e.g., the "docking" behavior of molecules at a potential binding site depends on the three-dimensional configuration of the molecule and the site), molecular structure data are very important. Graphs are one way of representing three-dimensional structure (e.g., of proteins), but ball-and-stick models of protein backbones provide a more intuitive representation.
- *Scalar and vector fields.* Scalar and vector field data are relevant to natural phenomena that vary continuously in space and time. In biology, scalar and vector field properties are associated with chemical concentration and electric charge across the volume of a cell, current fluxes across the surface of a cell or through its volume, and chemical fluxes across cell membranes, as well as data regarding charge, hydrophobicity, and other chemical properties that can be specified over the surface or within the volume of a molecule or a complex.
- *Patterns.* Within the genome are patterns that characterize biologically interesting entities. For example, the genome contains patterns associated with genes (i.e., sequences of particular genes) and with regulatory sequences (that determine the extent of a particular gene's expression). Proteins are characterized by particular genomic sequences. Patterns of sequence data can be represented as regular expressions, hidden Markov models (HMMs), stochastic context-free grammars (for RNA sequences), or other types of grammars. Patterns are also interesting in the exploration of protein structure data, microarray data, pathway data, proteomics data, and metabolomics data.
- *Constraints.* Consistency within a database is critical if the data are to be trustworthy, and biological databases are no exception. For example, individual chemical reactions in a biological pathway must locally satisfy the conservation of mass for each element involved. Reaction cycles in thermodynamic databases must satisfy global energy conservation constraints. Other examples of nonlocal constraints include the prohibition of cycles in overlap graphs of DNA sequence reads for linear chromosomes or in the directed graphs of conceptual or biological taxonomies.
- *Images.* Imagery, both natural and artificial, is an important part of biological research. Electron and optical microscopes are used to probe cellular and organ function. Radiographic images are used to highlight internal structure within organisms. Fluorescence is used to identify the expressions of genes. Cartoons are often used to simplify and represent complex phenomena. Animations and movies are used to depict the operation of biological mechanisms over time and to provide insight and intuitive understanding that far exceeds what is available from textual descriptions or formal mathematical representations.
- *Spatial information.* Real biological entities, from cells to ecosystems, are not spatially homogeneous, and a great deal of interesting science can be found in understanding how one spatial region is different from another. Thus, spatial relationships must be captured in machine-readable form, and other biologically significant data must be overlaid on top of these relationships.

- *Models.* As discussed in Section 5.3.4, computational models must be compared and evaluated. As the number of computational models grows, machine-readable data types that describe computational models—both the form and the parameters of the model—are necessary to facilitate comparison among models.
- *Prose.* The biological literature itself can be regarded as data to be exploited to find relationships that would otherwise go undiscovered. Biological prose is the basis for annotations, which can be regarded as a form of metadata. Annotations are critical for researchers seeking to assign meaning to biological data. This issue is discussed further in Chapter 4 (automated literature searching).
- *Declarative knowledge such as hypotheses and evidence.* As the complexity of various biological systems is unraveled, machine-readable representations of analytic and theoretical results as well as the underlying inferential chains that lead to various hypotheses will be necessary if relationships are to be uncovered in this enormous body of knowledge. This point is discussed further in Section 4.2.8.1.

In many instances, data on some biological entity are associated with many of these types: for example, a protein might have associated with it two-dimensional images, three-dimensional structures, one-dimensional sequences, annotations of these data structures, and so on.

Overlaid on these types of data is a temporal dimension. Temporal aspects of data types such as fields, geometric information, high-dimensional data, and even graphs—important for understanding dynamical behavior—multiply the data that must be managed by a factor equal to the number of time steps of interest (which may number in the thousands or tens of thousands). Examples of phenomena with a temporal dimension include cellular response to environmental changes, pathway regulation, dynamics of gene expression levels, protein structure dynamics, developmental biology, and evolution. As noted by Jagadish and Olken,[4] temporal data can be taken absolutely (i.e., measured on an absolute time scale, as might be the case in understanding ecosystem response to climate change) or relatively (i.e., relative to some significant event such as division, organism birth, or environmental insult). Note also that in complex settings such as disease progression, there may be many important events against which time is reckoned. Many traditional problems in signal processing involve the extraction of signal from temporal noise as well, and these problems are often found in investigating biological phenomena.

All of these different types of data are needed to integrate diverse witnesses of cellular behavior into a predictive model of cellular and organism function. Each data source, from high-throughput microarray studies to mass spectroscopy, has characteristic sources of noise and limited visibility into cellular function. By combining multiple witnesses, researchers can bring biological mechanisms into focus, creating models with more coverage that are far more reliable than models created from one source of data alone. Thus, data of diverse types including mRNA expression, observations of in vivo protein-DNA binding, protein-protein interactions, abundance and subcellular localization of small molecules that regulate protein function (e.g., second messengers), posttranslational modifications, and so on will be required under a wide variety of conditions and in varying genetic backgrounds. In addition, DNA sequence from diverse species will be essential to identify conserved portions of the genome that carry meaning.

3.2 DATA IN HIGH VOLUME

Data of all of the types described above contribute to an integrated understanding of multiple levels of a biological organism. Furthermore, since it is generally not known in advance how various components of an organism are connected or how they function, comprehensive datasets from each of these

[4] H.V. Jagadish and F. Olken, "Database Management for Life Science Research," *OMICS: A Journal of Integrative Biology* 7(1):131-137, 2003.

types are required. In cellular analysis, data comprehensiveness includes three aspects, as noted by Kitano:[5]

1. *Factor comprehensiveness*, which reflects the numbers of mRNA transcripts and proteins that can be measured at once;
2. *Time-line comprehensiveness*, which represents the time frame within which measurements are made (i.e., the importance of high-level temporal resolution); and
3. *Item comprehensiveness*—the simultaneous measurement of multiple items, such as mRNA and protein concentrations, phosphorylation, localization, and so forth.

For every one of the many proteins in a given cell type, information must be collected about protein identity, abundance, processing, chemical modifications, interactions, turnover time, and so forth. Spatial localization of proteins is particularly critical. To understand cellular function in detail, proteins must be localized on a scale finer than that of cell compartments; moreover, localization of specific protein assemblies to discrete subcellular sites through anchoring and scaffolding proteins is important.

All of these considerations suggest that in addition to being highly heterogeneous, biological data must be voluminous if they are to support comprehensive investigation.

3.3 DATA ACCURACY AND CONSISTENCY

All laboratories must deal with instrument-dependent or protocol-dependent data inconsistencies. For example, measurements must be calibrated against known standards, but calibration methods and procedures may change over time, and data obtained under circumstances of heterogeneous calibration may well not be comparable to each other. Experiments done by multiple independent parties almost always result in inconsistencies in datasets.[6] Different experimental runs with different technicians and protocols in different labs inevitably produce data that are not entirely consistent with each other, and such inconsistencies have to be noted and reconciled. Also, the absolute number of data errors that must be reconciled—both within a single dataset and across datasets—increases with the size of the dataset. For such reasons, statistical data analysis becomes particularly important in analyzing data acquired via high-throughput techniques.

To illustrate these difficulties, consider the replication of microarray experiments. Experience with microarrays suggests that such replication can be quite difficult. In principle, a microarray experiment is simple. The raw output of a microarray experiment is a listing of fluorescent intensities associated with spots in an array; apart from complicating factors, the brightness of these spots is an indication of the expression level of the transcript associated with them.

On the other hand, the complicating factors are many, and in some cases ignoring these factors can render one's interpretation of microarray data completely irrelevant. Consider the impact of the following:

- *Background effects*, which are by definition contributions to spot intensity that do not originate with the biological material being examined. For example, an empty microarray might result in some

[5] H. Kitano, "Systems Biology: A Brief Overview," *Science* 295(5560):1662-1664, 2002.

[6] As an example, there is only limited agreement between the datasets generated by multiple methods regarding protein-protein interactions in yeast. See, for example, the following set of papers: Y. Ho, A. Gruhler, A. Heilbut, G.D. Bader, L. Moore, S.L. Adams, A. Miller, et al., "Systematic Identification of Protein Complexes in *Saccharomyces cerevisiae* by Mass Spectrometry," *Nature* 415(6868):180-183, 2002; A.C. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, et al., "Functional Organization of the Yeast Proteome by Systematic Analysis of Protein Complexes," *Nature* 415(6868):141-147, 2002; T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, "A Comprehensive Two Hybrid Analysis to Explore the Yeast Protein Interactome," *Proceedings of the National Academy of Sciences* 98(8):4569-4574, 2001; P. Uetz, L. Giot, G. Cagney, T.A. Mansfield, R.S. Judson, J.R. Knight, D. Lockshon, et al., "A Comprehensive Analysis of Protein-Protein Interactions in *Saccharomyces cerevisiae*," *Nature* 403(6770):623-627, 2000.

background level of fluorescence and even some variation in background level across the entire surface of the array.

- *Noise dependent on expression levels of the sample*. For example, Tu et al. found that hybridization noise is strongly dependent on expression level, and in particular the hybridization noise is mostly Poisson-like for high expression levels but more complex at low expression levels.[7]
- *Differential binding strengths for different probe-target combinations*. The brightness of a spot is determined by the amount of target present at a probe site and the strength of the binding between probe and target. Held et al. found that the strength of binding is affected by the free energy of hybridization, which is itself a function of the specific sequence involved at the site, and they developed a model to account for this finding.[8]
- *Lack of correlation between mRNA levels and protein levels.* The most mature microarray technology measures mRNA levels, while the quantity of interest is often protein level. However, in some cases of interest, the correlation is small even if overall correlations are moderate. One reason for small correlations is likely to be the fact that some proteins are regulated after translation, as noted in Ideker et al.[9]
- *Lack of uniformity in the underlying glass surface of a microarray slide*. Lee et al. found that the specific location of a given probe on the surface affected the expression level recorded.[10]

Other difficulties arise when the results of different microarray experiments must be compared.[11]

- *Variations in sample preparation.* A lack of standardized procedure across experiments is likely to result in different levels of random noise—and procedures are rarely standardized very well when they are performed by humans in different laboratories. Indeed, sample preparation effects may dominate effects that arise from the biological phenomenon under investigation.[12]
- *Insufficient spatial resolution.* Because multiple cells are sampled in any microarray experiment, tissue inhomogeneities may result in more of a certain kind of cell being present, thus throwing off the final result.
- *Cell-cycle starting times.* Identical cells are likely to have more-or-less identical clocks, but there is no assurance that all of the clocks of all of the cells in a sample are started at the same time. Because expression profile varies over time, asynchrony in cell cycles may also throw off the final result.[13]

To deal with these difficulties, the advice offered by Lee et al. and Novak et al., among others, is fairly straightforward—repeat the experiment (assuming that the experiment is appropriately struc-

[7] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative Noise Analysis for Gene Expression Microarray Experiments," *Proceedings of the National Academy of Sciences* 99(22):14031-14036, 2002.

[8] G.A. Held, G. Grinstein, and Y. Tu, "Modeling of DNA Microarray Data by Using Physical Properties of Hybridization," *Proceedings of the National Academy of Sciences* 100(13):7575-7580, 2003.

[9] T. Ideker, V. Thornsson, J.A. Ranish, R. Christmas, J. Buhler, J.K. Eng, R. Bumgarner, et al., "Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network," *Science* 292(5518):929-934, 2001. (Cited in Rice and Stolovitzky, "Making the Most of It," 2004, Footnote 11.)

[10] M.L. Lee, F.C. Kuo, G.A. Whitmore, and J. Sklar, "Importance of Replication in Microarray Gene Expression Studies: Statistical Methods and Evidence from Repetitive cDNA Hybridizations," *Proceedings of the National Academy of Sciences* 97(18):9834-9839, 2000.

[11] J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2(2):70-77, 2004.

[12] J.P. Novak, R. Sladek, and T.J. Hudson, "Characterization of Variability in Large-scale Gene Expression Data: Implications for Study Design," *Genomics* 79(1):104-113, 2002.

[13] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, et al., "A Genome-wide Transcriptional Analysis of the Mitotic Cell Cycle," *Molecular Cell* 2(1):65-73, 1998; P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, et al., "Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization," *Molecular Biology of the Cell* 9(12):3273-3297, 1998. (Cited in Rice and Stolovitzky, "Making the Most of It," 2004, Footnote 11.)

tured and formulated in the first place). However, the expense of microarrays may be an inhibiting factor in this regard.

3.4 DATA ORGANIZATION

The acquiring of experimental data by some researcher is only the first step in making them useful to the wider biological research community. Data are useless if they are inaccessible or incomprehensible to others, and given the heterogeneity and large volumes of biological data, appropriate data organization is central to extracting useful information from the data. Indeed, it would not be an exaggeration to identify data management and organization issues as a key rate-limiting step in doing science for the small to medium-sized laboratory, where "science" covers the entire intellectual waterfront from laboratory experiment to data that are useful to the community at large. This is especially true in laboratories using high-throughput data acquisition technologies.

In recent years, biologists have taken significant steps in coming to terms with the need to think collectively about databases as research tools accessible to the entire community. In the field of molecular biology, the first widely recognized databases were the international archival repositories for DNA and genomic sequence information, including GenBank, the European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database, and the DNA Databank of Japan (DDJ). Subsequent databases have provided users with information that annotated the genomic sequence data, connecting regions of a genome with genes, identifying proteins associated with those genes, and assigning function to the genes and proteins. There are databases of scientific literature, such as PubMed; databases on single organisms, such as FlyBase (the *Drosophila* research database); and databases of protein interactions, such as the General Repository for Interaction Datasets (GRID). In their research, investigators typically access multiple databases (from the several hundred Web-accessible biological databases). Table 3.1 provides examples of key database resources in bioinformatics.

Data organization in biology faces significant challenges for the foreseeable future, given the levels of data being produced. Each year, workshops associated with major conferences in computational biology are held to focus on how to apply new techniques from computer science into computational biology. These include the Intelligent Systems for Molecular Biology (ISMB) Conference and the Conference on Research in Computational Biology (RECOMB), which have championed the cause of creating tools for database development and integration.[14] The long-term vision for biology is for a decentralized collection of independent and specialized databases that operate as one large, distributed information resource with common controlled vocabularies, related user interfaces, and practices. Much research will be needed to achieve this vision, but in the short term, researchers will have to make do with more specialized tools for the integration of diverse data types as described in Section 4.2.

What is the technological foundation for managing and organizing data? In 1998, Jeff Ullman noted that "the common characteristic of [traditional business databases] is that they have large amounts of data, but the operations to be performed on the data are simple," and also that under such circumstances, "the modification of the database scheme is very infrequent, compared to the rate at which queries and other data manipulations are performed."[15]

The situation in biology is the reverse. Modern information technologies can handle the volumes of data that characterize 21st century biology, but they are generally inadequate to provide a seamless integration of biological data across multiple databases, and commercial database technology has proven to have many limitations in biological applications.[16] For example, although relational databases have often been used for biological data management, they are clumsy and awkward to use in many ways.

[14] T. Head-Gordon and J. Wooley, "Computational Challenges in Structural and Functional Genomics," *IBM Systems Journal* 40(2):265-296, 2001.

[15] J.D. Ullman, *Principles of Database and Knowledge-Base Systems*, Vols. I and II, Computer Science Press, Rockville, MD, 1988.

[16] H.V. Jagadish and F. Olken, "Database Management for Life Science Research," *OMICS: A Journal of Integrative Biology* 7(1):131-137, 2003.

TABLE 3.1 Examples of Key Database Resources in Bioinformatics

Category	Databases and URLs
Comprehensive data center: broad content including sequence, structure, function, etc.	NCBI (National Center for Biotechnology and Information): http://www.ncbi.nlm.nih.gov/ EBI (European Bioinformatics Institute): http://www.ebi.ac.uk/ European Molecular Biology Laboratory (EMBL): http://www.emblheidelberg.de/ TIGR (the Institute of Genome Research): http://www.tigr.org/ Whitehead/Massachusetts Institute of Technology Genome Center: http://www-genome.wi.mit.edu/
DNA or protein sequence	GenBank: http://www.ncbi.nlm.nih.gov/Genbank DDBJ (DNA Data Bank of Japan): http://www.ddbj.nig.acjp/ EMBL Nucleotide Sequence Databank: http://www.ebi.ac.uk/embl/index.html PIR (Protein Information Resource): http://pir.georgetown.edu/ Swiss-Prot: http://www.expasy.ch/sprot/sprot-top.html
Biomolecular interactions	BIND (Biomolecular Interaction Network Database): http://www.blueprint.org/bind/bind.php The contents of BIND include high-throughput data submissions and hand-curated information gathered from the scientific literature.
Genomes: complete genome sequences and related information for specific organisms	Entrez complete genomes: http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html Complete genome at EBI: http://www.ebi.ac.uk/genomes/ University of California, Santa Cruz, Human Genome Working Draft: http://genome.ucsc.edu/ MGD (Mouse Genome Database): http://www.informaticsjax.org/ SGD (Saccharomyces Genome Database): http://genomewww.stanford.edu/Saccharomyces/ FlyBase (a database of the *Drosophila* genome): http://flybase.bio.indiana.edu/ WormBase (the genome and biology of *Caenorhabditis elegans):* http://www.wormbase.org/
Genetics: gene mapping, mutations, and diseases	GDB (Genome Database): http://gdbwww.gdb.org/gdb/ OMIM (Online Mendelian Inheritance in Man): http://www3.ncbi.nlm.nih.gov/Omim/searchomim.html HGMD (Human Gene Mutation Database): http://archive.uwcm.ac.uk/uwcm/mg/hgmdO.html

continued

TABLE 3.1 Continued

Category	Databases and URLs
Gene expression: microarray and cDNA gene expression	Unigene: http://www.ncbi.nlm.nih.gov/UniGene/
	dbEST (Expression Sequence Tag Database): http://www.ncbi.nlm.nih.gov/dbEST/index.html
	BodyMap: http://bodymap.ims.u-tokyo.ac.jp/
	GEO (Gene Expression Omnibus): http://www.ncbi.nlm.nih.gov/geo/
Structure: three-dimensional structures of small molecules, proteins, nucleic acids (both RNA and DNA) folding predictions	PDB (Protein Data Bank): http://www.rcsb.org/pdb/index.html
	NDB (Nucleic Acid Database): http://ndbserver.irutgers.edu/NDB/ndb.html
	CSD (Cambridge Structural Database): http://www.ccdc. cam. ac.uk/prods/csd/csd.html
Classification of protein family and protein domains	SCOP (Structure Classification of Proteins): http://scop.mrc-Imb.cam.ac.uk/scop/
	CATH (Protein Structure Classification Database): http://www.biochem.ucl.ac.uk/bsm/cath-new/index.html
	Pfam: http://pfam.wustl.edu/
	PROSITE database for protein family and domains: http://www.expasy.ch/prosite/
	BLOCK: http://www.blocks.fhcrc.org/
Protein pathway Protein-protein interactions and metabolic pathway	KEGG (Kyoto Encyclopedia of Genes and Genomes): http://www.genome.ad.jp/kegg/kegg2.html#pathway
	BIND (Biomolecular Interaction Network Database): http://www.binddb.org/
	DIP (Database of Interacting Proteins): http: Hdip.doe-mbi.ucla.edu/
	EcoCyc (Encyclopedia of *Escherichia coli* Genes and Metabolism): http://ecocyc.org/ecocyc/ecocyc.html
	WIT (Metabolic Pathway): http://Hwit.mcs.anl.gov/WIT2/
Proteomics: proteins, protein family	AFCS (Alliance for Cellular Signaling): http://cellularsignaling.org/
	JCSG (Joint Center for Structure Genomics): http://www.jcsg.org/scripts/prod/home.html
	PKR (Protein Kinase Resource): http://pkr.sdsc.edu/html/index.shtml

TABLE 3.1 Continued

Category	Databases and URLs
Pharmacogenomics, pharmaco genetics, single nucleotide polymorphism (SNP), genotyping	PharmGKB (Pharmacogenetics Knowledge Base): http://pharmgkb.org
	SNP Consortium: http://snp.cshl.org
	dbSNP (Single Nucleotide Polymorphism Database): http://www.ncbi.nlm.nih.gov/SNP/
	LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink
	AFRED (Allele Frequency Database): http://alfred.med.yale. edu/alfred/index.asp
	CEPH Genotype Database: http://www.cephb.fr/cephdb/
Tissues, organs, and organisms	Visible Human Project Database: http://www.nlm.nih.gov/research/visible/visible-human.html
	BRAID (Brain Image Database): http://Hbraid.rad.jhu.edu/interface.html
	NeuroDB (Neuroscience Federated Database): http://www.npaci.edu/DICE/Neuro/
	The Whole Brain Atlas: http://www.med.harvard.edu/AANLIB/home.html
Literature reference	PubMed MEDLINE: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi
	USPTO (U.S. Patent and Trademark Office): http://www.uspto.gov/

The size of biological objects is often not constant. More importantly, relational databases presume the existence of well-defined and known relationships between data records, whereas the reality of biological research is that relationships are imprecisely known—and this imprecision cannot be reduced to probabilistic measures of relationship that relational databases can handle.

Jagadish and Olken argue that without specialized life sciences enhancements, commercial relational database technology is cumbersome for constructing and managing biological databases, and most approximate sequence matching, graph queries on biopathways, and three-dimensional shape similarity queries have been performed outside of relational data management systems. Moreover, the relational data model is an inadequate abstraction for representing many kinds of biological data (e.g., pedigrees, taxonomies, maps, metabolic networks, food chains). Box 3.1 provides an illustration of how business database technology can be inadequate.

Object-oriented databases have some advantages over relational databases since the natural foci of study are in fact biological objects. Yet Jagadish and Olken note that object-oriented databases have also had limited success in providing efficient or extensible declarative query languages as required for specialized biological applications.

Because commercial database technology is of limited help, research and development of database technology that serves biological needs will be necessary. Jagadish and Olken provide a view of requirements that will necessitate further advances in data management technology, requirements that include

Box 3.1
Probabilistic One-to-Many Database Entry Linking

One purpose of database technology is the creation and maintenance of links between items in different databases. Thus, consider the problem in which a primary biological database of genes contains an object (call it A) that subsequent investigation and research reveal to be two objects. For example, what was thought to be a single gene might upon further study turn out to be two closely linked genes (A1 and A2) with a noncoding region in between (A3). Another database (e.g., a database of clones known to hybridize to various genes) may have contained a link to A—call the clone in question C. Research reveals that it is impossible for C to hybridize to both A1 and A2 individually, but that it does hybridize to the set taken collectively (i.e., A1, A2, and A3).

How should this relationship now be represented? Before the new discovery, the link was simple: C to A. Now that new knowledge requires that the primary database (or at least the entry for A) be restructured, how should this new knowledge be reflected in the original simple link? That is, what should one do with links connected to the previously single object, now that that single object has been divided into two?

The new information in the primary database has three components, A1, A2, and A3. To which of these, if any, should the original link be attached? If the link is discarded entirely, the database loses the fact that C hybridizes to the collection. If the link from C is now attached to all three equally, that link represents information contrary to fact, since experiment shows that C does not hybridize to both A1 and A2. The necessary relationship that must be reflected calls for the clone entry C to link to A1, A2, and A3 simultaneously but also probabilistically. That is, what must be represented is that the probability of the match in the set of three is one and that the probability of match for two or one in the set is zero.

As a general rule, such relationships (i.e., one-to-many relationships that are probabilistic) are not supported by business database technology. However, they are required in scientific databases once this kind of splitting operation has occurred on a hypothetical biological object—and such splitting is commonplace in scientific literature. As indicated, it can occur in the splitting of a gene, or in other cases, it can occur in the splitting of a species on the basis of additional findings on the biology of what was believed to be one species.

a great diversity of data types: sequences, graphs, three-dimensional structures, images; unconventional types of queries: similarity queries, (e.g., sequence similarity), pattern-matching queries, pattern-finding queries; ubiquitous uncertainty (and sometimes even inconsistency) in the data; data curation (data cleaning and annotation); large-scale data integration (hundreds of databases); detailed data provenance; extensive terminology management; rapid schema evolution; temporal data; and management for a variety of mathematical and statistical models of organisms and biological systems.

Data organization and management present major intellectual challenges in integration and presentation, as discussed in Chapter 4.

3.5 DATA SHARING

There is a reasonably broad consensus among scientists in all fields that reproducibility of findings is central to the scientific enterprise. One key component of reproducibility is thus the availability of data for community examination and inspection. In the words of the National Research Council (NRC) Committee on Responsibilities of Authorship in the Biological Sciences, "an author's obligation is not only to release data and materials to enable others to verify or replicate published findings but also to provide them in a form on which other scientists can build with further research."[17]

However, in practice, this ethos is not uniformly honored. An old joke in the life science research community comments on data mining in biology—"the data are mine, mine, mine." For a field whose roots are in empirical description, it is not hard to see the origins of such an attitude. For most of its history, the life sciences research community has granted primary intellectual credit to those who have collected data, a stance that has reinforced the sentiment that those that collect the data are its rightful owners. While some fields such as evolutionary biology generally have an ethos of data sharing, the data-sharing ethos is honored with much less uniformity in many other fields of biology. Requests for data associated with publications are sometimes (even often) denied, ignored, or fulfilled only after long delay or with restrictions that limit how the data may be used.[18]

The reasons for this state of affairs are multiple. The UPSIDE report called attention to the growing role of the for-profit sector (e.g., the pharmaceutical, biotechnology, research-tool, and bioinformatics companies) in basic and applied research over the last two decades, and the resulting circumstance that increasing amounts of data are developed by and held in private hands. These for-profit entities—whose primary responsibilities are to their investors—hope that their data will provide competitive advantages that can be exploited in the marketplace.

Nor are universities and other nonprofit research institutions immune to commercial pressures. An increasing amount of life sciences research in the nonprofit sector is supported directly by funds from the for-profit sector, thus increasing the prospect of potentially conflicting missions that can impede unrestricted data sharing as nonprofit researchers are caught up in commercial concerns. Universities themselves are encouraged as a matter of public law (the Bayh-Dole Act of 1980) to promote the use, commercialization, and public availability of inventions developed through federally funded research by allowing them to own the rights to patents they obtain on these inventions. University researchers also must confront the publish-or-perish issue. In particular, given the academic premiums on being first to publish, researchers are strongly motivated to take steps that will preserve their own ability to publish follow-up papers or the ability of graduate students, postdoctoral fellows, or junior faculty members to do the same.

Another contributing factor is that the nature of the data in question has changed enormously since the rise of the Human Genome Project. In particular, the enormous volumes of data collected are a continuing resource that can be productively "mined" for a long time and yield many papers. Thus, scientists who have collected such data can understandably view relinquishing control of them as a stiff penalty in light of the time, cost, and effort needed to do the research supporting the first publication.[19] Although some communities (notably the genomics, structural biology, and clinical trials communities) have established policies and practices to facilitate data sharing, other communities (e.g., those working in brain imaging or gene and protein expression studies) have not yet done so.

[17] National Research Council, *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*, National Academies Press, Washington, DC, 2003. Hereafter referred to as the UPSIDE report. Much of the discussion in Section 3.5 is based on material found in that report.

[18] For example, a 2002 survey of geneticists and other life scientists at 100 U.S. universities found that of geneticists who had asked other academic faculty for additional information, data, or materials regarding published research, 47 percent reported that at least one of their requests had been denied in the preceding 3 years. Twelve percent of geneticists themselves acknowledged denying a request from another academic researcher. See E.G. Campbell, B.R. Clarridge, M. Gokhale, L. Birenbaum, S. Hilgartner, N.A. Holtzen, and D. Blumenthal, "Data Withholding in Academic Genetics: Evidence from a National Survey," *Journal of the American Medical Association* 287(4):473-480, 2002. (Cited in the UPSIDE report; see Footnote 17.)

[19] Data provenance (the concurrent identification of the source of data along with the data itself as discussed in Section 3.7) has an impact on the social motivation to share data. If data sources are always associated with data, any work based on that data will automatically have a link to the original source; hence proper acknowledgment of intellectual credit will always be possible. Without automated data provenance, it is all too easy for subsequent researchers to lose the connection to the original source.

Finally, raw biological data are not the only commodities in question. Computational tools and models are increasingly the subject of publication in the life sciences (see Chapters 4 and 5), and it is inevitable that similar pressures will arise (indeed, have arisen) with respect to sharing the software and algorithms that underlie these artifacts. When software is at issue, a common concern is that the release of software—especially if it is released in source code—can enable another party to commercialize that code. Some have also argued that mandatory sharing of source code prevents universities from exercising their legal right to develop commercial products from federally funded research.

Considering these matters, the NRC Committee on Responsibilities of Authorship in the Biological Sciences concluded:

> The act of publishing is a quid pro quo in which authors receive credit and acknowledgment in exchange for disclosure of their scientific findings. All members of the scientific community—whether working in academia, government, or a commercial enterprise—have equal responsibility for upholding community standards as participants in the publication system, and all should be equally able to derive benefits from it.

The UPSIDE report also explicated three principles associated with sharing publication-related data and software:[20]

> - Authors should include in their publications the data, algorithms, or other information that is central or integral to the publication—that is, whatever is necessary to support the major claims of the paper and would enable one skilled in the art to verify or replicate the claims.
> - If central or integral information cannot be included in the publication for practical reasons (for example, because a dataset is too large), it should be made freely (without restriction on its use for research purposes and at no cost) and readily accessible through other means (for example, on line). Moreover, when necessary to enable further research, integral information should be made available in a form that enables it to be manipulated, analyzed, and combined with other scientific data. . . . [However, m]aking data that is central or integral to a paper freely obtainable does not obligate an author to curate and update it. While the published data should remain freely accessible, an author might make available an improved, curated version of the database that is supported by user fees. Alternatively, a value-added database could be licensed commercially.
> - If publicly accessible repositories for data have been agreed on by a community of researchers and are in general use, the relevant data should be deposited in one of these repositories by the time of publication. . . . [T]hese repositories help define consistent policies of data format and content, as well as accessibility to the scientific community. The pooling of data into a common format is not only for the purpose of consistency and accessibility. It also allows investigators to manipulate and compare datasets, synthesize new datasets, and gain novel insights that advance science.

When a publication explicitly involves software or algorithms to solve biological problems, the UPSIDE report pointed out that the principle enunciated for data should also apply: software or algorithms that are central or integral to a publication "should be made available in a manner that enables its use for replication, verification, and furtherance of science." The report also noted that one option is to provide in the publication a detailed description of the algorithm and its parameters. A second option is to make the relevant source code available to investigators who wish to test it, and either option upholds the spirit of the researcher's obligation.

Since the UPSIDE report was released in 2003, editors at two major life science journals, *Science* and *Nature*, have agreed in principle with the idea that publication entails a responsibility to make data freely available to the larger research community.[21] Nevertheless, it remains to be seen how widely the UPSIDE principles will be adopted in practice.

[20] The UPSIDE report contained five principles, but only three were judged relevant to the question of data sharing per se. The principles described in the text are quoted directly from the UPSIDE report.

[21] E. Marshall, "The UPSIDE of Good Behavior: Make Your Data Freely Available," *Science* 299(5609):990, 2003.

As for the technology to facilitate the sharing of data and models, the state of the art today is that even when the will to share is present, data or model exchange between researchers is generally a nontrivial exercise. Data and models from one laboratory or researcher must be accompanied by enough metadata that other researchers can query the data and use the model in meaningful ways without a lot of unproductive overhead in "futzing around doing stupid things." Technical dimensions of this point are discussed further in Section 4.2.

3.6 DATA INTEGRATION

As noted in Chapter 2, data are the sine qua non of biological science. The ability to share data widely increases the utility of those data to the research community and enables a higher degree of communication between researchers, laboratories, and even different subfields. Data incompatibilities can make data hard to integrate and to relate to information on other variables relevant to the same biological system. Further, when inquiries can be made across large numbers of databases, there is an increased likelihood that meaningful answers can be found. Large-scale data integration also has the salutary virtue that it can uncover inconsistencies and errors in data that are collected in disparate ways.

In digital form, all biological data are represented as bits, which are the underlying electronic representation of data. However, for these data to be useful, they must be interpretable according to some definitions. When there is a single point of responsibility for data management, the definitions are relatively easy to generate. When responsibility is distributed over multiple parties, they must agree on those definitions if the data of one party are to be electronically useful to another party. In other words, merely providing data in digital form does not necessarily mean that they can be shared readily—the semantics of differing data sets must be compatible as well.

Another complicating factor is the fact that nearly all databases—regardless of scale—have their origins in small-scale experimentation. Researchers almost always obtain relatively small amounts of data in their first attempts at experimentation. Small amounts of data can usually be managed in flat files—typically, spreadsheets. Flat files have the major advantage that they are quick and easy to implement and serve small-scale data management needs quite well.

However, flat files are generally impractical for large amounts of data. For example, queries involving multiple search criteria are hard to make when a flat-file database is involved. Relationships between entries are concealed in a flat-file format. Also, flat files are quite poor for handling heterogeneous data types.

There are a number of technologies and approaches, described below, that address such issues. In practice, however, the researcher is faced with the problem of knowing when to abandon the small-scale flat file in favor of a more capable and technically sophisticated arrangement that will inevitably entail higher overhead, at least initially.

The problem of large-scale data integration is extraordinarily complex and difficult to solve. In 2003, Lincoln Stein noted that "life would be much simpler if there was a single biological database, but this would be a poor solution. The diverse databases reflect the expertise and interests of the groups that maintain them. A single database would reflect a series of compromises that would ultimately impoverish the information resources that are available to the scientific community. A better solution would maintain the scientific and political independence of the databases, but allow the information that they contain to be easily integrated to enable cross-database queries. Unfortunately, this is not trivial."[22]

Consider, for example, what might be regarded as a straightforward problem—that of keeping straight vocabularies and terminologies and their associated concepts. In reality, when new biological structures, entities, and events have been uncovered in a particular biological context, they are often

[22] Reprinted by permission from L.D. Stein, "Integrating Biological Databases," *Nature Reviews Genetics* 4(5):337-345, 2003. Copyright 2005 Macmillan Magazines Ltd.

described with novel terminology or measurements that do not reveal much about how they might be related to similar entities in other contexts or how they quantitatively function in the contexts in which they exist, for example:

- Biological concepts may clash as users move from one database to another. Stein discusses several examples:[23]

> 1. To some research communities, "a pseudogene is a gene-like structure that contains in-frame stop codons or evidence of reverse transcription. To others, the definition of a pseudogene is expanded to include gene structures that contain full open reading frames (ORFs) but are not transcribed. Some members of the *Neisseria gonorrhea* research community, meanwhile, use pseudogene to mean a transposable cassette that is rearranged in the course of antigenic variation."
> 2. "The human genetics community uses the term allele to refer to any genomic variant, including silent nucleotide polymorphisms that lie outside of genes, whereas members of many model-organism communities prefer to reserve the term allele to refer to variants that change genes."
> 3. "Even the concept of the gene itself can mean radically different things to different research communities. Some researchers treat the gene as the transcriptional unit itself, whereas others extend this definition to include up- and downstream regulatory elements, and still others use the classical definitions of cistron and genetic complementation."

- Evolving scientific understandings may drive changes in terminology. For example, diabetes was once divided into the categories of juvenile and adult onset. As the role of insulin became clearer, the relevant categories evolved into "insulin dependent" and "non-insulin dependent." The relationship is that almost all juvenile cases of diabetes are insulin dependent, but a significant fraction of adult-onset cases are as well.
- Names of the same biological object may change across databases. "For example, consider the DNA-damage checkpoint-pathway gene that is named Rad24 in *Saccharomyces cerevisiae* (budding yeast). *[Schizo]saccharomyces pombe* (fission yeast) also has a gene named rad24 that is involved in the checkpoint pathway, but it is not the orthologue of the *S. cerevisiae* Rad24. Instead, the correct *S. pomb*e orthologue is rad17, which is not to be confused with the similarly named Rad17 gene in *S. cerevisiae*. Meanwhile, the human checkpoint-pathway genes are sometimes named after the *S. cerevisiae* orthologues, sometimes after the *S. pombe* orthologues, and sometimes have independently derived names. In *C. elegans*, there are a series of rad genes, none of which is orthologous to *S. cerevisiae* Rad17. The closest *C. elegans* match to Rad17 is, in fact, a DNA-repair gene named mrt-2."[24]
- Implicit meanings can be counterintuitive. For example, the International Classification of Disease (ICD) code for "angina" means "angina occurring in the past."[25] A condition of current angina is indicated by the code for "chest pain not otherwise specified."
- Data transformations from one database to another may destroy useful information. For example, a clinical order in a hospital may call for a "PA [posterior-anterior] and lateral chest X-ray." When that order is reflected in billing, it may be collapsed into "chest X-ray: 2 views."
- Metadata may change when databases originally created for different purposes are conceptually joined. For example, MEDLINE was developed to facilitate access to the printed paper literature by

[23] Reprinted by permission from L.D. Stein, "Integrating Biological Databases," *Nature Reviews Genetics* 4(5):337-345, 2003. Copyright 2005 Macmillan Magazines Ltd.

[24] Reprinted by permission from L.D. Stein, "Integrating Biological Databases," *Nature Reviews Genetics* 4(5):337-345, 2003. Copyright 2005 Macmillan Magazines Ltd.

[25] ICD codes refer to a standard international classification of diseases. For more information, see http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm.

scientists. The data were assembled in MEDLINE to help users find citations. As a result, authors in MEDLINE were originally treated as text strings, not as people. There was no effort, to identify individual people, so "Smith, J" could be John Smith, Jim Smith, or Joan Smith. However, the name of an individual is not necessarily constant over his or her professional lifetime. Thus, one cannot use MEDLINE to search for all papers authored by an individual who has undergone a name change without independent knowledge of the specifics of that change.

Experience suggests that left to their own devices, designers of individual databases generally make locally optimal decisions about data definitions and formats for entirely rational reasons, and local decisions are almost certain to be incompatible in some ways with other such decisions made in other laboratories by other researchers.[26] Nearly 10 years ago, Robbins noted that "a crisis occurred in the [biological] databases in the mid 1980s, when the data flow began to outstrip the ability of the database to keep up. A conceptual change in the relationship of databases to the scientific community, coupled with technical advances, solved the problem. . . . Now we face a data-integration crisis of the 1990s. Even if the various separate databases each keep up with the flow of data, there will still be a tremendous backlog in the integration of information in them. The implication is similar to that of the 1980s: either a solution will soon emerge or biological databases collectively will experience a massive failure."[27] Box 3.2 describes some of the ways in which community-wide use of biological databases continues to be difficult today.

Two examples of research areas requiring a large degree of data integration are cellular modeling and pharmacogenomics. In cellular modeling (discussed further in Section 5.4.2), researchers need to integrate the plethora of data available today about cellular function; such information includes the chemical, electrical, and regulatory features of cells; their internal pathways; mechanisms of cell motility; cell shape changes; and cell division. Box 3.3 provides an example of a cell-oriented database. In pharmacogenomics (the study of how an individual's genetic makeup affects his or her specific reaction to drugs, discussed in Section 9.7), databases must integrate data on clinical phenotypes (including both pharmacokinetic and pharmacodynamic data) and profiles (e.g., pulmonary, cardiac, and psychological function tests, and cancer chemotherapeutic side effects); DNA sequence data, gene structure, and polymorphisms in sequence (and information to track haploid, diploid, or polyploid alleles, alternative splice sites, and polymorphisms observed as common variants); molecular and cellular phenotype data (e.g., enzyme kinetic measurements); pharmacodynamic assays; cellular drug processing rates; and homology modeling of three-dimensional structures. Box 3.4 illustrates the Pharmacogenetics Research Network and Knowledge Base (PharmGKB), an important database for pharmacogenetics and pharmacogenomics.

3.7 DATA CURATION AND PROVENANCE[28]

Biological research is a fast-paced, quickly evolving discipline, and data sources evolve with it: new experimental techniques produce more and different types of data, requiring database structures to change accordingly; applications and queries written to access the original version of the schema must

[26] In particular, a scientist working on the cutting edge of a problem almost certainly requires data representations and models with more subtlety and more degrees of resolution in the data relevant to the problem than someone who has only a passing interest in that field. Almost every dataset collected has a lot of subtlety in some areas of the data model and less subtlety elsewhere. Merging these datasets into a common-denominator model risks throwing away the subtlety, where much of the value resides. Yet, merging these datasets into a uniformly data-rich model results in a database so rich that it is not particularly useful for general use. An example—biomedical databases for human beings may well include coding for gender as a variable. However, in a laboratory or medical facility that does a lot of work on transgendered individuals who may have undergone sex-change operations, the notion of gender is not necessarily as simple as "male" or "female."

[27] R.J. Robbins, "Comparative Genomics: A New Integrative Biology," in *Integrative Approaches to Molecular Biology*, J. Collado-Vides, B. Magasanik, and T.F. Smith, eds., MIT Press, Cambridge, MA, 1996.

[28] Section 3.7 embeds excerpts from S.Y. Chung and J.C. Wooley, "Challenges Faced in the Integration of Biological Information," *Bioinformatics: Managing Scientific Data*, Z. Lacroix and T. Critchlow, eds., Morgan Kaufmann, San Francisco, CA, 2003.

Box 3.2
Characteristics of Biological Databases

Biological databases have several characteristics that make them particularly difficult to use by the community at large. Biological databases are

- *Autonomous.* As a point of historical fact, most biological databases have been developed and maintained by individual research groups or research institutions. Initially, these databases were developed for individual use by these groups or institutions, and even when they proved to have value to the larger community, data management practices peculiar to those groups remained. As a result, biological databases almost always have their own governing body and infrastructure.
- *Inconsistent in format (syntax).* In addition to the heterogeneity of data types discussed in Section 3.1, databases that contain the same types of data still may be (and often are) syntactically heterogeneous. For example, the scientific literature, images, and other free-text documents are commonly stored in unstructured or semistructured formats (plain text files, HTML or XML files, binary files). Genomic, microarray gene expression, and proteomic data are routinely stored in conventional spreadsheet programs or in structured relational databases (Oracle, Sybase, DB2, Informix, etc.). Major data depository centers have also adopted different standards for data formats. For example, the U.S. National Center for Biotechnology Information (NCBI) has adopted the highly nested data ASN.1 (Abstract Syntax Notation) for the general storage of gene, protein, and genomic information, while the U.S. Department of Agriculture's Plant Genome Data and Information Center has adopted the object-oriented ACEDB data management systems and interface.
- *Inconsistent in meaning (semantics).* Biological databases containing the same types of data are also often semantically inconsistent. For example, in the database of biological literature known as MEDLINE, multiple aliases for genes are the norm, rather than the exception. There are cases in which the same name refers to different genes that have no relationship to each other. A gene that codes for an enzyme might be named according to its mutant phenotype by a geneticist and its enzymatic function by a biochemist. A vector to a molecular biologist refers to a vehicle, as in a cloning vector, whereas vector to a parasitologist is an organism that is an agent in the transmission of disease. Research groups working with different organisms will often give the same molecule a different name. Finally, biological knowledge is often represented only implicitly, in the shared assumptions of the community that produced the data source, and not explicitly via metadata that can be used either by human users or by integration software.
- *Dynamic and subject to continual change.* As biological research progresses and better understanding emerges, it is common that new data are obtained that contradict old data. Often, new data organizational schemes become necessary, even new data types or entirely new databases may become necessary.
- *Diverse in the query tools they support.* The queries supported by a database are what give the database its utility for a scientist, for only through the making of a query can the appropriate data be returned. Yet databases vary widely in the kinds of query they support—or indeed that they can support. User interfaces to query engines may require specific input and output formats. For example, BLAST (the basic local alignment search tool), the most frequently used program in the molecular biology community, requires a specific format (FASTA) for input sequence and outputs a list of pairwise sequence alignments to the end users. Output from one database query often is not suitable as direct input for a query on a different database. Finally, application semantics vary widely. Leaving aside the enormous variety of different applications for different biological problems (e.g., applications for nucleic and protein sequence analysis, genome comparison, protein structure prediction, biochemical pathway and genetic network analysis, construction of phylogenetic trees, modeling and simulation of biological systems and processes), even applications nominally designed for the same problem domain can make different assumptions about the underlying data and the meaning of answers to queries. At times, they require nontrivial domain knowledge from different fields. For example, protein folding can be approached using ab initio prediction based on first principles (physics) or using knowledge-based (computer science) threading methods.
- *Diverse in the ways they allow users to access data.* Some databases provide large text dumps of their contents, others offer access to the underlying database management system and still others provide only Web pages as their primary mode of access.

SOURCE: Derived largely from S.Y. Chung and J.C. Wooley, "Challenges Faced in the Integration of Biological Information," *Bioinformatics: Managing Scientific Data,* Z. Lacroix and T. Critchlow, eds., Morgan Kaufmann, San Francisco, CA, 2003.

Box 3.3
The Alliance for Cellular Signaling

The Alliance for Cellular Signaling (AfCS), partly supported by the National Institute of General Medical Sciences and partly by large pharmaceutical companies, seeks to build a publicly accessible, comprehensive database on cellular signaling that makes available virtually all significant information about molecules of interest. This database will also be one enabler for pathway analysis and facilitate an understanding of how molecules coordinate with one another during cellular responses. The database seeks to identify all of the proteins that constitute the various signaling systems, assess time-dependent information flow through the systems in both normal and pathological states, and reduce the mass of detailed data into a set of interacting theoretical models that describe cellular signaling. To the maximum extent possible, the information contained in the database is intended to be machine-readable.

The complete database is intended to enable researchers to:

- Query the database about complex relationships between molecules;
- View phenotype-altering mutations or functional domains in the context of protein structure;
- View or create de novo signaling pathways assembled from knowledge of interactions between molecules and the flow of information among the components of complex pathways;
- Evaluate or establish quantitative relationships among the components of complex pathways;
- View curated information about specific molecules of interest (e.g., names, synonyms, sequence information, biophysical properties, domain and motif information, protein family details, structure and gene data, the identities of orthologues and paralogues, BLAST results) through a "molecule home page" devoted to each molecule of interest, and
- Read comprehensive, peer-reviewed, expert-authored summaries, which will include highly structured information on protein states, interactions, subcellular localization, and function, together with references to the relevant literature.

The AFCS is motivated by a desire to understand as completely as possible the relationships between sets of inputs and outputs in signaling cells that vary both temporally and spatially. Yet because there are many researchers engaged in signaling research, the cultural challenge faced by the alliance is the fact that information in the database is collected by multiple researchers in different laboratories and from different organizations. Today, it involves more than 50 investigators from 20 academic and industrial institutions. However, as of this writing, it is reported that the NIGMS will reduce funding sharply for the Alliance following a mid-project review in early 2005 (see Z. Merali and J. Giles, "Databases in Peril," *Nature* 435:1010-1011, 23 June 2005).

be rewritten to match the new version. Incremental updates to data warehouses (as opposed to wholesale rebuilding of the warehouse from scratch) are difficult to accomplish efficiently, particularly when complex transformations or aggregations are involved.

A most important point is that most broadly useful databases contain both raw data and data that are either the result of analysis or derived from other databases. In this environment, databases become interdependent. Errors due to data acquisition and handling in one database can be propagated quickly into other databases. Data updated in one database may not be propagated immediately to related databases.

Thus, data curation is essential. Curation is the process through which the community of users can have confidence in the data on which they rely. So that these data can have enduring value, information related to curation must itself be stored within the database; such information is generally categorized as annotation data. Data provenance and data accuracy are central concerns, because the distinctions between primary data generated experimentally, data generated through the application of scientific

Box 3.4
The Pharmacogenetics Research Network and Knowledge Base

Supported by the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health, the Pharmacogenetics Research Network and Knowledge Base (PharmGKB) is intended as a national resource containing high-quality structured data linking genomic information, molecular and cellular phenotype information, and clinical phenotype information. The ultimate aim of this project is to produce a knowledge base that provides a public infrastructure for understanding how variations in the human genome lead to variations in clinical response to medications.

Sample inquiries to this database might include the following:

1. For gene X, show all observed polymorphisms in its sequence;
2. For drug Y, show the variability in pharmacokinetics; and
3. For phenotype Z, show the variability in association with drug Y and/or gene X.

Such queries require a database that can model key elements of the data, acquire data efficiently, provide query tools for analysis, and deliver the resulting system to the scientific community.

A central challenge for PharmGKB is that data contained it must be cross-referenced and integrated with a variety of other Web-accessible databases. Thus, PharmGKB provides mechanisms for surveillance of and integration with these databases, allowing users to submit one query with the assurance that other relevant databases are being accessed at the same time. For example, PharmGKB monitors dbSNP, the National Center for BioTechnology Information (NCBI)-supported repository for single nucleotide polymorphisms and short deletion and insertion polymorphisms. These monitoring operations search for new information about the genes of interest to the various research groups associated with the Pharmacogenetics Research Network. In addition, PharmGKB provides users with a tool for comparative genomic analysis between human and mouse that focuses on long-range regulatory elements. Such elements can be difficult to find experimentally, but are often conserved in syntenic regions between mice and humans, and may be useful in focusing polymorphism studies on noncoding areas that are more likely to be associated with detectable phenotypes.

Another important issue for the PharmGKB database is that because it contains clinical data derived from individual patients, it must have functionality that enforces the rights of those individuals to privacy and confidentiality. Thus, data flow must be limited both into and out of the knowledge base, based on evolving rules defining what can be stored in PharmGKB and what can be disseminated. No identifying information about an individual patient can be accepted into the knowledge base, and the data must be "massaged" so that patient identity cannot be reconstructed from publicly available data records.

analysis programs, and data derived from database searches are blurred. Users of databases containing these kinds of data must be concerned about where the data come from and how they are generated. A database may be a potentially rich information resource, but its value is diminished if it fails to keep an adequate description of the provenance of the data it contains.[29] Although proponents of online access

[29] P. Buneman, S. Khanna, and W.C. Tan, "Why and Where: A Characterization of Data Provenance," *8th International Conference on Database Theory (ICDT)*, pp. 316-330, 2001. Cited in Chung and Wooley, "Challenges Faced in the Integration of Biological Information," 2003, Footnote 28.

PharmGKB integrates data on clinical phenotypes (including both pharmacokinetic and pharmacodynamic data) and profiles (e.g., pulmonary, cardiac, and psychological function tests; cancer chemotherapeutic side effects), DNA sequence data, gene structure, and polymorphisms in sequence (and information to track haploid, diploid, or polyploid alleles; alternative splice sites; and polymorphisms observed as common variants), molecular and cellular phenotype data (e.g., enzyme kinetic measurements), pharmacodynamic assays, cellular drug processing rates, and homology modeling of three-dimensional structures. Figure 3.4.1 illustrates the complex relationships that are of interest for this knowledge base.

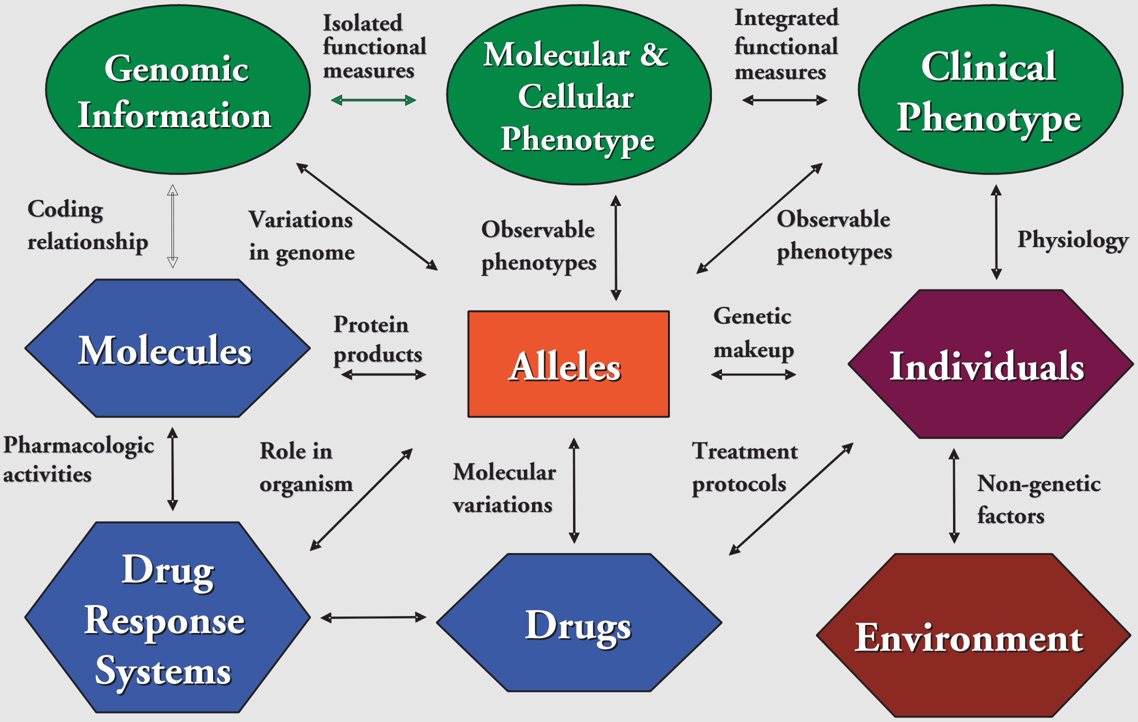


FIGURE 3.4.1 Complexity of relationships in pharmacogenetics.

SOURCE: Figure reprinted and text adapted by permission from T.E. Klein, J.T. Chang, M.K. Cho, K.L. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D.E. Oliver, D.L. Rubin, F. Shafa, J.M. Stuart, and R.B. Altman, "Integrating Genotype and Phenotype Information: An Overview of the PharmGKB Project," *The Pharmacogenomics Journal* 1:167-170, 2001. Copyright 2001 Macmillan Publishers Ltd.

to databases frequently tout it as an advantage that "the user does not need to know where the data came from or where the data are located," in fact it is essential for quality assurance reasons that the user be able to ascertain the source of all data accessed in such databases.

Data provenance addresses questions such as the following: Where did the characterization of a given GenBank sequence originate? Has an inaccurate legacy annotation been "transitively" propagated to similar sequences? What is the evidence for this annotation?

A complete record of a datum's history presents interesting intellectual questions. For example, it is difficult to justify filling a database with errata notices correcting simple errors when the actual entries can be updated. However, the original data themselves might be important, because subsequent research might have been based on them. One view is that once released, electronic database entries, like the pages of a printed journal, must stand for all time in their original condition, with errors and corrections noted only by the additional publication of errata and commentaries. However, this might quickly lead to a situation in which commentary outweighs original entries severalfold. On the other hand, occasional efforts to "improve" individual entries might inadvertently result in important information being mistakenly expunged. A middle ground might be to require that individual released entries be stable, no matter what the type of error, but that change entries be classified into different types (correction of data entry error, resubmission by original author, correction by different author, etc.), thus allowing the user to set filters to determine whether to retrieve all entries or just the most recent entry of a particular type.

To illustrate the need for provenance, consider that the output of a program used for scientific analysis is often highly sensitive to the parameters used and the specifics of the input datasets. In the case of genomic analysis, a finding that two sequences are "similar" or not may depend on the specific algorithms used and the different cutoff values used to parameterize matching algorithms, in which case other evidence is needed. Furthermore, biological conclusions derived by inference in one database will be propagated and may no longer be reliable after numerous transitive assertions. Repeated transitive assertions inevitably degrade data, whether the assertion is a transitive inference or the result of a simple "join" operation. In the absence of data perfection, additional degradation occurs with each connection.

For a new sequence that does not match any known sequence, gene prediction programs can be used to identify open reading frames, to translate DNA sequence into protein sequence, and to characterize promoter and regulatory sequence motifs. Gene prediction programs are also parameter-dependent, and the specifics of parameter settings must be retained if a future user is to make sense of the results stored in the database.

Neuroscience provides a good example of the need for data provenance. Consider the response of rat cortical cells to various stimuli. In addition to the "primary" data themselves—that is, voltages as a function of time—it is also important to record information about the rat: where the rat came from, how the rat was killed, how the brain was extracted, how the neurological preparation was made, what buffers were present, the temperature of the preparation, how much time elapsed between the sacrifice of the rat and the actual experiment being done, and so on. While all of this "extra" information seems irrelevant to the primary question, neuroscience has not advanced to the point where it is known which of these variables might have an effect on the response of interest—that is, on the evoked cortical potential.

Box 3.5 provides two examples of well-characterized and well-curated data repositories.

Finally, how far curation can be carried is an open question. The point of curation is to provide reliable and trustworthy data—what might be called biological truths. But the meaning of such "truths" may well change as more data is collected and more observations are made—suggesting a growing burden of constant editing to achieve accuracy and internal consistency. Indeed, every new entry in the database would necessarily trigger extensive validity checks of all existing entries individually and perhaps even for entries taken more than one at a time. Moreover, assertions about the real world may be initially believed, then rejected, then accepted again, albeit in a modified form. Catastrophism in geology is an example. Thus, maintaining a database of all biological truths would be an editorial nightmare, if not an outright impossibility—and thus the scope of any single database will necessarily be limited.

A database of biological observations and experimental results provides different challenges. An individual datum or result is a stand-alone contribution. Each datum or result has a recognized party responsible for it, and inclusion in the database means that it has been subject to some form of editorial review, which presumably assures its adherence to current scientific practices (and does not guarantee

Box 3.5
Two Examples of Well-Curated Data Repositories

GenBank

GenBank is a public database of all known nucleotide and protein sequences, distributed by the National Center for Biotechnology Information (NCBI), a division of the National Library of Medicine (NLM). As of January 2003, GenBank contained over 20 billion nucleotide bases in sequences from more than 55,000 species—human, mice, rat, nematode, fruit fly, and the model plant *Arabidopsis* are the most represented. GenBank and its collaborating European (EMBL) and Japanese (JPPL) databases are built with data submitted electronically by individual investigators (using BankIt or Sequin submission programs) and large-scale sequencing centers (using batch procedures). Each submission is reviewed for quality assurance and assigned an accession number; sequence updates are designated as new versions. The database is organized by a sequence-based taxonomy into divisions (e.g., bacteria, viruses, primates) and categories (e.g., expressed sequence tags, genome survey sequences, high-throughput genomic data). GenBank makes available derivative databases, for example of putative new genes, from these data.

Investigators use the Entrez retrieval system for cross-database searching of GenBank's collections of DNA, protein, and genome mapping sequence data, population sets, the NCBI taxonomy, protein structures from the Molecular Modeling Database (MMDB), and MEDLINE references (from the scientific literature). A popular tool is BLAST, the sequence alignment program, for finding GenBank sequences similar to a query sequence. The entire database is available by anonymous FTP in compressed flat-file format, updated every 2 months. NCBI offers its ToolKit to software developers creating their own interfaces and specialized analytical tools.

The Research Resource for Complex Physiologic Signals

The Research Resource for Complex Physiologic Signals was established by the National Center for Research Resources of the National Institutes of Health to support the study of complex biomedical signals. The creation of this three-part resource (PhysioBank, PhysioToolkit, and PhysioNet) overcomes long-standing barriers to hypothesis-testing research in this field by enabling access to validated, standardized data and software.[1]

PhysioBank comprises databases of multiparameter, cardiopulmonary, neural, and other biomedical signals from healthy subjects and patients with pathologies such as epilepsy, congestive heart failure, sleep apnea, and sudden cardiac death. In addition to fully characterized, multiply reviewed signal data, PhysioBank provides online access to archival data that underpin results reported in the published literature, significantly extending the contribution of that published work. PhysioBank provides theoreticians and software developers with realistic data with which to test new algorithms.

The PhysioToolkit includes software for the detection of physiologically significant events using both classic methods and novel techniques from statistical physics, fractal scaling analysis, and nonlinear dynamics; the analysis of nonstationary processes; interactive display and characterization of signals; the simulation of physiological and other signals; and the quantitative evaluation and comparison of analysis algorithms.

PhysioNet is an online forum for the dissemination and exchange of recorded biomedical signals and the software for analyzing such signals; it provides facilities for the cooperative analysis of data and the evaluation of proposed new algorithms. The database is available at http://www.physionet.org/physiobank.

[1] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, J.E. Mietus, G.B. Moody, C.K. Peng, and H.E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals," *Circulation* 101(23):E215-E220, 2000.

its absolute truth value). Without the existence of databases with differing editorial policies, some important but iconoclastic data or results might never be published. On the other hand, there is no guarantee of consistency among these data and results, which means that progress at the frontiers will depend on expert judgment in deciding which data and results will constitute the foundation from which to build.

In short, reconciling the tension between truth and diversity—both desirable, but for different reasons—is implicitly a part of the construction of every large-scale database.

4

Computational Tools

As a factual science, biological research involves the collection and analysis of data from potentially billions of members of millions of species, not to mention many trillions of base pairs across different species. As data storage and analysis devices, computers are admirably suited to the task of supporting this enterprise. Also, as algorithms for analyzing biological data have become more sophisticated and the capabilities of electronic computers have advanced, new kinds of inquiries and analyses have become possible.

4.1 THE ROLE OF COMPUTATIONAL TOOLS

Today, biology (and related fields such as medicine and pharmaceutics) are increasingly data-intensive—a trend that arguably began in the early 1960s.[1] To manage these large amounts of data, and to derive insight into biological phenomena, biological scientists have turned to a variety of computational tools.

As a rule, tools can be characterized as devices that help scientists do what they know they must do. That is, the problems that tools help solve are problems that are known by, and familiar to, the scientists involved. Further, such problems are concrete and well formulated. As a rule, it is critical that computational tools for biology be developed in collaboration with biologists who have deep insights into the problem being addressed.

The discussion below focuses on three generic types of computational tools: (1) databases and data management tools to integrate large amounts of heterogeneous biological data, (2) presentation tools that help users comprehend large datasets, and (3) algorithms to extract meaning and useful information from large amounts of data (i.e., to find meaningful a signal in data that may look like noise at first glance). (Box 4.1 presents a complementary view of advances in computer sciences needed for next-generation tools for computational biology.)

[1] The discussion in Section 4.1 is derived in part from T. Lenoir, "Shaping Biomedicine as an Information Science," *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, M.E. Bowden, T.B. Hahn, and R.V. Williams, eds., ASIS Monograph Series, Information Today, Inc., Medford, NJ, 1999, pp. 27-45.

Box 4.1
Tool Challenges for Computer Science

Data Representation

- Next-generation genome annotation system with accuracy equal to or exceeding the best human predictions
- Mechanism for multimodal representation of data

Analysis Tools

- Scalable methods of comparing many genomes
- Tools and analyses to determine how molecular complexes work within the cell
- Techniques for inferring and analyzing regulatory and signaling networks
- Tools to extract patterns in mass spectrometry datasets
- Tools for semantic interoperability

Visualization

- Tools to display networks and clusters at many levels of detail
- Approaches for interpreting data streams and comparing high-throughput data with simulation output

Standards

- Good software-engineering practices and standard definitions (e.g., a common component architecture)
- Standard ontology and data-exchange format for encoding complex types of annotation

Databases

- Large repository for microbial and ecological literature relevant to the "Genomes to Life" effort.
- Big relational database derived by automatic generation of semantic metadata from the biological literature
- Databases that support automated versioning and identification of data provenance
- Long-term support of public sequence databases

SOURCE: U.S. Department of Energy, *Report on the Computer Science Workshop for the Genomes to Life Program*, Gaithersburg, MD, March 6-7, 2002; available at http://DOEGenomesToLife.org/compbio/.

These examples are drawn largely from the area of cell biology. The reason is not that these are the only good examples of computational tools, but rather that a great deal of the activity in the field has been the direct result of trying to make sense out of the genomic sequences that have been collected to date. As noted in Chapter 2, the Human Genome Project—completed in draft in 2000—is arguably the first large-scale project of 21st century biology in which the need for powerful information technology was manifestly obvious. Since then, computational tools for the analysis of genomic data, and by extension data associated with the cell, have proliferated wildly; thus, a large number of examples are available from this domain.

4.2 TOOLS FOR DATA INTEGRATION[2]

As noted in Chapter 3, data integration is perhaps the most critical problem facing researchers as they approach biology in the 21st century.

[2] Sections 4.2.1, 4.2.4, 4.2.6, and 4.2.8 embed excerpts from S.Y. Chung and J.C. Wooley, "Challenges Faced in the Integration of Biological Information," in *Bioinformatics: Managing Scientific Data,* Z. Lacroix and T. Critchlow, eds., Morgan Kaufmann, San Francisco, CA, 2003. (Hereafter cited as Chung and Wooley, 2003.)

4.2.1 Desiderata

If researcher A wants to use a database kept and maintained by researcher B, the "quick and dirty" solution is for researcher A to write a program that will translate data from one format into another. For example, many laboratories have used programs written in Perl to read, parse, extract, and transform data from one form into another for particular applications.[3] Depending on the nature of the data involved and the structure of the source databases, writing such a program may require intensive coding.

Although such a fix is expedient, it is not scalable. That is, point-to-point solutions are not sustainable in a large community in which it is assumed that everyone wants to share data with everyone else. More formally, if there are N data sources to be integrated, and point-to-point solutions must be developed, $N (N - 1)/2$ translation programs must be written. If one data source changes (as is highly likely), $N - 1$ programs must be updated.

A more desirable approach to data integration is scalable. That is, a change in one database should not necessitate a change on the part of every research group that wants to use those data. A number of approaches are discussed below, but in general, Chung and Wooley argue that robust data integration systems must be able to

1. Access and retrieve relevant data from a broad range of disparate data sources;
2. Transform the retrieved data into a common data model for data integration;
3. Provide a rich common data model for abstracting retrieved data and presenting integrated data objects to the end-user applications;
4. Provide a high-level expressive language to compose complex queries across multiple data sources and to facilitate data manipulation, transformation, and integration tasks; and
5. Manage query optimization and other complex issues.

Sections 4.2.2, 4.2.4, 4.2.5, 4.2.6, and 4.2.8 address a number of different approaches to dealing with the data integration problem. These approaches are not, in general, mutually exclusive, and they may be usable in combination to improve the effectiveness of a data integration solution.

Finally, biological databases are always changing, so integration is necessarily an ongoing task. Not only are new data being integrated within the existing database structure (a structure established on the basis of an existing intellectual paradigm), but biology is a field that changes quickly—thus requiring structural changes in the databases that store data. In other words, biology does not have some "classical core framework" that is reliably constant. Thus, biological paradigms must be redesigned from time to time (on the scale of every decade or so) to keep up with advances, which means that no "gold standards" to organize data are built into biology. Furthermore, as biology expands its attention to encompass complexes of entities and events as well as individual entities and events, more coherent approaches to describing new phenomena will become necessary—approaches that bring some commonality and consistency to data representations of different biological entities—so that relationships between different phenomena can be elucidated.

As one example, consider the potential impact of "-omic" biology, biology that is characterized by a search for data completeness—the complete sequence of the human genome, a complete catalog of proteins in the human body, the sequencing of all genomes in a given ecosystem, and so on. The possibility of such completeness is unprecedented in the history of the life sciences and will almost certainly require substantial revisions to the relevant intellectual frameworks.

[3] The Perl programming language provides powerful and easy-to-use capabilities to search and manipulate text files. Because of these strengths, Perl is a major component of much bioinformatics programming. At the same time, Perl is regarded by many computer scientists as an unsafe language in which it is easy to make programs do dangerous things. In addition, many regard the syntax and structure of most Perl programs to be of a nature that is hard to understand much after the fact.

4.2.2 Data Standards

One obvious approach to data integration relies on technical standards that define representations of data and hence provide an understanding of data that is common to all database developers. For obvious reasons, standards are most relevant to future datasets. Legacy databases, which have been built around unique data definitions, are much less amenable to a standards-driven approach to data integration.

Standards are indeed an essential element of efforts to achieve data integration of future datasets, but the adoption of standards is a nontrivial task. For example, community-wide standards for data relevant to a certain subject almost certainly differ from those that might be adopted by individual laboratories, which are the focus of the "small-instrument, multi-data-source" science that characterizes most public-sector biological research.

Ideally, source data from these projects flow together into larger national or international data resources that are accessible to the community. Adopting community standards, however, entails local compromises (e.g., nonoptimal data structuring and semantics, greater expense), and the budgets that characterize small-instrument, single-data-source science generally do not provide adequate support for local data management and usually no support at all for contributions to a national data repository.

If data from such diverse sources are to be maintained centrally, researchers and laboratories must have incentives and support to adopt broader standards in the name of the community's greater good. In this regard, funding agencies and journals have considerable leverage and through techniques such as requiring researchers to deposit data in conformance to community standards may be able to provide such incentives.

At the same time, data standards cannot resolve the integration problem by themselves even for future datasets. One reason is that in some fast-moving and rapidly changing areas of science (such as biology), it is likely that the data standards existing at any given moment will not cover some new dimension of data. A novel experiment may make measurements that existing data standards did not anticipate. (For example, sequence databases—by definition—do not integrate methylation data; and yet methylation is an essential characteristic of DNA that falls outside primary sequence information.) As knowledge and understanding advance, the meaning attached to a term may change over time. A second reason is that standards are difficult to impose on legacy systems, because legacy datasets are usually very difficult to convert to a new data standard and conversion almost always entails some loss of information.

As a result, data standards themselves must evolve as the science they support changes. Because standards cannot be propagated instantly throughout the relevant biological community, database A may be based on Version 12.1 of a standard, and database B on Version 12.4 of the "same" standard. It would be desirable if the differences between Versions 12.1 and 12.4 were not large and a basic level of integration could still be maintained, but this is not ensured in an environment in which options vary within standards, different releases and versions of products, and so on. In short, much of the devil of ensuring data integration is in the detail of implementation.

Experience in the database world suggests that standards gaining widespread acceptance in the commercial marketplace tend to have a long life span, because the marketplace tends to weed out weak standards before they become widely accepted. Once a standard is widely used, industry is often motivated to maintain compliance with this accepted standard, but standards created by niche players in the market tend not to survive. This point is of particular relevance in a fragmented research environment and suggests that standards established by strong consortia of multiple players are more likely to endure.

4.2.3 Data Normalization[4]

An important issue related to data standards is data normalization. Data normalization is the process through which data taken on the "same" biological phenomenon by different instruments, procedures, or researchers can be rendered comparable. Such problems can arise in many different contexts:

[4] Section 4.2.3 is based largely on a presentation by C. Ball, "The Normalization of Microarray Data," presented at the AAAS 2003 meeting in Denver, Colorado.

- Microarray data related to a given cell may be taken by multiple investigators in different laboratories.
- Ecological data (e.g., temperature, reflectivity) in a given ecosystem may be taken by different instruments looking at the system.
- Neurological data (e.g., timing and amplitudes of various pulse trains) related to a specific cognitive phenomenon may be taken on different individuals in different laboratories.

The simplest example of the normalization problem is when different instruments are calibrated differently (e.g., a scale in George's laboratory may not have been zeroed properly, rendering mass measurements from George's laboratory noncomparable to those from Mary's laboratory). If a large number of readings have been taken with George's scale, one possible fix (i.e., one possible normalization) is to determine the extent of the zeroing required and to add or subtract that correction to the already existing data. Of course, this particular procedure assumes that the necessary zeroing was constant for each of George's measurements. The procedure is not valid if the zeroing knob was jiggled accidentally after half of the measurements had been taken.

Such biases in the data are systematic. In principle, the steps necessary to deal with systematic bias are straightforward. The researcher must avoid it as much as possible. Because complete avoidance is not possible, the researcher must recognize it when it occurs and then take steps to correct for it. Correcting for bias entails determining the magnitude and effect of the bias on data that have been taken and identifying the source of the bias so that the data already taken can be modified and corrected appropriately. In some cases, the bias may be uncorrectable, and the data must be discarded.

However, in practice, dealing with systematic bias is not nearly so straightforward. Ball notes that in the real world, the process goes something like this:

1. Notice something odd with data.
2. Try a few methods to determine magnitude.
3. Think of many possible sources of bias.
4. Wonder what in the world to do next.

There are many sources of systematic bias, and they differ depending on the nature of the data involved. They may include effects due to instrumentation, sample (e.g., sample preparation, sample choice), or environment (e.g., ambient vibration, current leakage, temperature). Section 3.3 describes a number of the systematic biases possible in microarray data, as do several references provided by Ball.[5]

There are many ways to correct for systematic bias, depending on the type of data being corrected. In the case of microarray studies, these ways include use of dye swap strategies, replicates and reference samples, experimental controls, consistent techniques, and sensible array and experiment design. Yet all

[5] Ball's AAAS presentation includes the following sources: T.B. Kepler, L. Crosby, and K.T. Morgan, "Normalization and Analysis of DNA Microarray Data by Self-consistency and Local Regression," *Genome Biololgy* 3(7), RESEARCH0037.1- RESEARCH0037.12, 2002. Available at http://genomebiology.org/2002/3/7/research/0037.1; R. Hoffmann, T. Seidl, M. Dugas. "Profound Effect of Normalization on Detection of Differentially Expressed Genes in Oligonucleotide Microarray Data Analysis," *Genome Biology* 3(7):RESEARCH0033.1-RESEARCH0033.1-11. Available at http://genomebiology.com/2002/3/7/research/0033; C. Colantuoni, G. Henry, S. Zeger, and J. Pevsner, "Local Mean Normalization of Microarray Element Signal Intensities Across an Array Surface: Quality Control and Correction of Spatially Systematic Artifacts," *Biotechniques* 32(6):1316-1320, 2002; B.P. Durbin, J.S. Hardin, D.M. Hawkins, and D.M. Rocke, "A Variance-Stabilizing Transformation for Gene-Expression Microarray Data," *Bioinformatics* 18 (Suppl. 1):S105-S110, 2002; P.H. Tran, D.A. Peiffer, Y. Shin, L.M. Meek, J.P. Brody, and K.W. Cho, "Microarray Optimizations: Increasing Spot Accuracy and Automated Identification of True Microarray Signals," *Nucleic Acids Research* 30(12):e54, 2002, available at http://nar.oupjournals.org/cgi/content/full/30/12/e54; M. Bilban, L.K. Buehler, S. Head, G. Desoye, and V. Quaranta, "Normalizing DNA Microarray Data," *Current Issues in Molecular Biology* 4(2):57-64, 2002; J. Quackenbush, "Microarray Data Normalization and Transformation," *Nature Genetics Supplement* 32:496-501, 2002.

of these approaches are labor-intensive, and an outstanding challenge in the area of data normalization is to develop approaches to minimize systematic bias that demand less labor and expense.

4.2.4 Data Warehousing

Data warehousing is a centralized approach to data integration. The maintainer of the data warehouse obtains data from other sources and converts them into a common format, with a global data schema and indexing system for integration and navigation. Such systems have a long track record of success in the commercial world, especially for resource management functions (e.g., payroll, inventory). These systems are most successful when the underlying databases can be maintained in a controlled environment that allows them to be reasonably stable and structured. Data warehousing is dominated by relational database management systems (RDBMS), which offer a mature and widely accepted database technology and a standard high-level standard query language (SQL).

However, biological data are often qualitatively different from the data contained in commercial databases. Furthermore, biological data sources are much more dynamic and unpredictable, and few public biological data sources use structured database management systems. Data warehouses are often troubled by a lack of synchronization between the data they hold and the original database from which those data derive because of the time lag involved in refreshing the data warehouse store. Data warehousing efforts are further complicated by the issue of updates. Stein writes:[6]

> One of the most ambitious attempts at the warehouse approach [to database integration] was the Integrated Genome Database (IGD) project, which aimed to combine human sequencing data with the multiple genetic and physical maps that were the main reagent for human genomics at the time. At its peak, IGD integrated more than a dozen source databases, including GenBank, the Genome Database (GDB) and the databases of many human genetic-mapping projects. The integrated database was distributed to end-users complete with a graphical front end. . . . The IGD project survived for slightly longer than a year before collapsing. The main reason for its collapse, as described by the principal investigator on the project (O. Ritter, personal communication, as relayed to Stein), was the database churn issue. On average, each of the source databases changed its data model twice a year. This meant that the IGD data import system broke down every two weeks and the dumping and transformation programs had to be rewritten—a task that eventually became unmanageable.

Also, because of the breadth and volume of biological databases, the effort involved in maintaining a comprehensive data warehouse is enormous—and likely prohibitive. Such an effort would have to integrate diverse biological information, such as sequence and structure, up to the various functions of biochemical pathways and genetic polymorphisms.

Still, data warehousing is a useful approach for specific applications that are worth the expense of intense data cleansing to remove potential errors, duplications, and semantic inconsistency.[7] Two current examples of data warehousing are GenBank and the International Consortium for Brain Mapping (ICBM) (the latter is described in Box 4.2).

4.2.5 Data Federation

The data federation approach to integration is not centralized and does not call for a "master" database. Data federation calls for scientists to maintain their own specialized databases encapsulating their particular areas of expertise and retain control of the primary data, while still making it available to other researchers. In other words, the underlying data sources are autonomous. Data federation often

[6] Reprinted by permission from L.D. Stein, "Integrating Biological Databases," *Nature Reviews Genetics* 4(5)337-345, 2003. Copyright 2005 Macmillan Magazines Ltd.

[7] R. Resnick, "Simplified Data Mining," pp. 51-52 in *Drug Discovery and Development,* 2000. (Cited in Chung and Wooley, 2003.)

Box 4.2
The International Consortium for Brain Mapping (ICBM): A Probabilistic Atlas and Reference System for the Human Brain

In the human population, the brain varies structurally and functionally in currently undefined ways. It is clear that the size, shape, symmetry, folding pattern, and structural relationships of the systems in the human brain vary from individual to individual. This has been a source of considerable consternation and difficulty in research and clinical evaluations of the human brain from both the structural and the functional perspective. Current atlases of the human brain do not address this problem. Cytoarchitectural and clinical atlases typically use a single brain or even a single hemisphere as the reference specimen or target brain to which other brains are matched, typically with simple linear stretching and compressing strategies. In 1992, John Mazziotta and Arthur Toga proposed the concept of developing a probabilistic atlas from a large number of normal subjects between the ages of 18 and 90. This data acquisition has now been completed, and the value of such an atlas is being realized for both research and clinical purposes. The mathematical and software machinery required to develop this atlas of normal subjects is now also being applied to patient populations including individuals with Alzheimer's disease, schizophrenia, autism, multiple sclerosis, and others.

Talairach Atlas

To date, more than 7,000 normal subjects have been entered into the Talairach atlas project and a wide range of datasets. These datasets contain detailed demographic histories of the subjects, results of general medical and neurological examinations, neuropsychiatric and neuropsychological evaluations, quantitative "handedness measurements", and imaging studies. The imaging studies include multispectra 1 mm^3 voxel-size magnetic resonance imaging (MRI) evaluations of the entire brain (T_1, T_2, and proton density pulse sequences). A subset of individuals also undergo functional MRI, cerebral blood flow position emission tomography (PET) and electroencephalogram (EEG) examinations (evoked potentials). Of these subjects, 5,800 individuals have also had their DNA collected and stored for future genotyping. As such, this database represents the most comprehensive evaluation of the structural and functional imaging phenotypes of the human brain in the normal population across a wide age span and very diverse social, economic, and racial groups. Participating laboratories are widely distributed geographically from Asia to Scandinavia, and include eight laboratories, in seven countries, on four continents.

World Map of Sites

A component of the World Map of Sites project involves the post mortem MRI imaging of individuals who have willed their bodies to science. Subsequent to MRI imaging, the brain is frozen and sectioned at a resolution of approximately 100 microns. Block face images are stored, and the sectioned tissue is stained for cytoarchitectural, chemoarchitectural, and differential myelin to produce microscopic maps of cellular anatomy, neuroreceptor or transmitter systems, and white matter tracts. These datasets are then incorporated into a target brain to which the in vivo brain studies are warped in three dimensions and labeled automatically. The 7,000 datasets are then placed in the standardized space, and probabilistic estimates of structural boundaries, volumes, symmetries, and shapes are computed for the entire population or any subpopulation (e.g., age, gender, race). In the current phase of the program, information is being added about in vivo chemoarchitecture (5-HT$_{2A}$ [5-hydroxytryptamine-2A] in vivo PET receptor imaging), in vivo white matter tracts (MRI-diffusion tensor imaging), vascular anatomy (magnetic resonance angiography and venography), and cerebral connections (transcranial magnetic stimulation-PET cerebral blood flow measurements).

Target Brain

The availability of 342 twin pairs in the dataset (half monozygotic and half dizygotic) along with DNA for genotyping provides the opportunity to understand structure-function relationships related to genotype and, therefore, provides the first large-scale opportunity to relate phenotype-genotype in behavior across a wide range of individuals in the human population.

continued

Box 4.2 Continued

The development of similar atlases to evaluate patients with well-defined disease states allows the opportunity to compare the normal brain with brains of patients having cerebral pathological conditions, thereby potentially leading to enhanced clinical trials, automated diagnoses, and other clinical applications. Such examples have already emerged in patients with multiple sclerosis and epilepsy. An example in Alzheimer's disease relates to a current hotly contested research question. Individuals with Alzheimer's disease have a greater likelihood of having the genotype ApoE 4 (as opposed to ApoE 2 or 3). Having this genotype, however, is neither sufficient nor required for the development of Alzheimer's disease. Individuals with Alzheimer's disease also have small hippocampi, presumably because of atrophy of this structure as the disease progresses. The question of interest is whether individuals with the high-risk genotype (ApoE 4) have small hippocampi to begin with. This would be a very difficult hypothesis to test without the dataset described above. With the ICBM database, it is possible to study individuals from, for example, ages 20 to 40 and identify those with the smallest (lowest 5 percent) and largest (highest 5 percent) hippocampal volumes. This relatively small number of subjects could then be genotyped for ApoE alleles. If individuals with small hippocampi all had the genotype ApoE 4 and those with large hippocampi all had the genotype ApoE 2 or 3, this would be strong support for the hypothesis that individuals with the high-risk genotype for the development of Alzheimer's disease have small hippocampi based on genetic criteria as a prelude to the development of Alzheimer's disease. Similar genotype-imaging phenotype evaluations could be undertaken across a wide range of human conditions, genotypes, and brain structures.

SOURCE: Modified from John C. Mazziotta and Arthur W. Toga, Department of Neurology, David Geffen School of Medicine, University of California, Los Angeles, personal communication to John Wooley, February 22, 2004.

calls for the use of object-oriented concepts to develop data definitions, encapsulating the internal details of the data associated with the heterogeneity of the underlying data sources.[8] A change in the representation or definition of the data then has minimal impact on the applications that access those data.

An example of a data federation environment is BioMOBY, which is based on two ideas.[9] The first is the notion that databases provide bioinformatics services that can be defined by their inputs and outputs. (For example, BLAST is a service provided by GenBank that can be defined by its input—that is, an uncharacterized sequence—and by its output, namely, described gene sequences deposited in GenBank.) The second idea is that all database services would be linked to a central registry (MOBY Central) of services that users (or their applications) would query. From MOBY Central, a user could move from one set of input-output services to the next—for example, moving from one database that, given a sequence (the input), postulates the identity of a gene (the output), and from there to a database that, given a gene (the input), will find the same gene in multiple organisms (the output), and so on, picking up information as it moves through database services. There are limitations to the BioMOBY system's ability to discriminate database services based the descriptions of inputs and outputs, and MOBY Central must be up and running 24 hours a day.[10]

[8] R.G.G. Cattell, *Object Data Management: Object-Oriented and Extended Relational Database Systems*, revised edition, Addison-Wiley, Reading, MA, 1994. (Cited in Chung and Wooley, 2003.)

[9] M.D. Wilkinson and M. Links, "BioMOBY: An Open-Source Biological Web Services Proposal," *Briefings In Bioinformatics* 3(4):331-341, 2002.

[10] L.D. Stein, "Integrating Biological Databases," *Nature Reviews Genetics* 4(5):337-345, 2003.

4.2.6 Data Mediators/Middleware

In the middleware approach, an intermediate processing layer (a "mediator") decouples the underlying heterogeneous, distributed data sources and the client layer of end users and applications.[11] The mediator layer (i.e., the middleware) performs the core functions of data transformation and integration, and communicates with the database "wrappers" and the user application layer. (A "wrapper" is a software component associated with an underlying data source that is generally used to handle the tasks of access to specified data sources, extraction and retrieval of selected data, and translation of source data formats into a common data model designed for the integration system.)

The common model for data derived from the underlying data sources is the responsibility of the mediator. This model must be sufficiently rich to accommodate various data formats of existing biological data sources, which may include unstructured text files, semistructured XML and HTML files, and structured relational, object-oriented, and nested complex data models. In addition, the internal data model must facilitate the structuring of integrated biological objects to present to the user application layer. Finally, the mediator also provides services such as filtering, managing metadata, and resolving semantic inconsistency in source databases.

There are many flavors of mediator approaches in life science domains. IBM's DiscoveryLink for the life sciences is one of the best known.[12] The Kleisli system provides an internal nested complex data model and a high-power query and transformation language for data integration.[13] K2 shares many design principles with Kleisli in supporting a complex data model, but adopts more object-oriented features.[14] OPM supports a rich object model and a global schema for data integration.[15] TAMBIS provides a global ontology (see Section 4.2.8 on ontologies) to facilitate queries across multiple data sources.[16] TSIMMIS is a mediation system for information integration with its own data model (Object-Exchange Model, OEM) and query language.[17]

4.2.7 Databases as Models

A natural progression for databases established to meet the needs and interests of specialized communities, such as research on cell signaling pathways or programmed cell death, is the evolution of

[11] G. Wiederhold, "Mediators in the Architecture of Future Information Systems," *IEEE Computer* 25(3):38-49, 1992; G. Wiederhold and M. Genesereth, "The Conceptual Basis for Mediation Services," *IEEE Expert, Intelligent Systems and Their Applications* 12(5):38-47, 1997. (Both cited in Chung and Wooley, 2003.)

[12] L.M. Haas et al., "DiscoveryLink: A System for Integrated access to Life Sciences Data Sources," *IBM Systems Journal* 40(2):489-511, 2001.

[13] S. Davidson, C. Overton, V. Tannen, and L. Wong, "BioKleisli: A Digital Library for Biomedical Researchers," *International Journal of Digital Libraries* 1(1):36-53, 1997; L. Wong, "Kleisli, a Functional Query System," *Journal of Functional Programming* 10(1):19-56, 2000. (Both cited in Chung and Wooley, 2003.)

[14] J. Crabtree, S. Harker, and V. Tannen, "The Information Integration System K2," available at http://db.cis.upenn.edu/K2/K2.doc; S.B. Davidson, J. Crabtree, B.P. Brunk, J. Schug, V. Tannen, G.C. Overton, and C.J. Stoeckert, Jr., "K2/Kleisli and GUS: Experiments in Integrated Access to Genomic Data Sources," *IBM Systems Journal* 40(2):489-511, 2001. (Both cited in Chung and Wooley, 2003.)

[15] I-M.A. Chen and V.M. Markowitz, "An Overview of the Object-Protocol Model (OPM) and OPM Data Management Tools," *Information Systems* 20(5):393-418, 1995; I-M.A. Chen, A.S. Kosky, V.M. Markowitz, and E. Szeto, "Constructing and Maintaining Scientific Database Views in the Framework of the Object-Protocol Model," *Proceedings of the Ninth International Conference on Scientific and Statistical Database Management*, Institute of Electrical and Electronic Engineers, Inc., New York, 1997, pp. 237–248. (Cited in Chung and Wooley, 2003.)

[16] N.W. Paton, R. Stevens, P. Baker, C.A. Goble, S. Bechhofer, and A. Brass, "Query Processing in the TAMBIS Bioinformatics Source Integration System," *Proceedings of the 11th International Conference on Scientific and Statistical Database Management*, IEEE, New York 1999, pp. 138-147; R. Stevens, P. Baker, S. Bechhofer, G. Ng, A. Jacoby, N.W. Paton, C.A. Goble, and A. Brass, "TAMBIS: Transparent Access to Multiple Bioinformatics Information Sources," *Bioinformatics* 16(2):184-186, 2000. (Both cited in Chung and Wooley, 2003.)

[17] Y. Papakonstantinou, H. Garcia-Molina, and J. Widom, "Object Exchange Across Heterogeneous Information Sources," *Proceedings of the IEEE Conference on Data Engineering*, IEEE, New York, 1995, pp. 251-260. (Cited in Chung and Wooley, 2003.)

databases into models of biological activity. As databases become increasingly annotated with functional and other information, they lay the groundwork for model formation.

In the future, such "database models" are envisioned as the basis of informed predictions and decision making in biomedicine. For example, physicians of the future may use biological information systems (BISs) that apply known interactions and causal relationships among proteins that regulate cell division to changes in an individual's DNA sequence, gene expression, and proteins in an individual tumor.[18] The physician might use this information together with the BIS to support a decision on whether the inhibition of a particular protein kinase is likely to be useful for treating that particular tumor.

Indeed, a major goal in the for-profit sector is to create richly annotated databases that can serve as testbeds for modeling pharmaceutical applications. For example, Entelos has developed PhysioLab, a computer model system consisting of a large set (more than 1,000) of ordinary nonlinear differential equations.[19] The model is a functional representation of human pathophysiology based on current genomic, proteomic, in vitro, in vivo, and ex vivo data, built using a top-down, disease-specific systems approach that relates clinical outcomes to human biology and physiology. Starting with major organ systems, virtual patients are explicit mathematical representations of a particular phenotype, based on known or hypothesized factors (genetic, life-style, environmental). Each model simulates up to 60 separate responses previously demonstrated in human clinical studies.

In the neuroscience field, Bower and colleagues have developed the Modeler's Workspace,[20] which is based on a notion that electronic databases must provide enhanced functionality over traditional means of distributing information if they are to be fully successful. In particular, Bower et al. believe that computational models are an inherently more powerful medium for the electronic storage and retrieval of information than are traditional online databases.

The Modeler's Workspace is thus designed to enable researchers to search multiple remote databases for model components based on various criteria; visualize the characteristics of the components retrieved; create new components, either from scratch or derived from existing models; combine components into new models; link models to experimental data as well as online publications; and interact with simulation packages such as GENESIS to simulate the new constructs.

The tools contained in the Workspace enable researchers to work with structurally realistic biological models, that is, models that seek to capture what is known about the anatomical structure and physiological characteristics of a neural system of interest. Because they are faithful to biological anatomy and physiology, structurally realistic models are a means of storing anatomical and physiological experimental information.

For example, to model a part of the brain, this modeling approach starts with a detailed description of the relevant neuroanatomy, such as a description of the three-dimensional structure of the neuron and its dendritic tree. At the single-cell level, the model represents information about neuronal morphology, including such parameters as soma size, length of interbranch segments, diameter of branches, bifurcation probabilities, and density and size of dendritic spines. At the neuronal network level, the model represents the cell types found in the network and the connectivity among them. The model must also incorporate information regarding the basic physiological behavior of the modeled structure—for example, by tuning the model to replicate neuronal responses to experimentally derived data.

With such a framework in place, a structural model organizes data in ways that make manifestly obvious how those data are related to neural function. By contrast, for many other kinds of databases it is not at all obvious how the data contained therein contribute to an understanding of function. Bower

[18] R. Brent and D. Endy, "Modelling Cellular Behaviour," *Nature* 409:391-395, 2001.

[19] See, for example, http://www.entelos.com/science/physiolabtech.html.

[20] M. Hucka, K. Shankar, D. Beeman, and J.M. Bower, "The Modeler's Workspace: Making Model-Based Studies of the Nervous System More Accessible," *Computational Neuroanatomy: Principles and Methods*, G.A. Ascoli, ed., Humana Press, Totowa, NJ, 2002, pp. 83-103.

and colleagues argue that "as models become more sophisticated, so does the representation of the data. As models become more capable, they extend our ability to explore the functional significance of the structure and organization of biological systems."[21]

4.2.8 Ontologies

Variations in language and terminology have always posed a great challenge to large-scale, comprehensive integration of biological findings. In part, this is due to the fact that scientists operate, with a data- and experience-driven intuition that outstrips the ability of language to describe. As early as 1952, this problem was recognized:

> Geneticists, like all good scientists, proceed in the first instance intuitively and . . . their intuition has vastly outstripped the possibilities of expression in the ordinary usages of natural languages. They know what they mean, but the current linguistic apparatus makes it very difficult for them to say what they mean. This apparatus conceals the complexity of the intuitions. It is part of the business of genetical methodology first to discover what geneticists mean and then to devise the simplest method of saying what they mean. If the result proves to be more complex than one would expect from the current expositions, that is because these devices are succeeding in making apparent a real complexity in the subject matter which the natural language conceals.[22]

In addition, different biologists use language with different levels of precision for different purposes. For instance, the notion of "identity" is different depending on context.[23] Two geneticists may look at a map of human chromosome 21. A year later, they both want to look at the same map again. But to one of them, "same" means exactly the same map (same data, bit for bit); to the other, it means the current map of the same biological object, even if all of the data in that map have changed. To a protein chemist, two molecules of beta-hemoglobin are the same because they are composed of exactly the same sequence of amino acids. To a biologist, the same two molecules might be considered different because one was isolated from a chimpanzee and the other from a human.

To deal with such context-sensitive problems, bioinformaticians have turned to ontologies. An ontology is a description of concepts and relationships that exist among the concepts for a particular domain of knowledge.[24] Ontologies in the life sciences serve two equally important functions. First, they provide controlled, hierarchically structured vocabularies for terminology that can be used to describe biological objects. Second, they specify object classes, relations, and functions in ways that capture the main concepts of and relationships in a research area.

4.2.8.1 Ontologies for Common Terminology and Descriptions

To associate concepts with the individual names of objects in databases, an ontology tool might incorporate a terminology database that interprets queries and translates them into search terms consistent with each of the underlying sources. More recently, ontology-based designs have evolved from static dictionaries into dynamic systems that can be extended with new terms and concepts without modification to the underlying database.

[21] M. Hucka, K. Shankar, D. Beeman, and J.M. Bower, "The Modeler's Workspace," 2002.

[22] J.H. Woodger, *Biology and Language*, Cambridge University Press, Cambridge, UK, 1952.

[23] R.J. Robbins, "Object Identity and Life Science Research," position paper submitted for the Semantic Web for Life Sciences Workshop, October 27-28 2004, Cambridge, MA, available at http://lists.w3.org/Archives/Public/public-swls-ws/2004Sep/att-0050/position-01.pdf.

[24] The term "ontology" is a philosophical term referring to the subject of existence. The computer science community borrowed the term to refer to "specification of a conceptualization" for knowledge sharing in artificial intelligence. See, for example, T.R. Gruber, "A Translation Approach to Portable Ontology Specification," *Knowledge Acquisition* 5(2):199-220, 1993. (Cited in Chung and Wooley, 2003.)

A feature of ontologies that facilitates the integration of databases is the use of a hierarchical structure that is progressively specialized; that is, specific terms are defined as specialized forms of general terms. Two different databases might not extend their annotation of a biological object to the same level of specificity, but the databases can be integrated by finding the levels within the hierarchy that share a common term.

The naming dimension of ontologies has been common to research in the life sciences for much of its history, although the term itself has not been widely used. Chung and Wooley note the following, for example:

- The Linnaean system for naming of species and organisms in taxonomy is one of the oldest ontologies.
- The nomenclature committee for the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Biochemistry and Molecular Biology (IUBMB) make recommendations on organic, biochemical, and molecular biology nomenclature, symbols, and terminology.
- The National Library of Medicine Medical Subject Headings (MeSH) provides the most comprehensive controlled vocabularies for biomedical literature and clinical records.
- A division of the College of American Pathologists oversees the development and maintenance of a comprehensive and controlled terminology for medicine and clinical information known as SNOMED (Systematized Nomenclature of Medicine).
- The Gene Ontology Consortium[25] seeks to create an ontology to unify work across many genomic projects—to develop controlled vocabulary and relationships for gene sequences, anatomy, physical characteristics, and pathology across the mouse, yeast, and fly genomes.[26] The consortium's initial efforts focus on ontologies for molecular function, biological process, and cellular components of gene products across organisms and are intended to overcome the problems associated with inconsistent terminology and descriptions for the same biological phenomena and relationships.

Perhaps the most negative aspect of ontologies is that they are in essence standards, and hence take a long time to develop—and as the size of the relevant community affected by the ontology increases, so does development time. For example, the ecological and biodiversity communities have made substantial progress in metadata standards, common taxonomy, and structural vocabulary with the help of National Science Foundation and other government agencies.[27] By contrast, the molecular biology community is much more diverse, and reaching a community-wide consensus has been much harder.

An alternative to seeking community-wide consensus is to seek consensus in smaller subcommunities associated with specific areas of research such as sequence analysis, gene expression, protein pathways, and so on.[28] These efforts usually adopt a use-case and open-source approach for community input. The ontologies are not meant to be mandatory, but instead to serve as a reference framework from which further development can proceed.

[25] See www.geneontology.org.

[26] M. Ashburner, C.A. Ball, J.A. Blacke, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, et al., "Gene Ontology: Tool for the Unification of Biology," *Nature Genetics* 25(1):25–29, 2000. (Cited in Chung and Wooley, 2003.)

[27] J.L. Edwards, M.A. Lane, and E.S. Nielsen, "Interoperability of Biodiversity Databases: Biodiversity Information on Every Desk," Science 289(5488):2312-2314, 2000; National Biological Information Infrastructure (NBII), available at http://www.nbii.gov/disciplines/systematics.html; Federal Geographic Data Committee (FGDC), available at http://www.fgdc.gov/. (All cited in Chung and Wooley, 2003.)

[28] Gene Expression Ontology Working Group, see http://www.mged.org/; P.D. Karp, M. Riley, S.M. Paley, and A. Pellegrini-Toole, "The MetaCyc Database," *Nucleic Acids Research* 30(1):59-61, 2002; P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, J. Collado-Vides, S.M. Paley, A. Pellegrini-Toole, et al., "The EcoCyc Database," *Nucleic Acids Research* 30(1):56-58, 2002; D.E. Oliver, D.L. Rubin, J.M. Stuart, M. Hewett, T.E. Klein, and R.B. Altman, "Ontology Development for a Pharmacogenetics Knowledge Base," *Pacific Symposium on Biocomputing* 65-76, 2002. (All cited in Chung and Wooley, 2003.)

An ontology developed by one subcommunity inevitably leads to interactions with related ontologies and the need to integrate. For example, consider the concept of homology. In traditional evolutionary biology, "analogy" is used to describe things that are identical by function and "homology" is used to identify things that are identical by descent. However, in considering DNA, function and descent are both captured in the DNA sequence, and therefore to molecular biologists, homology has come to mean simply similarity in sequence, regardless of whether this is due to convergence or ancestry. Thus, the term "homologous" means different things in molecular biology and evolutionary biology.[29] More broadly, a brain ontology will inevitably relate to ontologies of other anatomic structures or at the molecular level sharing ontologies for genes and proteins.[30]

Difficulties of integrating diverse but related databases thus are transformed into analogous difficulties in integrating diverse but related ontologies, but since each ontology represents the integration of multiple databases relevant to the field, the integration effort at the higher level is more encompassing. At the same time, it is also more difficult, because the implications of changes in fundamental concepts—which will be necessary in any integration effort—are much more far-reaching than analogous changes in a database. That is, design compromises in the development of individual ontologies might make it impossible to integrate the ontologies without changes to some of their basic components. This would require undoing the ontologies, then redoing them to support integration.

These points relate to semantic interoperability, which is an active area of research in computer science.[31] Information integration across multiple biological disciplines and subdisciplines would depend on the close collaborations of domain experts and information technology professionals to develop algorithms and flexible approaches to bridge the gaps between multiple biological ontologies. In recent years, a number of life science researchers have come to believe in the potential of the Semantic Web for integrating biological ontologies, as described in Box 4.3.

A sample collection of ontology resources for controlled vocabulary purposes in the life sciences is listed in Table 4.1.

4.2.8.2 Ontologies for Automated Reasoning

Today, it is standard practice to store biological data in databases; no one would deny that the volume of available data is far beyond the capabilities of human memory or written text. However, even as the volume of analytic and theoretical results drawn from these data (such as inferred genetic regulatory, metabolic, and signaling network relationships) grows, it will become necessary to store such information as well in a format suitable for computational access.

The essential rationale underlying automated reasoning is that reasoning one's way through all of the complexity inherent in biological organisms is very difficult, and indeed may be, for all practical purposes, impossible for the knowledge bases that are required to characterize even the simplest organisms. Consider, for example, the networks related to genetic regulation, metabolism, and signaling of an organism such as *Escherichia coli*. These networks are too large for humans to reason about in their totality, which means that it is increasingly difficult for scientists to be certain about global network properties. Is the model complete? Is it consistent? Does it explain all of the data? For example, the database of known molecular pathways in *E. coli* contains many hundreds of connections, far more than most researchers could remember, much less reason about.

[29] For more on the homology issue, see W.M. Fitch, "Homology: A Personal View on Some of the Problems," *Trends in Genetics* 16(5):227-231, 2000.

[30] A. Gupta, B. Ludäscher, and M.E. Martone, "Knowledge-Based Integration of Neuroscience Data Sources" *Conference on Scientific and Statistical Database Management*, Berlin, IEEE Computer Society, July 2000. (Cited in Chung and Wooley, 2003.)

[31] P. Mitra, G. Wiederhold, and M. Kersten, "A Graph-oriented Model for Articulation of Ontology Interdependencies," *Proceedings of Conference on Extending Database Technology Konstanz*, Germany, March 2000. (Cited in Chung and Wooley, 2003.)

Box 4.3
Biological Data and the Semantic Web

The Semantic Web seeks to create a universal medium for the exchange of machine-understandable data of all types, including biological data. Using Semantic Web technology, programs can share and process data even when they have been designed totally independently. The semantic web involves a Resource Description Framework (RDF), an RDF Schema language, and the Web Ontology language (OWL). RDF and OWL are Semantic Web standards that provide a framework for asset management, enterprise integration and the sharing and reuse of data on the Web. Furthermore, a standardized query language for RDF enables the "joining" of decentralized collections of RDF data. The underlying technology foundation of these languages is that of URLs, XML, and XML name spaces.

Within the life sciences, the notion of a life sciences identifier (LSID) is intended to provide a straightforward approach to naming and identifying data resources stored in multiple, distributed data stores in a manner that overcomes the limitations of naming schemes in use today. LSIDs are persistent, location-independent, resource identifiers for uniquely naming biologically significant resources including but not limited to individual genes or proteins, or data objects that encode information about them.

The life sciences pose a particular challenge for data integration because the semantics of biological knowledge are constantly changing. For example, it may be known that two proteins bind to each other. But this fact could be represented at the cellular level, the tissue level, and the molecular level depending on the context in which that fact was important.

The Semantic Web is intended to allow for evolutionary change in the relevant ontologies as new science emerges without the need for consensus. For example, if Researcher A states (and encodes using Semantic Web technology) a relationship between a protein and a signaling cascade with which Researcher B disagrees, Researcher B can instruct his or her computer to ignore (perhaps temporarily) the relationship encoded by Researcher A in favor (perhaps) of a relationship that is defined only locally.

An initiative coordinated by the World Wide Web Consortium seeks to explore how Semantic Web technologies can be used to reduce the barriers and costs associated with effective data integration, analysis, and collaboration in the life sciences research community, to enable disease understanding, and to accelerate the development of therapies. A meeting in October 2004 on the Semantic Web and the life sciences concluded that work was needed in two high-priority areas.

- In the area of ontology development, collaborative efforts were felt required to define core vocabularies that can bridge data and ontologies developed by individual communities of practice. These vocabularies would address provenance and context (e.g., identifying data sources, authors, publications names, and collection conditions), terms for cross-references in publication and other reporting of experimental results, navigation, versioning, and geospatial/temporal quantifiers.
- With respect to LSIDs, the problem of sparse implementation was regarded as central, and participants believed that work should focus on how to implement LSIDs in a manner that leverages existing Web resource resolution mechanisms such as http servers.

SOURCES: The Semantic Web Activity Statement, available at http://www.w3.org/2001/sw/Activity; *Life Sciences Identifiers RFP Response*, OMG Document lifesci/2003-12-02, January 12, 2004, available at http://www.omg.org/docs/lifesci/03-12-02.doc#_Toc61702471; John Wilbanks, Science Commons, Massachusetts Institute of Technology, personal communication, April 4, 2005.

TABLE 4.1 Biological Ontology Resources

Organization	Descriptions
Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC): http://www.gene.ucl.ac.uk/nomenclature/	HGNC is responsible for the approval of a unique symbol for each gene and designate description of genes. Aliases for genes are also listed in the database.
Gene Ontology Consortium (GO): http://www.geneontology.org	The purpose of GO is to develop ontologies describing the molecular function, biological process, and cellular component of genes and gene products for eukaryotes. Members include genome databases of fly, yeast, mouse, worm, and *Arabidopsis.*
Plant Ontology Consortium: http://www.plantontology.org	This consortium will produce structured, controlled vocabularies applied to plant-based database information.
Microarrey Gene Expression Data (MGED) Society Ontology Working Group: http://www.mged.org/	The MGED group facilitates the adoption of standards for DNA-microarray experiment annotation and data representation, as well as the introduction of standard expertmental controls and data normalization methods.
NIBII (National Biological Information Infrastructure): http://www.nbii.gov/disciplines/systematics.html	NBII provides links to taxonomy sites for all biological disciplines.
ITIS (Integrated Taxonomic Information System): http://www.itis.usda.gov/	ITIS provides taxonomic information on plants, animals, and microbes of North America and the world.
MeSH (Medical Subject Headings): http://www.nlm.nih.gov/mesh/meshhome.html	MeSH is a controlled vocabulary established by the National Library of Medicine (NLM) and used for indexing articles, cataloging books and other holdings, and searching MeSH-indexed databases, including MEDLINE.
SNOMED (Systematized Nomenclature of Medicine): http://www.snomed.org/	SNOMED is recognized globally as a comprehensive, multiaxial, controlled terminology created for the indexing of the entire medical record.
International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM): http://www.cdc.gov/nchs/about/otheract/lcd9/abtlcd9.htm	ICD-9-CM is the official system of assigning codes to diagnoses and procedures associated with hospital utilization in the United States. It is published by the U.S. National Center for Health Statistics.
International Union of Pure and Applied Chemistry (IUPAQ)	IUPAC and IUBMB make recommendations on organic, biochemical, and molecular biology nomenclature, symbols, and terminology.
International Union of Biochemistry and Molecular Biology (IUBMB) Nomenclature Committee: http://www.chem.q-mul.ac.uk/iubmb/	
PharmGKB (Pharmacogenetics Knowledge Base: http://pharmgkb.org/	PharmGKB, develops ontologies for pharmacogenetics and pharmacogenomics.

continued

TABLE 4.1 Continued

Organization	Descriptions
mmCEF (Macromolecular Crystallographic Information File): http://pdb.rutgers.edu/mmcif/ http://www.iucr.ac.ukliucr-top/cif/index.html	The information file mmCEF is sponsored by IUCr (International Union of Crystallography) to provide a dictionary for data items relevant to macromolecular crystallographic experiments.
LocusLink: http://www.ncbi.nlm.nih.gov/LocusLink/	LocusLink contains gene-centered resources, including nomenclature and aliases for genes.
Protégé-2000: http://protege.stanford.edu	Protégé-2000 is a tool that allows the user to construct a domain ontology that can be extended to access embedded applications in other knowledge-based systems. A number of biomedical ontologies have been constructed with this system, but it can be applied to other domains as well.
TAMBIS: http://imgproj.cs.man.ac.uk/tambis/	TAMBIS aims to aid researchers in the biological sciences by providing a single access point for biological information sources around the world. The access point will be a single Web-based interface that acts as a single information source. It will find appropriate sources of information for user queries and phrase the user questions for each source, returning the results in a consistent manner which will include details of the information source.

By representing working hypotheses, derived results, and the evidence that supports and refutes them in machine-readable representations, researchers can uncover correlations in and make inferences about independently conducted investigations of complex biological systems that would otherwise remain undiscovered by relying simply on serendipity or their own reasoning and memory capacities.[32] In principle, software can read and operate on these representations, determining properties in a way similar to human reasoning, but able to consider hundreds or thousands of elements simultaneously. Although automated reasoning can potentially predict the response of a biological system to a particular stimulus, it is particularly useful for discovering inconsistencies or missing relations in the data, establishing global properties of networks, discovering predictive relationships between elements, and inferring or calculating the consequences of given causal relationships.[33] As the number of discovered pathways and molecular networks increases and the questions of interest to researchers become more about global properties of organisms, automated reasoning will become increasingly useful.

Symbolic representations of biological knowledge—ontologies—are a foundation for such efforts. Ontologies contain names and relationships of the many objects considered by a theory, such as genes, enzymes, proteins, transcription, and so forth. By storing such an ontology in a symbolic machine-

[32] L. Hunter, "Ontologies for Programs, Not People," *Genome Biology* 3(6):1002.1-1002.2, 2002.

[33] As shown in Chapter 5, simulations are also useful for predicting the response of a biological system to various stimuli. But simulations instantiate procedural knowledge (i.e., *how to do* something), whereas the automated reasoning systems discussed here operate on declarative knowledge (i.e., knowledge *about* something). Simulations are optimized to answer a set of questions that is narrower than those that can be answered by automated reasoning systems—namely, predictions about the subsequent response of a system to a given stimulus. Automated reasoning systems can also answer such questions (though more slowly), but in addition they can answer questions such as, What part of a network is responsible for this particular response?, presuming that such (declarative) knowledge is available in the database on which the systems operate.

readable form and making use of databases of biological data and inferred networks, software based on artificial intelligence research can make complex inferences using these encoded relationships, for example, to consider statements written in that ontology for consistency or to predict new relationships between elements.[34] Such new relationships might include new metabolic pathways, regulatory relationships between genes, signaling networks, or other relationships. Other approaches rely on logical frameworks more expressive than database queries and are able to reason about explanations for a given feature or suggest plans for intervention to reach a desired state.[35]

Developing an ontology for automated reasoning can make use of many different sources. For example, inference from gene-expression data using Bayesian networks can take advantage of online sources of information about the likely probabilistic dependencies among expression levels of various genes.[36] Machine-readable knowledge bases can be built from textbooks, review articles, or even the *Oxford Dictionary of Molecular Biology*. The rapidly growing volume of publications in the biological literature is another important source, because inclusion of the knowledge in these publications helps to uncover relationships among various genes, proteins, and other biological entities referenced in the literature.

An example of ontologies for automated reasoning is the ontology underlying the EcoCyc database. The EcoCyc Pathway Database (http://ecocyc.org) describes the metabolic transport, and genetic regulatory networks of *E. coli*. EcoCyc structures a scientific theory about *E. coli* within a formal ontology so that the theory is available for computational analysis.[37] Specifically, EcoCyc describes the genes and proteins of *E. coli* as well as its metabolic pathways, transport functions, and gene regulation. The underlying ontology encodes a diverse array of biochemical processes, including enzymatic reactions involving small molecule substrates and macromolecular substrates, signal transduction processes, transport events, and mechanisms of regulation of gene expression.[38]

4.2.9 Annotations and Metadata

Annotation is auxiliary information associated with primary information contained in a database. Consider, for example, the human genome database. The primary database consists of a sequence of some 3 billion nucleotides, which contains genes, regulatory elements, and other material whose function is unknown. To make sense of this enormous sequence, the identification of significant patterns within it is necessary. Various pieces of the genome must be identified, and a given sequence might be annotated as translation (e.g., "stop"), transcription (e.g., "exon" or "intron"), variation ("insertion"), structural ("clone"), similarity, repeat, or experimental (e.g., "knockout," "transgenic"). Identifying a particular nucleotide sequence as a gene would itself be an annotation, and the protein corresponding to it, including its three-dimensional structure characterized as a set of coordinates of the protein's atoms, would also be an annotation. In short, the sequence database includes the raw sequence data, and the annotated version adds pertinent information such as gene coded for, amino acid sequence, or other commentary to the database entry of raw sequence of DNA bases.[39]

[34] P.D. Karp, "Pathway Databases: A Case Study in Computational Symbolic Theories," *Science* 293(5537):2040-2044, 2001.

[35] C. Baral, K. Chancellor, N. Tran, N.L. Tran, A. Joy, and M. Berens, "A Knowledge Based Approach for Representing and Reasoning About Signaling Networks," *Bioinformatics* 20(Suppl. 1):I15-I22, 2004.

[36] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller, "Rich Probabilistic Models for Gene Expression," *Bioinformatics* 17(Supp. 1):S243-S252, 2001. (Cited in Hunter, "Ontologies for Programs, Not People," 2002, Footnote 32.)

[37] P.D. Karp, "Pathway Databases: A Case Study in Computational Symbolic Theories," *Science* 293(5537):2040-2044, 2001; P.D. Karp, M. Riley, M. Saier, I.T. Paulsen, J. Collado-Vides, S.M. Paley, A Pellegrini-Toole, et al., "The EcoCyc Database," *Nucleic Acids Research* 30(1):56-58, 2002.

[38] P.D. Karp, "An Ontology for Biological Function Based on Molecular Interactions," *Bioinformatics* 16(3):269–285, 2000.

[39] See http://www.biochem.northwestern.edu/holmgren/Glossary/Definitions/Def-A/Annotation.html.

Although the genomic research community uses annotation to refer to auxiliary information that has biological function or significance, annotation could also be used as a way to trace the provenance of data (discussed in greater detail in Section 3.7). For example, in a protein database, the utility of an entry describing the three-dimensional structure of a protein would be greatly enhanced if entries also included annotations that described the quality of data (e.g., their precision), uncertainties in the data, the physical and chemical properties of the protein, various kinds of functional information (e.g., what molecules bind to the protein, location of the active site), contextual information such as where in a cell the protein is found and in what concentration, and appropriate references to the literature.

In principle, annotations can often be captured as unstructured natural language text. But for maximum utility, machine-readable annotations are necessary. Thus, special attention must be paid to the design and creation of languages and formats that facilitate machine processing of annotations. To facilitate such processing, a variety of metadata tools are available. Metadata—or literally "data about data"—are anything that describes data elements or data collections, such as the labels of the fields, the units used, the time the data were collected, the size of the collection, and so forth. They are invaluable not only for increasing the life span of data (by making it easier or even possible to determine the meaning of a particular measurement), but also for making datasets comprehensible to computers. The National Biological Information Infrastructure (NBII)[40] offers the following description:

> Metadata records preserve the usefulness of data over time by detailing methods for data collection and data set creation. Metadata greatly minimize duplication of effort in the collection of expensive digital data and foster sharing of digital data resources. Metadata supports local data asset management such as local inventory and data catalogs, and external user communities such as Clearinghouses and websites. It provides adequate guidance for end-use application of data such as detailed lineage and context. Metadata makes it possible for data users to search, retrieve, and evaluate data set information from the NBII's vast network of biological databases by providing standardized descriptions of geospatial and biological data.

A popular tool for the implementation of controlled metadata vocabularies is the extensible markup language (XML).[41] XML offers a way to serve and describe data in a uniform and automatically parsable format and provides an open-source solution for moving data between programs. Although XML is a language for describing data, the descriptions of data are articulated in XML-based vocabularies.

Such vocabularies are useful for describing specific biological entities along with experimental information associated with those entities. Some of the vocabularies have been developed in association with specialized databases established by the community. Because of their common basis in XML, however, one vocabulary can be translated to another using various tools, for example, the XML style sheet language transformation, or XSLT.[42]

Examples of such XML-based dialects include the BIOpolymer Markup Language (BIOML),[43] designed for annotating the sequences of biopolymers (e.g., genes, proteins), in such a way that all information about a biopolymer can be logically and meaningfully associated with it. Much like HTML, the language uses tags such as <protein>, <subunit>, and <peptide> to describe elements of a biopolymer along with a series of attributes.

The Microarray Markup Language (MAML) was created by a coalition of developers (www.beahmish.lbl.gov) to meet community needs for sharing and comparing the results of gene expression experiments. That community proposed the creation of a Microarray Gene Expression Database and defined the minimum information about a microarray experiment (MIAME) needed to enable

[40] See http://www.nbii.gov/datainfo/metadata/.
[41] H. Simon, *Modern Drug Discovery*, American Chemical Society, Washington, DC, 2001, pp. 69-71.
[42] See http://www.w3c./TR/xslt.
[43] See http://www.bioml.com/BIOML.

sharing. Consistent with the MIAME standards proposed by microarray users, MAML can be used to describe experiments and results from all types of DNA arrays.

The Systems Biology Markup Language, (SBML) is used to represent and model information in systems simulation software, so that models of biological systems can be exchanged by different software programs (e.g., E-Cell, StochSim). The SBML language, developed by the Caltech ERATO Kiranto systems biology Project,[44] is organized around five categories of information: model, compartment, geometry, specie, and reaction.

A downside of XML is that only a few of the largest and most used databases (e.g., a GenBank) support an XML interface. Other databases whose existence predates XML keep most of their data in flat files. But this reality is changing, and database researchers are working to create conversion tools and new database platforms based on XML. Additional XML-based vocabularies and translation tools are needed.

The data annotation process is complex and cumbersome when large datasets are involved, and some efforts have been made to reduce the burden of annotation. For example, the Distributed Annotation System (DAS) is a Web service for exchanging genome annotation data from a number of distributed databases. The system depends on the existence of a "reference sequence" and gathers "layers" of annotation about the sequence that reside on third-party servers and are controlled by each annotation provider. The data exchange standard (the DAS XML specification) enables layers to be provided in real time from the third-party servers and overlaid to produce a single integrated view by a DAS client. Success in the effort depends on the willingness of investigators to contribute annotation information recorded on their respective servers, and on users' learning about the existence of a DAS server (e.g., through ad hoc mechanisms such as link lists). DAS is also more or less specific to sequence annotation and is not easily extended to other biological objects.

Today, when biologists archive a newly discovered gene sequence in GenBank, for example, they have various types of annotation software at their disposal to link it with explanatory data. Next-generation annotation systems will have to do this for many other genome features, such as transcription-factor binding sites and single nucleotide polymorphisms (SNPs), that most of today's systems don't cover at all. Indeed, these systems will have to be able to create, annotate, and archive models of entire metabolic, signaling, and genetic pathways. Next-generation annotation systems will have to be built in a highly modular and open fashion, so that they can accommodate new capabilities and new data types without anyone's having to rewrite the basic code.

4.2.10 A Case Study: The Cell Centered Database[45]

To illustrate the notions described above, it is helpful to consider an example of a database effort that implements many of them. Techniques such as electron tomography are generating large amounts of exquisitely detailed data on cells and their macromolecular organization that have to be exposed to the greater scientific community. However, very few structured data repositories for community use exist for the type of cellular and subcellular information produced using light and electron microscopy. The Cell Centered Database (CCDB) addresses this need by developing a database for three-dimensional light and electron microscopic information.[46]

[44] See http://www.cds.caltech.edu/erato.

[45] Section 4.2.10 is adapted largely from M.E. Martone, S.T. Peltier, and M.H. Ellisman, "Building Grid Based Resources for Neurosciences," National Center for Microscopy and Imaging Research, Department of Neurosciences, University of California, San Diego, unpublished and undated working paper.

[46] M.E. Martone, A. Gupta, M. Wong, X. Qian, G. Sosinsky, B. Ludascher, and M.H. Ellisman, "A Cell-Centered Database for Electron Tomographic Data," *Journal of Structural Biology* 138(1-2):145-155, 2002; M.E. Martone, S. Zhang, S. Gupta, X. Qian, H. He, D.A. Price, M. Wong, et al., "The Cell Centered Database: A Database for Multiscale Structural and Protein Localization Data from Light and Electron Microscopy," *Neuroinformatics* 1(4):379-396, 2003.

The CCDB contains structural and protein distribution information derived from confocal, multiphoton, and electron microscopy, including correlated microscopy. Its main mission is to provide a means to make high-resolution data derived from electron tomography and high-resolution light microscopy available to the scientific community, situating itself between whole brain imaging databases such as the MAP project[47] and protein structures determined from electron microscopy, nuclear magnetic resonance (NMR) spectroscopy, and X-ray crystallography (e.g., the Protein Data Bank and EMBL).

The CCDB serves as a research prototype for investigating new methods of representing imaging data in a relational database system so that powerful data-mining approaches can be employed for the content of imaging data. The CCDB data model addresses the practical problem of image management for the large amounts of imaging data and associated metadata generated in a modern microscopy laboratory. In addition, the data model has to ensure that data within the CCDB can be related to data taken at different scales and modalities.

The data model of the CCDB was designed around the process of three-dimensional reconstruction from two-dimensional micrographs, capturing key steps in the process from experiment to analysis. (Figure 4.1 illustrates the schema-entity relationship for the CCDB.) The types of imaging data stored in the CCDB are quite heterogeneous, ranging from large-scale maps of protein distributions taken by confocal microscopy to three-dimensional reconstruction of individual cells, subcellular structures, and organelles. The CCDB can accommodate data from tissues and cultured cells regardless of tissue of origin, but because of the emphasis on the nervous system, the data model contains several features specialized for neural data. For each dataset, the CCDB stores not only the original images and three-dimensional reconstruction, but also any analysis products derived from these data, including segmented objects and measurements of quantities such as surface area, volume, length, and diameter. Users have access to the full resolution imaging data for any type of data, (e.g., raw data, three-dimensional reconstruction, segmented volumes), available for a particular dataset.

For example, a three-dimensional reconstruction is viewed as one interpretation of a set of raw data that is highly dependent on the specimen preparation and imaging methods used to acquire it. Thus, a single record in the CCDB consists of a set of raw microscope images and any volumes, images, or data derived from it, along with a rich set of methodological details. These derived products include reconstructions, animations, correlated volumes, and the results of any segmentation or analysis performed on the data. By presenting all of the raw data, as well as reconstructed and processed data with a thorough description of how the specimen was prepared and imaged, researchers are free to extract additional content from micrographs that may not have been analyzed by the original author or employ additional alignment, reconstruction, or segmentation algorithms to the data.

The utility of image databases depends on the ability to query them on the basis of descriptive attributes and on their contents. Of these two types of query, querying images on the basis of their contents is by far the most challenging. Although the development of computer algorithms to identify and extract image features in image data is advancing,[48] it is unlikely that any algorithm will be able to match the skill of an experienced microscopist for many years.

The CCDB project addresses this problem in two ways. One currently supported way is to store the results of segmentations and analyses performed by individual researchers on the data sets stored in the CCDB. The CCDB allows each object segmented from a reconstruction to be stored as a separate object in the database along with any quantitative information derived from it. The list of segmented objects and their morphometric quantities provides a means to query a dataset based on features contained in the data such as object name (e.g., dendritic spine) or quantities such as surface area, volume, and length.

[47] A. MacKenzie-Graham, E.S. Jones, D.W. Shattuck, I. Dinov, M. Bota, and A.W. Toga, "The Informatics of a C57BL/6 Mouse Brain Atlas," *Neuroinformatics* 1(4):397-410, 2003.

[48] U. Sinha, A. Bui, R. Taira, J. Dionisio, C. Morioka, D. Johnson, and H. Kangarloo, "A Review of Medical Imaging Informatics," *Annals of the New York Academy of Sciences* 980:168-197, 2002.

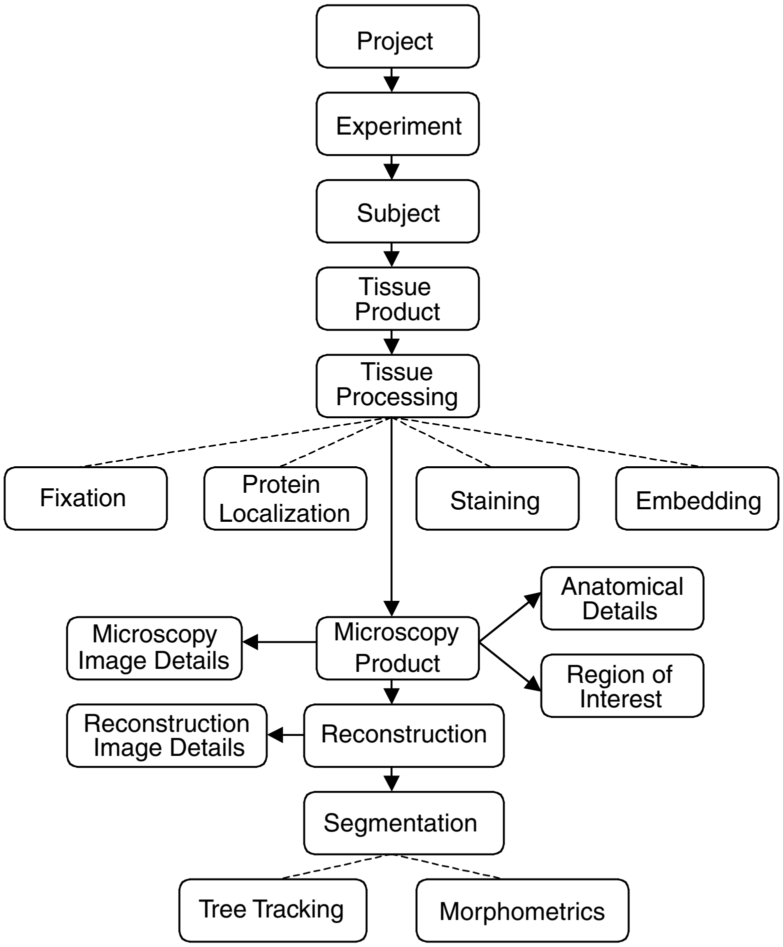


FIGURE 4.1 The schema and entity relationship in the Cell Centered Database.
SOURCE: See http://ncmir.ucsd.edu/CCDB.

It is also desirable to exploit information in the database that is not explicitly represented in the schema.[49] Thus, the CCDB project team is developing specific data types around certain classes of segmented objects contained in the CCDB. For example, the creation of a "surface data type" will enable users to query the original surface data directly. The properties of the surfaces can be determined through very general operations at query time that allow the user to query on characteristics not explicitly modeled in the schema (e.g., dendrites from striatal medium spiny cells where the diameter of the dendritic shaft shows constrictions of at least 20 percent along its length). In this example, the schema does not contain explicit indication of the shape of the dendritic shaft, but these characteristics can be computed as part of the query processing. Additional data types are being developed for volume data and protein distribution data. A data type for tree structures generated by Neurolucida has recently been implemented.

The CCDB is being designed to participate in a larger, collaborative virtual data federation. Thus, an approach to reconciling semantic differences between various databases must be found.[50] Scientific

[49] Z. Lacroix, "Issues to Address While Designing a Biological Information System," pp. 4-5 in *Bioinformatics: Managing Scientific Data,* Z.T. Lacroix , ed., Morgan Kaufmann, San Francisco, 2003.

[50] Z. Lacroix, "Issues to Address While Designing a Biological Information System," pp. 4-5 in *Bioinformatics: Managing Scientific Data*, 2003.

terminology, particularly neuroanatomical nomenclature, is vast, nonstandard, and confusing. Anatomical entities may have multiple names (e.g., caudate nucleus, *nucleus caudates)*, the same term may have multiple meanings (e.g., spine [spinal cord] versus spine [dendritic spine]), and worst of all, the same term may be defined differently by different scientists (e.g., basal ganglia). To minimize semantic confusion and to situate cellular and subcellular data from the CCDB in a larger context, the CCDB is mapped to several shared knowledge sources in the form of ontologies.

Concepts in the CCDB are being mapped to the Unified Medical Language System (UMLS), a large metathesaurus and knowledge source for the biomedical sciences.[51] The UMLS assigns each concept in the ontology a unique identifier (ID); thus, all synonymous terms can then be assigned the same ID. For example, the UMLS ID number for the synonymous terms Purkinje cell, cerebellar Purkinje cell, and Purkinje's corpuscle is C0034143. Thus, regardless of which term is preferred by a given individual, if they share the same ID, they are asserted to be the same. Conversely, even if two terms share the same name, they are distinguishable by their unique IDs. In the example given above, spine (spinal cord) = C0037949, whereas spine (dendritic spine) = C0872341.

In addition, an ontology can support the linkage of concepts by a set of relationships. These relationships may be simple "is a" and "has a" relationships (e.g., Purkinje cell is a neuron, neuron has a nucleus), or they may be more complex.[52] From the above statements, a search algorithm could infer that "Purkinje cell has a nucleus" if the ontology is encoded in a form that would allow such reasoning to be performed. Because the knowledge required to link concepts is contained outside of the source database, the CCDB is relieved of the burden of storing exhaustive taxonomies for individual datasets, which may become obsolete as new knowledge is discovered.

The UMLS has recently incorporated the NeuroNames ontology[53] as a source vocabulary. NeuroNames is a comprehensive resource for gross brain anatomy in the primate. However, for the type of cellular and subcellular data contained in the CCDB, the UMLS does not contain sufficient detail. Ontologies for areas such as neurocytology and neurological disease are being built on top of the UMLS, utilizing existing concepts wherever possible and constructing new semantic networks and concepts as needed.[54]

In addition, imaging data in the CCDB is mapped to a higher level of brain organization by registering their location in the coordinate system of a standard brain atlas. Placing data into an atlas-based coordinate systems provides one method by which data taken across scales and distributed across multiple resources can reliably be compared.[55]

Through the use of computer-based atlases and associated tools for warping and registration, it is possible to express the location of anatomical features or signals in terms of a standardized coordinate system. While there may be disagreement among neuroscientists about the identity of a brain area giving rise to a signal, its location in terms of spatial coordinates is at least quantifiable. The expression of brain data in terms of atlas coordinates also allows them to be transformed spatially to offer alternative views that may provide additional information (such as flat maps or additional parcellation

[51] B.L. Humphreys, D.A. Lindberg, H.M. Schoolman, and G.O. Barnett, "The Unified Medical Language System: An Informatics Research Collaboration," *Journal of the American Medical Informatics Association* 5(1):1-11, 1998.

[52] A. Gupta, B. Ludascher, J.S. Grethe, and M.E. Martone, "Towards a Formalization of a Disease Specific Ontology for Neuroinformatics," *Neural Networks* 16(9):1277-1292, 2003.

[53] D.M. Bowden and M.F. Dubach, "NeuroNames 2002," *Neuroinformatics* 1:43-59, 2002.

[54] A. Gupta, B. Ludascher, J.S. Grethe, and M.E. Martone, "Towards a Formalization of a Disease Specific Ontology for Neuroinformatics," *Neural Networks* 6(9):1277-1292, 2003.

[55] A. Brevik, T.B. Leergaard M. Svanevik, J.G. Bjaalie, "Three-dimensional Computerised Atlas of the Rat Brain Stem Precerebellar System: Approaches for Mapping, Visualization, and Comparison of Spatial Distribution Data," *Anatomy and Embryology* 204(4):319-332, 2001; J.G. Bjaalie, "Opinion: Localization in the Brain: New Solutions Emerging," *Nature Reviews: Neuroscience* 3(4):322-325, 2003; D.C. Van Essen, H.A. Drury, J. Dickson, J. Harwell, D. Hanlon, and C.H. Anderson, "An Integrated Software Suite for Surface-based Analyses of Cerebral Cortex," *Journal of the American Medical Informatics Association* 8(5):443-459, 2001; D.C. Van Essen, "Windows on the Brain: The Emerging Role of Atlases and Databases in Neuroscience," *Current Opinion in Neurobiology* 12(5):574-579, 2002.

schemes).[56] Finally, because individual experiments can study only a few aspects of a brain region at one time, a standard coordinate system allows the same brain region to be sampled repeatedly to allow data to be accumulated over time.

4.2.11 A Case Study: Ecological and Evolutionary Databases

Although genomic databases such as GenBank receive the majority of attention, databases and algorithms that operate on databases are key tools in research into ecology and biodiversity as well. These tools can provide researchers with access to information regarding all identified species of a given type, such as AlgaeBase[57] or FishBase;[58] they also serve as a repository for submission of new information and research. Other databases go beyond species listings to record individuals: for example, the ORNIS database of birds seeks to provide access to nearly 5 million individual specimens held in natural history collections, which includes data such as recordings of vocalizations and egg and nest holdings.[59]

The data associated with ecological research are gathered from a wide variety of sources: physical observations in the wild by both amateurs and professionals; fossils; natural history collections; zoos, botanical gardens, and other living collections; laboratories; and so forth. In addition, these data must placed into contexts of time, geographic location, environment, current and historical weather and climate, and local, regional, and global human activity. Needless to say, these data sources are scattered throughout many hundreds or thousands of different locations and formats, even when they are in digitally accessible format. However, the need for integrated ecological databases is great: only by being able to integrate the totality of observations of population and environment can certain key questions be answered. Such a facility is central to endangered species preservation, invasive species monitoring, wildlife disease monitoring and intervention, agricultural planning, and fisheries management, in addition to fundamental questions of ecological science.

The first challenge in building such a facility is to make the individual datasets accessible by networked query. Over the years, hundreds of millions of specimens have been recorded in museum records. In many cases, however, the data are not even entered into a computer; they may be stored as a set of index cards dating from the 1800s. Natural history collections, such as a museum's collection of fossils, may not even be indexed, and they are available to researchers only by physically inspecting the drawers. Very few specimens have been geocoded.

Museum records carry a wealth of image and text data, and digitizing these records in a meaningful and useful way remains a serious challenge. For this reason, funding agencies such as the National Science Foundation (NSF) are emphasizing integrating database creation, curation, and sharing into the process of ecological science: for example, the NSF Biological Databases and Informatics program[60] (which includes research into database algorithms and structures, as well as developing particular databases) and the Biological Research Collections program, which provides around $6 million per year for computerizing existing biological data. Similarly, the NSF Partnerships for Enhancing Expertise in Taxonomy (PEET) program,[61] which emphasizes training in taxonomy, requires that recipients of funding incorporate collected data into databases or other shared electronic formats.

[56] D.C. Van Essen, "Windows on the Brain: The Emerging Role of Atlases and Databases in Neuroscience," *Current Opinion in Neurobiology* 12:574-579, 2002.

[57] See http://www.algaebase.org.

[58] See http://www.fishbase.org.

[59] See http://www.ornisnet.org.

[60] NSF Program Announcement NSF 02-058; see http://www.nsf.gov/pubsys/ods/getpub.cfm?nsf02058.

[61] See http://web.nhm.ku.edu/peet/.

Ecological databases also rely on metadata to improve interoperability and compatibility among disparate data collections.[62] Ecology is a field that demands access to large numbers of independent datasets such as geographic information, weather and climate records, biological specimen collections, population studies, and genetic data. These datasets are collected over long periods of time, possibly decades or even centuries, by a diverse set of actors for different purposes. A commonly agreed-upon format and vocabulary for metadata is essential for efficient cooperative access.

Furthermore, as data increasingly are collected by automated systems such as embedded systems and distributed sensor networks, the applications that attempt to fuse the results into formats amenable to algorithmic or human analysis must deal with high (and always on) data rates, likely contained in shifting standards for representation. Again, early agreement on a basic system for sharing metadata will be necessary for the feasibility of such applications.

In attempting to integrate or cross-query these data collections, a central issue is the naming of species or higher-level taxa. The Linnean taxonomy is the oldest such effort in biology, of course, yet because there is not yet (nor likely can ever be) complete agreement on taxa identification, entries in different databases may contain different tags for members of the same species, or the same tag for members that were later determined to be of different species. Taxa are often moved into different groups, split, or merged with others; names are sometimes changed. A central effort to manage this is the Integrated Taxonomic Information System (ITIS),[63] which began life as a U.S. interagency task force, but today is a global cooperative effort between government agencies and researchers to arrive at a repository for agreed-upon species names and taxonomic categorization. ITIS data are of varying quality, and entries are tagged with three different quality indicators: credibility, which indicates whether or not data have been reviewed; latest review, giving the year of the last review; and global completeness, which records whether all species belonging to a taxon were included at the last review. These measurements allow researchers to evaluate whether the data are appropriate for their use.

In constructing such a database, many data standards questions arise. For example, ITIS uses naming standards from the International Code of Botanical Nomenclature and the International Code of Zoological Nomenclature. However, for the kingdom Protista, which at various times in biological science has been considered more like an animal and more like a plant, both standards might apply. Dates and date ranges provide another challenge: while there are many international standards for representing a calendar date, in general these did not foresee the need to represent dates occurring millions or billions of years ago. ITIS employs a representation for geologic ages, and this illustrates the type of challenge encountered when stretching a set of data standards to encompass many data types and different methods of collection.

For issues of representing observations or collections, an important element is the Darwin Core, a set of XML metadata standards for describing a biological specimen, including observations in the wild and preserved items in natural history collections. Where ITIS attempts to improve communicability by achieving agreement on precise name usage, Darwin Core[64] (and similar metadata efforts) concentrates the effort on labeling and markup of data. This allows individual databases to use their own data structures, formats, and representations, as long as the data elements are labeled by Darwin Core keywords. Since the design demands on such databases will be substantially different, this is a useful approach. Another attempt to standardize metadata for ecological data is the Access to Biological Collections Data (ABCD) Schema,[65] which is richer and contains more information. These two approaches indicate a common strategic choice: simpler standards are easier to adopt, and thus will likely be more widespread, but are limited in their expressiveness; more complex standards can successfully

[62] For a more extended discussion of the issues involved in maintaining ecological data, see W.K. Michener and J.W. Brunt, eds., *Ecological Data: Design, Management and Processing, Methods in Ecology*, Blackwell Science, Maryland, 2000. A useful online presentation can be found at http://www.soest.hawaii.edu/PFRP/dec03mtg/michener.pdf.

[63] See http://www.itis.usda.gov.

[64] See http://speciesanalyst.net/docs/dwc/.

[65] See http://www.bgbm.org/TDWG/CODATA/Schema/default.htm.

support a wider variety of queries and data types, but may be slower to gain adoption. Another effort to accomplish agreement on data and metadata standards is the National Biological Information Initiative (NBII), a program of the U.S. Geological Survey's Center for Biological Informatics.

Agreement on standard terminology and data labeling would accomplish little if the data sources were unknown. The most significant challenge in creating large-scale ecological information is the integration and federation of the potentially vast number of relevant databases. The Global Biodiversity Information Facility (GBIF)[66] is an attempt to offer a single-query interface to cooperating data providers; in December of 2004, it consisted of 95 providers totaling many tens of millions of individual records. GBIF accomplishes this query access through the use of data standards (such as the Darwin Core) and Web services, an information technology (IT) industry standard way of requesting information from servers in a platform-independent fashion. A similar international effort is found at the Clearinghouse Mechanism (CHM),[67] an instrumentality of the Convention on Biodiversity. The CHM is intended as a way for information on biodiversity to be shared among signatory states and made available as a way to monitor compliance and as a tool for policy.

Globally integrated ecological databases are still in embryonic form, but as more data become digitized and made available by the Internet in standard fashions, their value will increase. Integration with phylogenetic and molecular databases will add to their value as research tools, in both the ecological and the evolutionary fields.

4.3 DATA PRESENTATION

4.3.1 Graphical Interfaces

Biological processes can take place over a vast array of spatial scales, from the nanoscale inhabited by individual molecules, to the everyday, meter-sized human world. They can take place over an even vaster range of time scales, from the nanosecond gyrations of a folding protein molecule to the seven decade (or so) span of a human life—and far beyond, if evolutionary time is included. They also can be considered at many levels of organization, from the straightforward realm of chemical interaction to the abstract realm of, say, signal transduction and information processing.

Much of 21st century biology must deal with these processes at every level and at every scale, resulting in data of high dimensionality. Thus, the need arises for systems that can offer vivid and easily understood visual metaphors to display the information at each level, showing the appropriate amount of detail. (Such a display would be analogous to, say, a circuit diagram, with its widely recognized icons for diodes, transistors, and other such components.) A key element of such systems is easily understood metaphors that present signals containing multiple colors over time on more than one axis. As an empirical matter, these metaphors are hard to find. Indeed, the problem of finding a visually (or intellectually!) optimal display layout for high-dimensional data is arguably combinatorially hard, because in the absence of a well-developed theory of display, it requires exploring every possible combination of data in a multitude of arrangements.

The system would likewise offer easy and intuitive ways to navigate between levels, so that the user could drill down to get more detail or pop up to higher abstractions as needed. Also, it would offer good ways to visualize the dynamical behavior of the system over time—whatever the appropriate time scale might be. Current-generation visualization systems such as those associated with BioSPICE[68] and Cytoscape[69] are a good beginning—but, as their developers themselves are the first to admit, only a beginning.

[66] See http://www.gbif.org/.
[67] See http://www.biodiv.org/chm/default.aspx.
[68] See http://biospice.lbl.gov/home.html.
[69] See http://www.cytoscape.org/.

Biologists use a variety of different data representations to help describe, examine, and understand data. Biologists often use cartoons as conceptual, descriptive models of biological events or processes. A cartoon might show a time line of events: for example, the time line of the phosphorylation of a receptor that allows a protein to bind to it. As biologists take into account the simultaneous interactions of larger numbers of molecules, events over time become more difficult to represent in cartoons. New ways to "see" interactions and associations are therefore needed in life sciences research.

The most complex data visualizations are likely to be representations of networks. The complete graph in Figure 4.2 contains 4,543 nodes of approximately 6,000 proteins encoded by the yeast genome, along with 12,843 interactions. The graph was developed using the Osprey network visualization system.

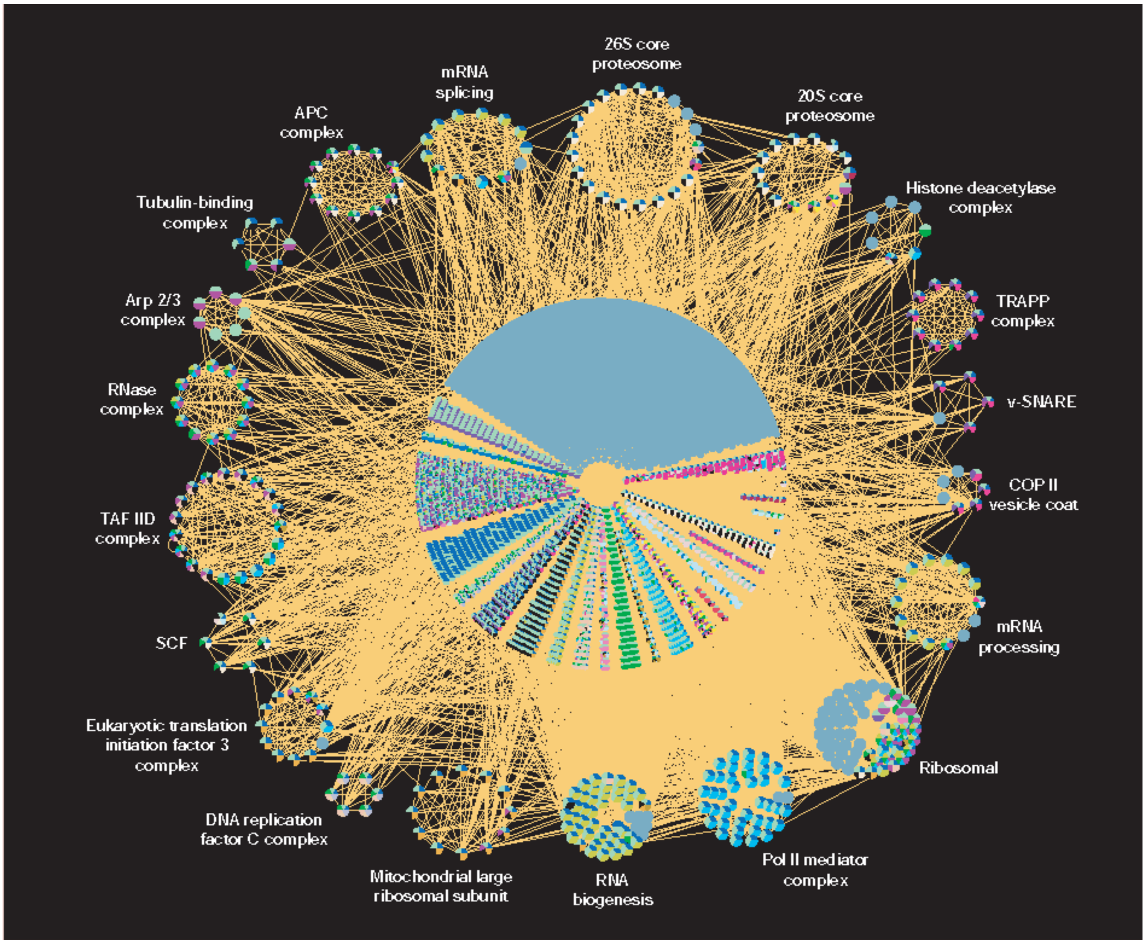


FIGURE 4.2 From genomics to proteomics. Visualization of combined, large-scale interaction data sets in yeast. A total of 14,000 physical interactions obtained from the GRID database were represented with the Osprey network visualization system (see http://biodata.mshri.on.ca/grid). Each edge in the graph represents an interaction between nodes, which are colored according to Gene Ontology (GO) functional annotation. Highly connected complexes within the dataset, shown at the perimeter of the central mass, are built from nodes that share at least three interactions within other complex members. The complete graph contains 4,543 nodes of ~6,000 proteins encoded by the yeast genome, 12,843 interactions and an average connectivity of 2.82 per node. The 20 highly connected complexes contain 340 genes, 1,835 connections, and an average connectivity of 5.39.
SOURCE: Reprinted by permission from M. Tyers and M. Mann, "From Genomics to Proteomics," *Nature* 422:193-197, 2003. Copyright 2003 Macmillan Magazines Ltd.

Other diagrammatic simulations of complex cell networks use tools such as the Diagrammatic Cell Language (DCL) and Visual Cell. These software tools are designed to read, query, and edit cell pathways, and to visualize data in a pathway context. Visual Cell creates detailed drawings by compactly formatting thousands of molecular interactions. The software uses DCL, which can visualize and simulate large-scale networks such as interconnected signal transduction pathways and the gene expression networks that control cell proliferation and apoptosis. DCL can visualize millions of chemical states and chemical reactions.

A second approach to diagrammatic simulation has been developed by Efroni et al.[70] These researchers use the visual language of Statecharts, which makes specification of the simulation precise, legible, and machine-executable. Behavior in Statecharts is described by using states and events that cause transitions between states. States may contain substates, thus enabling description at multiple levels and zooming in and zooming out between levels. States may also be divided into orthogonal states, thus modeling concurrency, allowing the system to reside simultaneously in several different states. A cell, for example, may be described orthogonally as expressing several receptors, no receptors, or any combination of receptors at different stages of the cell cycle and in different anatomical compartments. Furthermore, transitions take the system from one state to another. In cell modeling, transitions are the result of biological processes or the result of user intervention. A biological process may be the result of an interaction between two cells or between a cell and various molecules. Statecharts provide a controllable way to handle the enormous dataset of cell behavior by enabling the separation of that dataset into orthogonal states and allowing transitions.

Still another kind of graphical interface is used for molecular visualization. Interesting biomolecules usually consist of thousands of atoms. A list of atomic coordinates is useful for some purposes, but an actual image of the molecule can often provide much more insight into its properties—and an image that can be manipulated (e.g., viewed from different angles) is even more useful. Virtual reality techniques can be used to provide the viewer with a large field of view, and to enable the viewer to interact with the virtual molecule and compare it to other molecules. However, many problems in biomolecular visualization tax the capability of current systems because of the diversity of operations required and because many operations do not fit neatly into the current architectural paradigm.

4.3.2 Tangible Physical Interfaces

As useful as graphical visualizations are, even in simulated three-dimensional virtual reality they are still two-dimensional. Tangible, physical models that a human being can manipulate directly with his or her hands are an extension of the two-dimensional graphical environment. A project at the Molecular Graphics Laboratory at the Scripps Research Institute is developing tangible interfaces for molecular biology.[71] These interfaces use computer-driven autofabrication technology (i.e., three-dimensional printers) and result in physical molecular representations that one can hold in one's hand.

These efforts have required the development and testing of software for the representation of physical molecular models to be built by autofabrication technologies, linkages between molecular descriptions and computer-aided design and manufacture approaches for enhancing the models with additional physical characteristics, and integration of the physical molecular models into augmented-reality interfaces as inputs to control computer display and interaction.

[70] S. Efroni, D. Harel, and I.R. Cohen, "Toward Rigorous Comprehension of Biological Complexity: Modeling, Execution, and Visualization of Thymic T-Cell Maturation," *Genome Research* 13(11):2485-2497, 2003.

[71] A. Gillet, M. Sanner, D. Stoffler, D. Goodsell, and A. Olson, "Augmented Reality with Tangible Auto-Fabricated Models for Molecular Biology Applications," *Proceedings of the IEEE Visualization 2004 (VIS'04)*, October 10-15, 2004, Austin, pp. 235-242.

Box 4.4
Text Mining and Populating a Network Model of Intracellular Interaction

Other methods [for the construction of large-scale topological maps of cellular networks] have sought to mine MEDLINE/PubMed abstracts that are considered to contain concise records of peer-reviewed published results. The simplest methods, often called 'guilt by association,' seek to find co-occurrence of genes or protein names in abstracts or even smaller structures such as sentences or phrases. This approach assumes that co-occurrences are indicative of functional links, although an obvious limitation is that negative relations (e.g., A does not regulate B) are counted as positive associations. To overcome this problem, other natural language processing methods involve syntactic parsing of the language in the abstracts to determine the nature of the interactions. There are obvious computation costs in these approaches, and the considerable complexity in human language will probably render any machine-based method imperfect. Even with limitations, such methods will probably be required to make knowledge in the extant literature accessible to machine-based analyses. For example, PreBIND used support vector machines to help select abstracts likely to contain useful biomolecular interactions to 'backfill' the BIND database.

SOURCE: Reprinted by permission from J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2(2):70-77. Copyright 2004 Elsevier. (References omitted.)

4.3.3 Automated Literature Searching[72]

Still another form of data presentation is journal publication. It has not been lost on the scientific bioinformatics community that vast amounts of functional information that could be used to annotate gene and protein sequences are embedded in the written literature. Rice and Stolovitzky go so far as to say that mining the literature on biomolecular interactions can assist in populating a network model of intracellular interaction (Box 4.4).[73]

So far, however, the availability of full-text articles in digital formats such as PDF, HTML, or TIF files has limited the possibilities for computer searching and retrieval of full text in databases. In the future, wider use of structured documents tagged with XML will make intelligent searching of full text feasible, fast, and informative and will allow readers to locate, retrieve, and manipulate specific parts of a publication.

In the meantime, however, natural language provides a considerable, though not insurmountable, challenge for algorithms to extract meaningful information from natural text. One common application of natural language processing involves the extraction from the published literature of information about proteins, drugs, and other molecules. For example, Fukuda et al. (1998) pioneered identification of protein names using properties of the text such as the occurrence of uppercase letters, numerals, and special endings to pinpoint protein names.[74]

Other work has investigated the feasibility of recognizing interactions between proteins and other molecules. One approach is based on simultaneous occurrences of gene names and their use to predict their connections based on their occurrence statistics.[75] A second approach to pathway discovery was

[72] The discussion in Section 4.3.3 is based on excerpts from L. Hirschman, J.C. Park, J. Tsujii, L. Wong, and C.H. Wu, "Accomplishments and Challenges in Literature Data Mining for Biology," *Bioinformatics Review* 18(12):1553-1561, 2002. Available at http://pir.georgetown.edu/pirwww/aboutpir/doc/data_mining.pdf.

[73] J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2(2):70-77, 2004.

[74] K. Fukuda, et al., "Toward Information Extraction: Identifying Protein Names from Biological Papers," *Pacific Symposium on Biocomputing 1998*, 707-718. (Cited in Hirschman et al., 2002.)

[75] B. Stapley and G. Benoit, "Biobibliometrics: Information Retrieval and Visualization from Co-occurrences of Gene Names in MEDLINE Abstracts," *Pacific Symposium on Biocomputing 2000*, 529-540; J. Ding et al., "Mining MEDLINE: Abstracts, Sentences, or Phrases?" *Pacific Symposium on Biocomputing 2002*, 326-337. (Cited in Hirschman et al., 2002.)

based on templates that matched specific linguistic structures to recognize and extract of protein interaction information from MEDLINE documents.[76] More recent work goes beyond the analysis of single sentences to look at relations that span multiple sentences through the use of co-reference. For example, Putejovsky and Castano focused on relations of the word *inhibit* and showed that it was possible to extract biologically important information from free text reliably, using a corpus-based approach to develop rules specific to a class of predicates.[77] Hahn et al. described the MEDSYNDIKATE system for acquiring knowledge from medical reports, a system capable of analyzing co-referring sentences and extracting new concepts given a set of grammatical constructs.[78]

Box 4.5 describes a number of other information extraction successes in biology. In a commentary in *EMBO Reports* on publication mining, Les Grivell, manager of the European electronic publishing initiative, E-BioSci, sums up the challenges this way:[79]

> The detection of gene symbols and names, for instance, remains difficult, as researchers have seldom followed logical rules. In some organisms—the fruit fly *Drosophila* is an example—scientists have enjoyed applying gene names with primary meaning outside the biological domain. Names such as *vamp*, *eve*, *disco*, *boss*, *gypsy*, *zip* or *ogre* are therefore not easily recognized as referring to genes.[80]
>
> Also, both synonymy (many different ways to refer to the same object) and polysemy (multiple meanings for a given word) cause problems for search algorithms. Synonymy reduces the number of recalls of a given object, whereas polysemy causes reduced precision. Another problem is ambiguities of a word's sense. The word insulin, for instance, can refer to a gene, a protein, a hormone or a therapeutic agent, depending on the context. In addition, pronouns and definite articles and the use of long, complex or negative sentences or those in which information is implicit or omitted pose considerable hurdles for full-text processing algorithms.

Grivell points out that algorithms exist (e.g., the Vector Space Model) to undertake text analysis, theme generation, and summarization of computer-readable texts, but adds that "apart from the considerable computational resources required to index terms and to precompute statistical relationships for several million articles," an obstacle to full-text analysis is the fact that scientific journals are owned by a large number of different publishers, so computational analysis will have to be distributed across multiple locations.

[76] S.K. Ng and M. Wong, "Toward Routine Automatic Pathway Discovery from Online Scientific Text Abstracts," *Genome Informatics* 10:104-112, 1999. (Cited in Hirschman et al., 2002.)

[77] J. Putejovsky and J. Castano, "Robust Relational Parsing over Biomedical Literature: Extracting Inhibit Relations," *Pacific Symposium on Biocomputing 2002*, 362-373. (Cited in Hirschman et al., 2002.)

[78] U. Hahn, et al., "Rich Knowledge Capture from Medical Documents in the MEDSYNDIKATE System," *Pacific Symposium on Biocomputing 2002*, 338-349. (Cited in Hirschman et al., 2002.)

[79] L. Grivell, "Mining the Bibliome: Searching for a Needle in a Haystack? New Computing Tools Are Needed to Effectively Scan the Growing Amount of Scientific Literature for Useful Information," *EMBO Report* 3(3):200-203, 2002.

[80] D. Proux, F. Rechenmann, L. Julliard, V. Pillet. and B. Jacq, "Detecting Gene Symbols and Names in Biological Texts: A First Step Toward Pertinent Information Extraction," *Genome Informatics* 9:72-80, 1999. (Cited in Grivell, 2002.) Note also that while gene names are often italicized in print (so that they are more readily recognized as genes), neither verbal discourse nor text search recognizes italicization. In addition, because some changes of name are made for political rather than scientific reasons, and because these political revisions are done quietly, even identifying the need for synonym tracking can be problematic. An example is a gene mutation, discovered in 1963, that caused male fruit flies to court other males. Over time, the assigned gene name of "fruity" came to be regarded as offensive, and eventually the genes name was changed to "fruitless" after much public disapproval. A similar situation arose more recently, when scientists at Princeton University found mutations in flies that caused them to be learning defective or, in the vernacular of the investigators, "vegged out." They assigned names such as cabbage, rutabaga, radish, and turnip—which some other scientists found objectionable. See, for example, M. Vacek, "A Gene by Any Other Name," *American Scientist* 89(6), 2001.

Box 4.5
Selected Information Extraction Successes in Biology

Besides the recognition of protein interactions from scientific text, natural language processing has been applied to a broad range of information extraction problems in biology.

Capturing of Specific Relations in Databases.

. . . We begin with systems that capture specific relations in databases. Hahn et al. (2002) used natural language techniques and nomenclatures of the Unified Medical Language System (UMLS) to learn ontological relations for a medical domain. Baclawski et al. (2000) is a diagrammatic knowledge representation method called keynets. The UMLS ontology was used to build keynets.

Using both domain-independent and domain-specific knowledge, keynets parsed texts and resolved references to build relationships between entities. Humphreys et al. (2000) described two information extraction applications in biology based on templates: EMPathIE extracted from journal articles details of enzyme and metabolic pathways; PASTA extracted the roles of amino acids and active sites in protein molecules. This work illustrated the importance of template matching, and applied the technique to terminology recognition. Rindflesch et al. (2000) described EDGAR, a system that extracted relationships between cancer-related drugs and genes from biomedical literature. EDGAR drew on a stochastic part-of-speech tagger, a syntactic parser able to produce partial parses, a rule-based system, and semantic information from the UMLS. The metathesaurus and lexicon in the knowledge base were used to identify the structure of noun phrases in MEDLINE texts. Thomas et al. (2000) customized an information extraction system called Highlight for the task of gathering data on protein interactions from MEDLINE abstracts. They developed and applied templates to every part of the texts and calculated the confidence for each match. The resulting system could provide a cost-effective means for populating a database of protein interactions.

Information Retrieval and Clustering.

The next papers [in this volume] focus on improving retrieval and clustering in searching large collections. Chang et al. (2001) modified PSI-BLAST to use literature similarity in each iteration of its search. They showed that supplementing sequence similarity with information from biomedical literature search could increase the accuracy of homology search result. Illiopoulos et al. (2001) gave a method for clustering MEDLINE abstracts based on a statistical treatment of terms, together with stemming, a "go-list," and unsupervised machine learning. Despite the minimal semantic analysis, clusters built here gave a shallow description of the documents and supported concept discovery.

Wilbur (2002) formalized the idea of a "theme" in a set of documents as a subset of the documents and a subset of the indexing terms so that each element of the latter had a high probability of occurring in all elements of the former. An algorithm was given to produce themes and to cluster documents according to these themes.

Classification.

. . . text processing has been used for classification. Stapley et al. (2002) used a support vector machine to classify terms derived by standard term weighting techniques to predict the cellular location of proteins from description in abstracts. The accuracy of the classifier on a benchmark of proteins with known cellular locations was better than that of a support vector machine trained on amino acid composition and was comparable to a handcrafted rule-based classifier (Eisenhaber and Bork, 1999).

SOURCE: Reprinted by permission from L. Hirschman, J.C. Park, J. Tsujii, L. Wong, and C.H. Wu, "Accomplishments and Challenges in Literature Data Mining for Biology, *Bioinformatics Review* 18(12):1553-1561, 2002, available at http://pir.georgetown.edu/pirwww/aboutpir/doc/data_mining.pdf. Copyright 2002 Oxford University Press.

4.4 ALGORITHMS FOR OPERATING ON BIOLOGICAL DATA

4.4.1 Preliminaries: DNA Sequence as a Digital String

The digital nature of DNA is a central evolutionary innovation for many reasons—that is, the "values" of the molecules making up the polymer are discrete and indivisible units. Just as an electronic digital computer abstracts various continuous voltage levels as 0 and 1, DNA abstracts a three-dimensional organization of atoms as A, T, G, and C. This has important biological benefits, including very high-accuracy replication, common and simplified ways for associated molecules to bind to sites, and low ambiguity in coding for proteins.

For human purposes in bioinformatics, however, the use of the abstraction of DNA as a digital string has had other equally significant and related benefits. It is easy to imagine the opposite case, in which DNA is represented as the three-dimensional locations of each atom in the macromolecule, and comparison of DNA sequences is a painstaking process of comparing the full structures. Indeed, this is very much the state of the art in representing proteins (which, although they can be represented as a digital string of peptides, are more flexible than DNA, so the digital abstraction leaves out the critically important features of folding). The digital abstraction includes much of the essential information of the system, without including complicating higher- and lower-order biochemical properties.[81] The comparison of the state of the art in computational analysis of DNA sequences and protein sequences speaks in part to the enormous advantage that the digital string abstraction offers when appropriate.

The most basic feature of the abstraction is that it treats the arrangement of physical matter as information. An important advantage of this is that information-theoretic techniques can be applied to specific DNA strings or to the overall alphabet of codon-peptide associations. For example, computer science-developed concepts such as Hamming distance, parity, and error-correcting codes can be used to evaluate the resilience of information in the presence of noise and close alternatives.[82]

A second and very practical advantage is that as strings of letters, DNA sequences can be stored efficiently and recognizably in the same format as normal text.[83] An entire human genome, for example, can be stored in about 3 gigabytes, costing a few dollars in 2003. More broadly, this means that a vast array of tools, software, algorithms, and software packages that were designed to operate on text could be adapted with little or no effort to operate on DNA strings as well. More abstract examples include the long history of research into algorithms to efficiently search, compare, and transform strings. For example, in 1974, an algorithm for identifying the "edit distance" of two strings was discovered,[84] measuring the minimum number of changes, transpositions, and insertions necessary to transform one string into another. Although this algorithm was developed long before the genome era, it is useful to DNA analysis nonetheless.[85]

Finally, the very foundation of computational theory is the Turing machine, an abstract model of symbolic manipulation. Some very innovative research has shown that the DNA manipulations of some single-celled organisms are Turing-complete,[86] allowing the application of a large tradition of formal language analysis to problems of cellular machinery.

[81] A. Regev and E. Shapiro, "Cellular Abstractions: Cells as Computation," *Nature* 419(6905): 343, 2002.

[82] D.A. MacDonaill, "A Parity Code Interpretation of Nucleotide Alphabet Composition," *Chemical Communications* 18:2062-2063, 2002.

[83] Ideally, of course, a nucleotide could be stored using only two bits (or three to include RNA nucleotides as well). ASCII typically uses eight bits to represent characters.

[84] R.A. Wagner and M.J. Fischer, "The String-to-String Correction Problem," *Journal of the Association for Computing Machinery* 21(1):168-173, 1974.

[85] See for example, American Mathematical Society, "Mathematics and the Genome: Near and Far (Strings)," April 2002. Available at http://www.ams.org/new-in-math/cover/genome5.html; M.S. Waterman, *Introduction to Computational Biology: Maps, Sequences and Genomes*, Chapman and Hall, London, 1995; M.S. Waterman, "Sequence Alignments," *Mathematical Methods for DNA Sequences*, CRC, Boca Raton, FL, 1989, pp. 53-92.

[86] L.F. Landweber and L. Kari, "The Evolution of Cellular Computing: Nature's Solution to a Computational Problem," *Biosystems* 52(1-3):3-13, 1999.

These comments should not be taken to mean that the abstraction of DNA into a digital string is cost-free. Although digital coding of DNA is central to the mechanisms of heredity, the nucleotide sequence cannot deal with nondigital effects that also play important roles in protein synthesis and function. Proteins do not necessarily bind only to one specific sequence; the overall proportions of AT versus CG in a region affect its rate of transcription; and the state of methylation of a region of DNA is an important mechanism for the epigenetic control of gene expression (and can indeed be inherited just as the digital code can be inherited).[87] There are also numerous posttranslational modifications of proteins by processes such as acetylation, glycosylation, and phosphorylation, which by definition are not inherent in the genetic sequence.[88] The digital abstraction also cannot accommodate protein dynamics or kinetics. Because these nondigital properties can have important effects, ignoring them puts a limit on how far the digital abstraction can support research related to gene finding and transcription regulation.

Last, DNA is often compared to a computer program that drives the functional behavior of a cell. Although this analogy has some merit, it is not altogether accurate. Because DNA specifies which proteins the cell must assemble, it is at least one step removed from the actual behavior of a cell, since the proteins—not the DNA—that determine (or at least have a great influence on) cell behavior.

4.4.2 Proteins as Labeled Graphs

A significant problem in molecular biology is the challenge of identifying meaningful substructural similarities among proteins. Although proteins, like DNA, are composed of strings made from a sequence of a comparatively small selection of types of component molecules, unlike DNA, proteins can exist in a huge variety of three-dimensional shapes. Such shapes can include helixes, sheets, and other forms generally referred to as secondary or tertiary structure.

Since the structural details of a protein largely determine its functions and characteristics, determining a protein's overall shape and identifying meaningful structural details is a critical element of protein studies. Similar structure may imply similar functionality or receptivity to certain enzymes or other molecules that operate on specific molecular geometry. However, even for proteins whose three-dimensional shape has been experimentally determined through X-ray crystallography or nuclear magnetic resonance, finding similarities can be difficult due to the extremely complex geometries and large amount of data.

A rich and mature area of algorithm research involves the study of graphs, abstract representations of networks of relationships. A graph consists of a set of nodes and a set of connections between nodes called "edges." In different types of graphs, edges may be one-way (a "directed graph") or two-way ("undirected"), or edges may also have "weights" representing the distance or cost of the connection. For example, a graph might represent cities as nodes and the highways that connect them as edges weighted by the distance between the pair of cities.

Graph theory has been applied profitably to the problem of identifying structural similarities among proteins.[89] In this approach, a graph represents a protein, with each node representing a single amino acid residue and labeled with the type of residue, and edges representing either peptide bonds or close spatial proximity. Recent work in this area has combined graph theory, data mining, and information theoretic techniques to efficiently identify such similarities.[90]

[87] For more on the influence of DNA methylation on genetic regulation, see R. Jaenisch and A. Bird, "Epigenetic Regulation of Gene Expression: How the Genome Integrates Intrinsic and Environmental Signals," *Nature Genetics* 33 (Suppl):245-254, 2003.

[88] Indeed, some work even suggests that DNA methylation and histone acetylation may be connected. See J.R. Dobosy and E.U. Selker, "Emerging Connections Between DNA Methylation and Histone Acetylation," *Cellular and Molecular Life Sciences* 58(5-6):721-727, 2001.

[89] E.M. Mitchell, P.J. Artymiuk, D.W. Rice, and P. Willet, "Use of Techniques Derived from Graph Theory to Compare Secondary Structure Motifs in Proteins," *Journal of Molecular Biology* 212(1):151-166, 1989.

[90] J. Huan, W. Wang, A. Washington, J. Prins, R. Shah, and A. Tropsha, "Accurate Classification of Protein Structural Families Using Coherent Subgraph Analysis," *Pacific Symposium on Biocomputing* 2004:411-422, 2004.

A significant computational aspect of this example is that since the general problem of identifying subgraphs is NP-complete,[91] the mere inspiration of using graph theory to represent proteins is insufficient; sophisticated algorithmic research is necessary to develop appropriate techniques, data representations, and heuristics that can sift through the enormous datasets in practical times. Similarly, the problem involves subtle biological detail (e.g., what distance represents a significant spatial proximity, which amino acids can be classified together), and could not be usefully attacked by computer scientists alone.

4.4.3 Algorithms and Voluminous Datasets

Algorithms play an increasingly important role in the process of extracting information from large biological datasets produced by high-throughput studies. Algorithms are needed to search, sort, align, compare, contrast, and manipulate data related to a wide variety of biological problems and in support of models of biological processes on a variety of spatial and temporal scales. For example, in the language of automated learning and discovery, research is needed to develop algorithms for active and cumulative learning; multitask learning; learning from labeled and unlabeled data; relational learning; learning from large datasets; learning from small datasets; learning with prior knowledge; learning from mixed-media data; and learning causal relationships.[92]

The computational algorithms used for biological applications are likely to be rooted in mathematical and statistical techniques used widely for other purposes (e.g., Bayesian networks, graph theory, principal component analysis, hidden Markov models), but their adaptation to biological questions must address the constraints that define biological events. Because critical features of many biological systems are not known, algorithms must operate on the basis of working models and must frequently contend with a lack of data and incomplete information about the system under study (though sometimes simulated data suffices to test an algorithm). Thus, the results they provide must be regarded as approximate and provisional, and the performance of algorithms must be tested and validated by empirical laboratory studies. Algorithm development, therefore, requires the joint efforts of biologists and computer scientists.

Sections 4.4.4 through 4.4.9 describe certain biological problems and the algorithmic approaches to solving them. Far from giving a comprehensive description, these sections are intended to illustrate the complex substrate on which algorithms must operate and, further, to describe areas of successful and prolific collaboration between computer scientists and biologists.

Some of the applications described below are focused on identifying or measuring specific attributes, such as the identity of a gene, the three-dimensional structure of a protein, or the degree of genetic variability in a population. At the heart of these lines of investigation is the quest to understand biological function, (e.g., how genes interact, the physical actions of proteins, the physiological results of genetic differences). Further opportunities to address biological questions are likely to be as diverse as biology itself, although work on some of those questions is only nascent at this time.

4.4.4 Gene Recognition

Although the complete genomic sequences of many organisms have been determined, not all of the genes within those genomes have been identified. Difficulties in identifying genes from sequences of uncharacterized DNA stem mostly from the complexity of gene organization and architecture. Just a small fraction of the genome of a typical eukaryote consists of exons, that is, blocks of DNA that, when arranged according to their sequence in the genome, constitute a gene; in the human genome, the fraction is estimated at less than 3 percent.

[91] The notion of an NP-complete problem is rooted in the theory of computational complexity and has a precise technical definition. For purposes of this report, it suffices to understand an NP-complete problem as one that is very difficult and would take a long time to solve.

[92] S. Thurn, C. Faloutsos, T. Mitchell, and L. Wasseterman, "Automated Learning and Discovery: State-of-the-Art and Research Topics in a Rapidly Growing Field," *Summary of a Conference on Automated Learning and Discovery*, Center for Automated Learning and Discovery, Carnegie Mellon University, Pittsburgh, PA, 1998.

Regions of the genome that are not transcribed from DNA into RNA include biological signals (such as promoters) that flank the coding sequence and regulate the gene's transcription. Other untranscribed regions of unknown purpose are found between genes or interspersed within coding sequences.

Genes themselves can occasionally be found nested within one another, and overlapping genes have been shown to exist on the same or opposite DNA strands.[93] The presence of pseudogenes (nonfunctional sequences resembling real genes), which are distributed in numerous copies throughout a genome, further complicates the identification of true protein-coding genes.[94] Finally, it is known that most genes are ultimately translated into more than one protein through a process that is not completely understood. In the process of transcription, the exons of a particular gene are assembled into a single mature mRNA. However, in a process known as alternate splicing, various splicings omit certain exons, resulting in a family of variants ("splice variants") in which the exons remain in sequence, but some are missing. It is estimated that at least a third of human genes are alternatively spliced,[95] with certain splicing arrangements occurring more frequently than others. Protein splicing and RNA editing also play an important role. To understand gene structures completely, all of these sequence features have to be anticipated by gene recognition tools.

Two basic approaches have been established for gene recognition: the sequence similarity search, or lookup method, and the integrated compositional and signal search, or template method (also known as ab initio gene finding).[96] Sequence similarity search is a well-established computational method for gene recognition based on the conservation of gene sequences (called homology) in evolutionarily related organisms. A sequence similarity search program compares a query sequence (an uncharacterized sequence) of interest with already characterized sequences in a public sequence database (e.g., databases of the Institute of Genomic Research (TIGR)[97]) and then identifies regions of similarity between the sequences. A query sequence with significant similarity to the sequence of an annotated (characterized) gene in the database suggests that the two sequences are homologous and have common evolutionary origin. Information from the annotated DNA sequence or the protein coded by the sequence can potentially be used to infer gene structure or function of the query sequence, including promoter elements, potential splice sites, start and stop codons, and repeated segments. Alignment tools, such as BLAST,[98] FASTA, and Smith-Waterman, have been used to search for the homologous genes in the database.

Although sequence similarity search has been proven useful in many cases, it has fundamental limitations. Manning et al. note in their work on the protein kinase complement of the human genome

[93] I. Dunham, L.H. Matthews, J. Burton, J.L. Ashurst, K.L. Howe, K.J. Ashcroft, D.M. Beare, et al., "The DNA Sequence of Human Chromosome 22," *Nature* 402(6982):489-495, 1999.

[94] A mitigating factor is that pseudogenes are generally not conserved between species (see, for example, S. Caenepeel, G. Charydezak, S. Sudarsanam, T. Hunter, and G. Manning, "The Mouse Kinome: Discovery and Comparative Genomics of All Mouse Protein Kinases," *Proceedings of the National Academy of Sciences* 101(32):11707-11712, 2004). This fact provides another clue in deciding which sequences represent true genes and which represent pseudogenes.

[95] D. Brett, J. Hanke, G. Lehmann, S. Haase, S. Delbruck, S. Krueger, J. Reich, and P. Bork, "EST Comparison Indicates 38% of Human mRNAs Contain Possible Alternative Splice Forms," *FEBS Letters* 474(1):83-86, 2000.

[96] J.W. Fickett, "Finding Genes by Computer: The State of the Art," *Trends in Genetics* 12(8):316-320, 1996.

[97] See http://www.tigr.org/tdb/.

[98] The BLAST 2.0 algorithm, perhaps the most commonly used tool for searching large databases of gene or protein sequences, is based on the idea that sequences that are truly homologous will contain short segments that will match almost perfectly. BLAST was designed to be fast while maintaining the sensitivity needed to detect homology in distantly related sequences. Rather than aligning the full length of a query sequence against all of the sequences in the reference database, BLAST fragments the reference sequences into sub-sequences or "words" (11 nucleotides long for gene search) constituting a dictionary against which a query sequence is matched. The program creates a list of all the reference words that show up in the query sequence and then looks for pairs of those words that occur at adjacent positions on different sequences in the reference database. BLAST uses these "seed" positions to narrow candidate matches and to serve as the starting point for the local alignment of the query sequence. In local alignment, each nucleotide position in the query receives a score relative to how well the query and reference sequence match; perfect matches score highest, substitutions of different nucleotides incur different penalties. Alignment is continued outward from the seed positions until the similarity of query and reference sequences drops below a predetermined threshold. The program reports the highest scoring alignments, described by an E-value, the probability that an alignment with this score would be observed by chance. See, for example, S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman, "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs," *Nucleic Acids Research* 25(17):3389-3402, 1997.

that although "all 518 [kinase] genes are covered by some EST [Expressed Sequence Tag] sequence, and ~90% are present in gene predictions from the Celera and public genome databases, . . . those predictions are often fragmentary or inaccurate and are frequently misannotated."[99]

There are several reasons for these limitations. First, only a fraction of newly discovered sequences have identifiable homologous genes in the current databases.[100] The proportion of vertebrate genes with no detectable similarity in other phyla is estimated to be about 50 percent,[101] and this is supported by a recent analysis of human chromosome 22, where only 50 percent of the proteins are found to be similar to previously known proteins.[102] Also, the most prominent vertebrate organisms in GenBank have only a fraction of their genomes present in finished (versus draft, error-prone) sequences. Hence, it is obvious that sequence similarity search within vertebrates is currently limited. Second, sequence similarity searches are computationally expensive when query sequences have to be matched against a large number of sequences in the databases.

To resolve this problem, a dictionary-based method, such as Identifier of Coding Exons (ICE), is often employed. In this method, gene sequences in the reference database are fragmented into subsequences of length *k*, and these subsequences make up the dictionary against which a query sequence is matched. If the subsequences corresponding to a gene have at least *m* consecutive matches with a query sequence, the gene is selected for closer examination. Full-length alignment techniques are then applied to the selected gene sequences. The dictionary-based approach significantly reduces the processing time (down to seconds per gene).

In compositional and signal search, a model (typically a hidden Markov model) is constructed that integrates coding statistics (measures indicative of protein coding functions) with signal detection into one framework. An example of a simple hidden Markov model for a compositional and signal search for a gene in a sequence sampled from a bacterial genome is shown in Figure 4.3. The model is first "trained" on sequences from the reference database and generates the probable frequencies of different nucleotides at any given position on the query sequence to estimate the likelihood that a sequence is in a different "state" (such as a coding region). The query sequence is predicted to be a gene if the product of the combined probabilities across the sequence exceeds a threshold determined by probabilities generated from sequences in the reference database.

The discussion above has presumed that biological understanding does not play a role in gene recognition. This is often untrue—gene-recognition algorithms make errors of omission and commission when run against genomic sequences in the absence of experimental biological data. That is, they fail to recognize genes that are present, or misidentify starts or stops of genes, or mistakenly insert or delete segments of DNA into the putative genes. Improvements in algorithm design will help to reduce these difficulties, but all the evidence to date shows that knowledge of some of the underlying science helps even more to identify genes properly.[103]

[99] G. Manning, D.B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The Protein Kinase Complement of the Human Genome," *Science* 298(5600):1912-1934, 2002.

[100] I. Dunham, N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, R. Bruskiewich, et al. "The DNA Sequence of Human Chromosome 22," *Nature* 402(6761):489-495, 1999.

[101] J.M. Claverie, "Computational Methods for the Identification of Genes in Vertebrate Genomic Sequences," *Human Molecular Genetics* 6(10):1735-1744, 1999.

[102] I. Dunham, N. Shimizu, B.A. Roe, S. Chissoe, A.R. Hunt, J.E. Collins, R. Bruskiewich, et al., "The DNA Sequence of Human Chromosome 22," *Nature* 402(6761):489-495, 1999.

[103] This discussion is further complicated by the fact that there is no scientific consensus on the definition of a gene. Robert Robbins (Vice President for Information Technology at the Fred Hutchinson Cancer Research Center in Seattle, Washington, personal communication, December 2003) relates the following story: "Several times, I've experienced a situation where something like the following happens. First, you get biologists to agree on the definition of a gene so that a computer could analyze perfect data and tell you how many genes are present in a region. Then you apply the definition to a fairly complex region of DNA to determine the number of genes (let's say the result is 11). Then, you show the results to the biologists who provided the rules and you say, 'According to your definition of a gene there are eleven genes present in this region.' The biologists respond, 'No, there are just three. But they are related in a very complicated way.' When you then ask for a revised version of the rules that would provide a result of three in the present example, they respond, 'No, the rules I gave you are fine.'" In short, Robbins argues with considerable persuasion that if biologists armed with perfect knowledge and with their own definition of a gene cannot produce rules that will always identify how many genes are present in a region of DNA, computers have no chance of doing so.

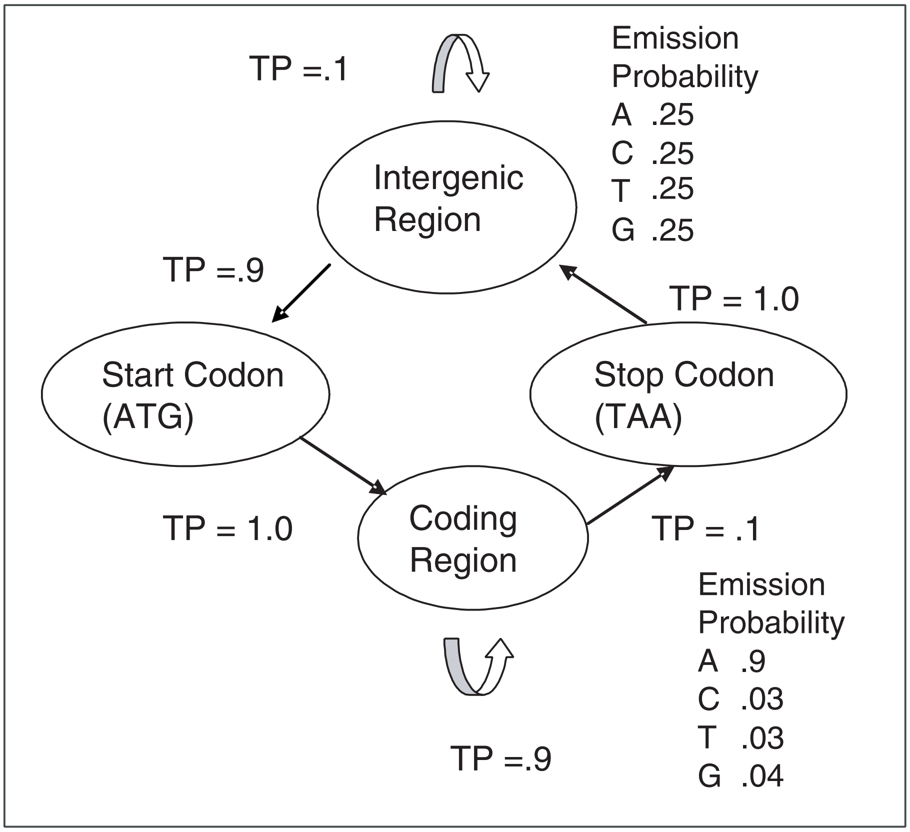


FIGURE 4.3 Hidden Markov model of a compositional signal and search approach for finding a gene in a bacterial genome.

The model has four features: (1) state of the sequence, of which four states are possible (coding, intergenic, start, and stop); (2) outputs, defined as the possible nucleotide(s) that can exist at any given state (A, C, T, G at coding and intergenic states; ATG and TAA at start and stop states, respectively); (3) emission possibilities—the probability that a given nucleotide will be generated in any particular state; and (4) transition probability (TP)—the probability that the sequence is in transition between two states.

To execute the model, emission and transition probabilities are obtained by training on the characterized genes in the reference database. The set of all possible combinations of states for the query sequence is then generated, and an overall probability for each combination of states is calculated. If the combination having the highest overall probability exceeds a threshold determined using gene sequences in the reference database, the query sequence is concluded to be a gene.

4.4.5 Sequence Alignment and Evolutionary Relationships

A remarkable degree of similarity exists among the genomes of living organisms.[104] Information about the similarities and dissimilarities of different types of organisms presents a picture of relatedness between species (i.e., between reproductive groups), but also must provide useful clues to the importance, structure, and function of genes and proteins carried or lost over time in different species. "Comparative genomics" has become a new discipline within biology to study these relationships.

[104] For example, 9 percent of *E. coli* genes, 9 percent of rice genes, 30 percent of yeast genes, 43 percent of mosquito genes, 75 percent of zebrafish genes, and 94 percent of rat genes have homologs in humans. See http://iubio.bio. Indiana.edu:8089/all/hgsummary.html (Summary Table August 2005).

Alignments of gene and protein sequences from many different organisms are used to find diagnostic patterns to characterize protein families; to detect or demonstrate homologies between new sequences and existing families of sequences; to help predict the secondary and tertiary structures of new sequences; and to serve as an essential prelude to molecular evolutionary analysis.

To visualize relationships between genomes, evolutionary biologists develop phylogenetic trees that portray groupings of organisms, characteristics, genes, or proteins based on their common ancestries and the set of common characters they have inherited. One type of molecular phylogenetic tree, for example, might represent the amino acid sequence of a protein found in several different species. The tree is created by aligning the amino acid sequences of the protein in question from different species, determining the extent of differences between them (e.g., insertions, deletions, or substitutions of amino acids), and calculating a measure of relatedness that is ultimately reflected in a drawing of a tree with nodes and branches of different lengths.

The examination of phylogenetic relationships of sequences from several different species generally uses a method known as progressive sequence alignment, in which closely related sequences are aligned first, and more distant ones are added gradually to the alignment. Attempts at tackling multiple alignments simultaneously have been limited to small numbers of short sequences because of the computational power needed to resolve them. Therefore, alignments are most often undertaken in a stepwise fashion. The algorithm of one commonly used program (ClustalW) consists of three main stages. First, all pairs of sequences are aligned separately in order to calculate a distance matrix giving the divergence of each pair of sequences; second, a guide tree is calculated from the distance matrix; and third, the sequences are progressively aligned according to the branching order in the guide tree.

Alignment algorithms that test genetic similarity face several challenges. The basic premise of a multiple sequence alignment is that, for each column in the alignment, every residue from every sequence is homologous (i.e., has evolved from the same position in a common ancestral sequence). In the process of comparing any two amino acid sequences, the algorithm must place gaps or spaces at points throughout the sequences to get the sequences to align. Because inserted gaps are carried forward into subsequent alignments with additional new sequences, the cumulative alignment of multiple sequences can become riddled with gaps that sometimes result in an overall inaccurate picture of relationships between the proteins. To address this problem, gap penalties based on a weight matrix of different factors are incorporated into the algorithm. For example, the penalty for introducing a gap in aligning two similar sequences is greater that that for aligning two dissimilar sequences. Gap penalties differ depending on the length of the sequence, the types of sequence, and different regions of sequence. Based on the weight matrix and rules for applying penalties, the algorithm compromises in the placement of gaps to obtain the lowest penalty score for each alignment.

The placement of a gap in a protein sequence may represent an evolutionary change—if a gap, reflecting the putative addition or subtraction of an amino acid to a protein's structure, is introduced, the function of the protein may change, and the change may have evolutionary benefit. However, the change may also be insignificant from a functional point of view. Today, it is known that most insertions and deletions occur in loops on the surface of the protein or between domains of multidomain proteins, which means that knowledge of the three-dimensional structure or the domain structure of the protein can be used to help identify functionally important deletions and insertions.

As the structures of different protein domains and families are increasingly determined by other means, alignment algorithms that incorporate such information should become more accurate. More recently, stochastic and iterative optimization methods are being used to refine individual alignments. Also, some algorithms (e.g., Bioedit) allow users to manually edit the alignment when other information or "eyeballing" suggests logical placement of gaps.

Exploitation of complete genomic knowledge across closely related species can play an important role in identifying the functional elements encoded in a genome. Kellis et al. undertook a comparative analysis of the yeast *Saccharomyces cerevisiae* based on high-quality draft sequences of three related species (*S. paradoxus*, *S. mikatae,* and *S. bayanus*).[105] This analysis resulted in significant revisions of the yeast gene catalogue, affecting approximately 15 percent of all genes and reducing the total count by about 500 genes. Seventy-two genome-wide elements were identified, including most known regulatory motifs and numerous new motifs, and a putative function was inferred for most of these motifs. The power of the comparative genomic approach arises from the fact that sequences that are positively selected (i.e., confer some evolutionary benefit or have some useful function) tend to be conserved as a species evolves, while other sequences are not conserved. By comparing a given genome of interest to closely related genomes, conserved sequences become much more obvious to the observer than if the functional elements had to be identified only by examination of the genome of interest. Thus, it is possible, at least in principle, that functional elements can be identified on the basis of conservation alone, without relying on previously known groups of co-regulated genes or without using data from gene expression or transcription factor binding experiments.

Molecular phylogenetic trees that graphically represent the differences between species are usually drawn with branch lengths proportional to the amount of evolutionary divergence between the two nodes they connect. The longer the distance between branches, the more relatively divergent are the sequences they represent. Methods for calculating phylogenetic trees fall into two general categories: (1) distance-matrix methods, also known as clustering or algorithmic methods, and (2) discrete data methods. In distance-matrix methods, the percentage of sequence difference (or distance) is calculated for pairwise combinations of all points of divergence; then the distances are assembled into a tree. In contrast, discrete data methods examine each column of the final alignment separately and look for the tree that best accommodates all of the information, according to optimality criteria—for example, the tree that requires the fewest character state changes (maximum parsimony), the tree that best fits an evolutionary model (maximum likelihood), or the tree that is most probable, given the data (Bayesian inference). Finally, "bootstrapping" analysis tests whether the whole dataset supports the proposed tree structure by taking random subsamples of the dataset, building trees from each of these, and calculating the frequency with which the various parts of the proposed tree are reproduced in each of the random subsamples.

Among the difficulties facing computational approaches to molecular phylogeny is the fact that some sequences (or segments of sequences) mutate more rapidly than others.[106] Multiple mutations at the same site obscure the true evolutionary difference between sequences. Another problem is the tendency of highly divergent sequences to group together when being compared regardless of their true relationships. This occurs because of a background noise problem—with only a limited number of possible sequence letters (20 in the case of amino acid sequences), even divergent sequences will not infrequently present a false phylogenetic signal due strictly to chance.

4.4.6 Mapping Genetic Variation Within a Species

The variation that occurs between different species represents the product of reproductive isolation and population fission over very long time scales during which many mutational changes in genes and proteins occur. In contrast, variation within a single species is the result of sexual reproduction, genetic

[105] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E.S. Lander, "Sequencing and Comparison of Yeast Species to Identify Genes and Regulatory Elements," *Nature* 423(6937):241-254, 2003.

[106] A number of interesting references to this problem can be found in the following: M.T. Holder and P.O. Lewis, "Phylogeny Estimation: Traditional and Bayesian Approaches," *Nature Reviews Genetics* 4:275-284, 2003; I. Holmes and W.J. Bruno, "Evolutionary HMMs: A Bayesian approach to multiple alignment," *Bioinformatics* 17(9):803-820, 2001; A. Siepel and D. Haussler, "Combining Phylogenetic and Hidden Markov Models in Biosequence Analysis," in *Proceedings of the Seventh Annual International Conference on Computational Molecular Biology*, Berlin, Germany, pp. 277-286, 2003; R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis—Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, New York, 1998.

recombination, and smaller numbers of relatively recent mutations.[107] Examining the variation of gene or protein sequences between different species helps to draw a picture of the pedigree of a particular gene or protein over evolutionary time, but scientists are also interested in understanding the practical significance of such variation within a single species.

Geneticists have been trying for decades to identify the genetic variation among individuals in the human species that result in physical differences between them. There is an increasing recognition of the importance of genetic variation for medicine and developmental biology and for understanding the early demographic history of humans.[108] In particular, variation in the human genome sequence is believed to play a powerful role in the origins of and prognoses for common medical conditions.[109]

The total number of unique mutations that might exist collectively in the entire human population is not known definitively and has been estimated at upward of 10 million,[110] which in a 3 billion base-pair genome corresponds to a variant every 300 bases or less. Included in these are single-nucleotide polymorphisms (SNPs), that is, single-nucleotide sites in the genome where two or more of the four bases (A, C, T, G) occur in at least 1 percent of the population. Many SNPs were discovered in the process of overlapping the ends of DNA sequences used to assemble the human genome, when these sequences came from different individuals or from different members of a chromosome pair from the same individual. The average number of differences observed between the DNA of any two unrelated individuals represented at 1 percent or more in the population is one difference in every 1,300 bases; this leads to the estimation that individuals differ from one another at 2.4 million places in their genomes.[111]

In rare cases, a single SNP has been directly associated with a medical condition, such as sickle cell anemia or cystic fibrosis. However, most common diseases such as diabetes, cancer, stroke, heart disease, depression, and arthritis (to name a few) appear to have complex origins and involve the participation of multiple genes along with environmental factors. For this reason there is interest in identifying those SNPs occurring across the human genome that might be correlated with common medical conditions. SNPs found within exons that contain genes are of greatest interest because they are believed to be potentially related to changes in proteins that affect a predisposition to disease, but because most of the genome does not code for proteins (and indeed a number of noncoding SNPs have been found[112]), the functional impact of many SNPs is unknown.

Armed with rapid DNA sequencing tools and the ability to detect single-base differences, an international consortium looked for SNPs in individuals over the last several years, ultimately identifying more than 3 million unique SNPs and their locations on the genome in a public database. SNP maps of the human genome with a density of about one SNP per thousand nucleotides have been developed. An effort under way in Iceland known as deCODE seeks to correlate SNPs with human diseases.[113] However, determining which combinations of the 10 million SNPs are associated with particular disease states, predisposition to disease, and genes that contribute to disease remains a formidable challenge.

Some research on this problem has recently on focused on the discovery that specific combinations of SNPs on a chromosome (called "haplotypes") occur in blocks that are inherited together; that is, they

[107] D. Posada and K.A. Crandall, "Intraspecific Gene Genealogies: Trees Grafting into Networks," *Trends in Ecology and Evolution* 16(1):37-45, 2001.

[108] L.L. Cavalli-Sforza and M.W. Feldman, "The Application of Molecular Genetic Approaches to the Study of Human Evolution," *Nature Genetics* 33 (Suppl.):266-275, 2003.

[109] S.B. Gabriel, S.F. Schaffner, H. Nguyen, J.M. Moore, J. Roy, B. Blumenstiel, J. Higins, et al., "The Structure of Haplotype Blocks in the Human Genome," *Science* 296(5576):2225-2229, 2002.

[110] L. Kruglyak and D.A. Nickerson, "Variation Is the Spice of Life," *Nature Genetics* 27(3):234-236, 2001, available at http://nucleus.cshl.edu/agsa/Papers/snp/Kruglyak_2001.pdf.

[111] The International SNP Map Working Group, "A Map of Human Genome Sequence Variation Containing 1.42 Million Single Nucleotide Polymorphisms," *Nature* 409:928-933, 2001.

[112] See, for example, D. Trikka, Z. Fang, A. Renwick, S.H. Jones, R. Chakraborty, M. Kimmel, and D.L. Nelson, "Complex SNP-based Haplotypes in Three Human Helicases: Implications for Cancer Association Studies," *Genome Research* 12(4):627-639, 2002.

[113] See www.decode.com.

are unlikely to be separated by recombination that takes place during reproduction. Further, only a relatively small number of haplotype patterns appear across portions of a chromosome in any given population.[114] This discovery potentially simplifies the problem of associating SNPs with disease because a much smaller number of "tag" SNPs (500,000 versus the estimated 10 million SNPs) might be used as representative markers for blocks of variation in initial studies to find correlations between parts of the genome and common diseases. In October 2002, the National Institutes of Health (NIH) launched the effort to map haplotype patterns (the HapMap) across the human genome.

Developing a haplotype map requires determination of all of the possible tag SNP combinations that are common in a population, and therefore relies on data from high-throughput screening of SNPs from a large number of individuals. A difficulty is that a haplotype represents a specific group of SNPs on a single chromosome. However, with the exception of gametes (sperm and egg), human cells contain two copies of each chromosome (one inherited from each parent). High-throughput studies generally do not permit the separate, parallel examination of each SNP site on both members of an individual's pair of chromosomes. SNP data obtained from individuals represent a combination of information (referred to as the genotype) from both of an individual's chromosomes. For example, genotyping an individual for the presence of a particular SNP will result in two data values (e.g., A and T). Each value represents an SNP at the same site on both chromosomes, and recently it has become possible to determine the specific chromosomes to which A and T belong.[115]

There are two problems in creating a HapMap. The first is to extract haplotype information computationally from genotype information for any individual. The second is to estimate haplotype frequencies in a population. Although good approaches to the first problem are known,[116] the second remains challenging. Algorithms such as the expectation-maximization approach, Gibbs sampling method, and partition-ligation methods have been developed to tackle this problem.

Some algorithmic programs rely on the concept of evolutionary coalescence or a perfect phylogeny—that is, a rooted tree whose branches describe the evolutionary history of a set of sequences (or haplotypes) in sample individuals. In this scenario, each sequence has a single ancestor in the previous generation, under the presumption that the haplotype blocks have not been subject to recombination, and takes as a given that only one mutation will have occurred at any one SNP site. Given a set of genotypes, the algorithm attempts to find a set of haplotypes that fit a perfect phylogeny (i.e., could have originated from a common ancestor). The performance of algorithms for haplotype prediction generally improves as the number of individuals sampled and the number of SNPs included in the analysis increases. This area of algorithm development will continue to be a robust area of research in the future as scientists and industry seek to associate genetic variation with common diseases.

Direct haplotyping is also possible, and can circumvent many of the difficulties and ambiguities encountered when a statistical approach is used.[117] For example, Ding and Cantor have developed a technique that enables direct molecular haplotyping of several polymorphic markers separated by as many as 24 kb.[118] The haplotype is directly determined by simultaneously genotyping several polymorphic markers in the same reaction with a multiplex PCR and base extension reaction. This approach does not rely on pedigree data and does not require previous amplification of the entire genomic region containing the selected markers.

[114] E.S. Lander, L.M. Linton, B. Birren, C. Nusbaum, M.C. Zody, J. Baldwin, et al., "Initial Sequencing and Analysis of the Human Genome," *Nature* 409(6822):860-921, 2001.

[115] C. Ding and C.R. Cantor, "Direct Molecular Haplotyping of Long-range Genomic DNA with M1-PCR," *Proceedings of the National Academy of Sciences* 100(13):7449-7453, 2003.

[116] See, for example, D. Gusfield, "Inference of Haplotypes from Samples of Diploid Populations: Complexity and Algorithms," *Journal of Computational Biology* 8(3):305-323, 2001.

[117] J. Tost, O. Brandt, F. Boussicault, D. Derbala, C. Caloustian, D. Lechner, and I.G. Gut, "Molecular Haplotyping at High Throughput," *Nucleic Acids Research* 30(19):e96, 2002.

[118] C. Ding and C.R. Cantor, "Direct Molecular Haplotyping of Long-range Genomic DNA with M1-PCR," *Proceedings of the National Academy of Sciences* 100(13):7449-7453, 2003.

Finally, in early 2005, National Geographic and IBM announced a collaboration known as the the Genographic Project to probe the migratory history of the human species.[119] The project seeks to collect 100,000 blood samples from indigenous populations, with the intent of analyzing DNA in these samples. Ultimately, the project will create a global database of human genetic variation and associated anthropological data (language, social customs, etc.) that provides a snapshot of human genetic variation before the cultural context of indigenous populations is lost—a context that is needed to make sense of the variations in DNA data.

4.4.7 Analysis of Gene Expression Data

Although almost all cells in an organism contain the same genetic material (the genomic blueprint for the entire organism), only about one-third of a given cell's genes are expressed or "switched on"—that is, are producing proteins—at a given time. Expressed genes account for differences in cell types; for example, DNA in skin cells produces a different set of proteins than DNA in nerve cells. Similarly, a developing embryo undergoes rapid changes in the expression of its genes as its body structure unfolds. Differential expression in the same types of cells can represent different cellular "phenotypes" (e.g., normal versus diseased), and modifying a cell's environment can result in changed levels of expression of a cell's genes. In fact, the ability to perturb a cell and observe the consequential changes in expression is a key to understanding linkages between genes and can be used to model cell signaling pathways.

A powerful technology for monitoring the activity of all the genes in a cell is the DNA microarray (described in Box 7.5 in Chapter 7). Many different biological questions can be asked with microarrays, and arrays are now constructed in many varieties. For example, instead of DNA across an entire genome, the array might be spotted with a specific set of genes from an organism or with fabricated sequences of DNA (oligonucleotides) that might represent, for example, a particular SNP or a mutated form of a gene. More recently, protein arrays have been developed as a new tool that extends the reach of gene expression analysis.

The ability to collect and analyze massive sets of data about the transcriptional states of cells is an emerging focus of molecular diagnostics as well as drug discovery. Profiling the activation or suppression of genes within cells and tissues provides telling snapshots of function. Such information is critical not only to understand disease progression, but also to determine potential routes for disease intervention. New technologies that are driving the field include the creation of "designer" transcription factors to modulate expression, use of laser microdissection methods for isolation of specific cell populations, and technologies for capturing mRNA. Among the questions asked of microarrays (and the computational algorithms to decipher the results) are the discrimination of genes with significant changes in expression relative to the presence of a disease, drug regimen, or chemical or hormonal exposure.

To illustrate the power of large-scale analysis of gene data, an article in *Science* by Gaudet and Mango is instructive.[120] A comparison of microarray data taken from *Caenorhabditis elegans* embryos lacking a pharynx with microarray data from embryos having excess pharyngeal tissue identified 240 genes that were preferentially expressed in the pharynx, and further identified a single gene as directly regulating almost all of the pharynx-specific genes that were examined in detail. These results suggest the possibility that direct transcriptional regulation of entire gene networks may be a common feature of organ-specification genes.[121]

[119] More information on the project can be found at http://www5.nationalgeographic.com/genographic/.

[120] J. Gaudet and S.E. Mango, "Regulation of Organogenesis by the *Caenorhabditis elegans* FoxA Protein PHA-4," *Science* 295(5556):821-825, 2002.

[121] For example, it is known that a specific gene activates other genes that function at two distinct steps of the regulatory hierarchy leading to wing formation in *Drosophila* (K.A. Guss, C.E. Nelson, A. Hudson, M. E. Kraus and S. B. Carroll, "Control of a Genetic Regulatory Network by a Selector Gene," *Science* 292(5519):1164-1167, 2001), and also that the presence of specific factor is both necessary and sufficient for specification of eye formation in *Drosophila* imaginal discs, where it directly activates the expression of both early- and late-acting genes (W.J. Gehring and K. Ikeo, "Pax 6: Mastering Eye Morphogenesis and Evolution," *Trends in Genetics* 15(9):371-377, 1999).

Many analytic techniques have been developed and applied to the problem of revealing biologically significant patterns in microarray data. Various statistical tests (e.g., t-test, F-test) have been developed to identify genes with significant changes in expression (out of thousands of genes); such genes have had widespread attention as potential diagnostic markers or drug targets for disease, stages of development, and other cellular phenotypes. Many classification tools (e.g., Fisher's Discriminant Analysis, Bayesian classifier, artificial neural networks, tools from signal processing) have also been developed to build a phenotype classifier with the genes differentially expressed. These classification tools are generally used to discriminate known sample groups from each other using differentially expressed genes selected by statistical testing.

Other algorithms are necessary because data acquired through microarray technology often have problems that must be managed prior to use. For example, the quality of microarray data is highly dependent on the way in which a sample is prepared. Many factors can affect the extent to which a dot fluoresces, of which the transcription level of the particular gene involved is only one. Such extraneous factors include the sample's spatial homogeneity, its cleanliness (i.e., lack of contamination), the sensitivity of optical detectors in the specific instrument, varying hybridization efficiency between clones, relative differences between dyes, and so forth. In addition, because different laboratories (and different technicians) often have different procedures for sample preparation, datasets taken from different laboratories may not be strictly comparable. Statistical methods of analysis of variance (ANOVA) have been applied to deal with these problems, using models to estimate the various contributions to relative signal from the many potential sources. Importantly, these models not only allow researchers to attach measures of statistical significance to data, but also suggest improved experimental designs.[122]

An important analytical task is to identify groups of genes with similar expression patterns. These groups of genes are more likely to be involved in the same cellular pathways, and many data-driven hypotheses about cellular regulatory mechanisms (e.g., disease mechanisms) have been drawn under this assumption. For this purpose, various clustering methods, such as hierarchical clustering methods, self-organizing maps (trained neural networks), and COSA (Clustering Objects on Subsets of Attributes), have been developed. The goal of cluster analysis is to partition a dataset of *N* objects into subgroups such that these objects are more similar to those in their subgroups than to those in other groups. Clustering tools are generally used to identify groups of genes that have similar expression pattern across samples; thus, it is reasonable to suppose that the genes in each group (or cluster) are involved in the same biological pathway. Most clustering methods are iterative and involve the calculation of a notional distance between any two data points; this distance is used as the measure of similarity. In many implementations of clustering, the distance is a function of all of the attributes of each sample.

Agglomerative hierarchical clustering begins with assigning *N* clusters for *N* samples, where all samples are defined as different individual clusters. Potential clusters are arranged in a hierarchy displayed as a binary tree or "dendrogram." Euclidian distance or Pearson correlation is used with "average linking" to develop the dendrogram. For example, two clusters that are closest to each other in terms of Euclidean distance are combined to form a new cluster, which is represented as the average of two groups combined (average linkage). This process is continued until there is one cluster to which all samples belong. In the process of forming the single cluster, the overall structure of clusters is evaluated for whether the merging of two clusters into one new cluster decreases both the sum of the similarity within all of the clusters and the sum of differences between all of the clusters. The clustering procedure stops at the level at which these are equal.

Self-organizing maps (SOMs)[123] are another form of cluster analysis. With SOMs, a number of desired clusters is decided in advance, and a geometry of nodes (such as an $N \times M$ grid) is created, where each node represents a single cluster. The nodes are randomly placed in the data space. Then, in

[122] M. Kerr, M. Martin, and G. Churchill, "Analysis of Variance for Gene Expression Microarray Data," *Journal of Computational Biology* 7(6):819-837, 2000.

[123] T. Kohonen, *Self-Organizing Maps*, Second Edition, Springer, Berlin, 1997.

a random order, each data point is selected. At each iteration, the nodes move closer to the selected data point, with the distance moved influenced by the distance from the data point to the node and the iteration number. Thus, the closest node will move the most. Over time, the initial geometry of the nodes will deform and each node will represent the center of an identified cluster. Experimentation is often necessary to arrive at a useful number of nodes and geometry, but since SOMs are computationally tractable, it is feasible to run many sessions. The properties of SOMs—partially structured, scalable to large datasets, unsupervised, easily visualizable—make them well suited for analysis of microarray data, and they have been used successfully to detect patterns of gene expression.[124]

In contrast to the above two methods, COSA is based on the assumption that better clustering can be achieved if only relevant genes are used in individual clusters. This is consistent with the idea of identifying differentially expressed genes (relevant genes) and then using only those genes to build a classifier. The search algorithm in COSA identifies an optimal set of variables that should be used to group individual clusters and which clusters should be merged when their similarity is assessed using the optimal set of variables identified. This idea was implemented by adding weights reflecting contributions of all genes to producing a particular set of sample clusters, and the search algorithm is then formulated as an optimization problem. The clustering results by COSA indicate that a subset of genes makes a greater contribution to a particular sample cluster than to other clusters.[125]

Clustering methods are being used in many types of studies. For example, they are particularly useful in modeling cell networks and in clustering disparate kinds of data (e.g., RNA data and non-RNA data; sequence data and protein data). Clustering can be applied to evaluate how feasible a given network structure is. Also, clustering is often combined with perturbation analysis to explore a set of samples or genes for a particular purpose. In general, clustering can be useful in any study in which local analyses with groups of samples or genes identified by clustering improve the understanding of the overall system.

Biclustering is an alternate approach to revealing meaningful patterns in the data.[126] It seeks to identify submatrices in which the set of values has a low mean-squared residue, meaning that the each value is reasonably coherent with other members in its row and column. (However, excluding meaningless solutions with zero area, this problem is unfortunately NP-complete.) Advantages of this approach include that it can reveal clusters based on a subset of attributes, it simultaneously clusters genes with similar expression patterns and conditions with similar expression patterns, and most importantly, clusters can overlap. Since genes are often involved in multiple biological pathways, this can be used to reveal linkages that otherwise would be obscured by traditional cluster analysis.

While many analyses of microarray data consider a single snapshot in time, of course expression levels vary over time, especially due to the cellular life cycle. A challenge in analyzing microarray time-series data is that cell cycles may be unsynchronized, making it difficult to correctly identify correlations between data samples that have similar expression behavior. Statistical techniques can identify periodicity in series and look for phase-shifted correlations between pairs of samples,[127] as well as more traditional clustering analysis.

A separate set of analytic techniques is referred to as supervised methods, in contrast to clustering and similar methods that run with no incoming assumptions. Supervised methods, in contrast, use existing knowledge of the dataset to classify data into one of a set of classes. In general, these techniques

[124] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewwan, E. Dmitrovsky, E.S. Lander, and T.R. Golub, "Interpreting Patterns of Gene Expression with Self-organizing maps: Methods and Application to Hematopoietic Differentiation," *Proceedings of the National Academy of Sciences* 96(6):2907-2912, 1999.

[125] J.H. Friedman and J.J. Meulman, "Clustering Objects on Subsets of Attributes," *Journal of the Royal Statistical Society Series B* 66(4):815-849(34), 2004.

[126] Y. Cheng and G.M. Church, "Biclustering of Expression Data," *Proceedings of the Eighth International Conference on Intelligent Systems for Molecular Biology* 8:93-103, 2000.

[127] V. Filkov, S. Skiena, and J. Zhi, "Analysis Techniques for Microarray Time-Series Data," *Journal of Computational Biology* 9(2):317-330. Available at http://www.cs.ucdavis.edu/~filkov/papers/spellmananalysis.pdf.

rely on training sets provided by the researchers, where the class membership of data is provided. Then, when presented with experimental data, supervised methods apply the learning from the training set to perform similar classifications. One such technique is support vector machines (SVMs), which are useful for highly multidimensional data. SVMs map the data into a "feature space" and then create (through one of a large number of possible algorithms) a hyperplane that separates the classes. Another common method is Artificial Neural Nets (see XREF), which train on a dataset with defined class membership; if the neural network classifies a member of the training set incorrectly, the error back-propagates through the system and updates the weightings. Unsupervised and supervised methods can be combined for "semisupervised" learning methods, in which heterogeneous training data can be both classified and unclassified.[128]

However, there is no analytic method optimal to any dataset. Thus, it would be useful to develop a scheme that can guide users to choose an appropriate method (e.g., in hierarchical clustering, an appropriate set of similarity measure, linkage method, and the measure used to determine the number of clusters) to achieve a reasonable analysis of their own datasets.

Ultimately, it is desirable to go beyond correlations and associations in the analysis of gene expression data to seek causal relationships. It is an elementary truism of statistics that indications of correlation are not by themselves indicators of causality—an experimental manipulation of one of more variables is always necessary to conclude a causal relationship. Nevertheless, analysis of microarray data can be helpful in suggesting experiments that might be particularly fruitful in uncovering causal relationships. Bayesian analysis allows one to make inferences about the possible structure of a genetic regulatory pathway on the basis of microarray data, but even advocates of such analysis recognize the need for experimental test. One work goes so far as to suggest that it is possible that automated processing of microarray data can suggest interesting experiments that will shed light on causal relationships, even if the existing data themselves don't support causal inferences.[129]

4.4.8 Data Mining and Discovery

4.4.8.1 The First Known Biological Discovery from Mining Databases[130]

By the early 1970s, the simian sarcoma virus had been determined to cause cancer in certain species of monkeys. In 1983, the responsible oncogene within the virus was sequenced. At around the same time, and entirely independently, a partial amino acid sequence of an important growth factor in humans—the platelet-derived growth factor (PDGF) was also determined. PDGF was known to cause cultured cells to proliferate in a cancer-like manner. Russell Doolittle compared the two sequences and found a high degree of similarity between them, indicating a possible connection between an oncogene and a normal human gene. In this case, the indication was that the simian sarcoma virus acted on cells in monkeys in a manner similar to the action of PDGF on human cells.

[128] T. Li, S. Zhu, Q. Li, and M. Ogihara, "Gene Functional Classification by Semisupervised Learning from Heterogeneous Data," pp. 78-82 in *Proceedings of the ACM Symposium on Applied Computing*, ACM Press, New York, 2003.

[129] C. Yoo and G. Cooper, "An Evaluation of a System That Recommends Microarray Experiments to Perform to Discover Gene-regulation Pathways," *Artificial Intelligence in Medicine* 31(2):169-182, 2004, available at http://www.phil.cmu.edu/projects/genegroup/papers/yoo2003a.pdf.

[130] Adapted from S.G.E. Andersson and L. Klasson, "Navigating Through the Databases," available at http://artedi.ebc.uu.se/course/overview/navigating_databases.html. The original Doolittle article was published as R.F. Doolittle, M.W. Hunkapiller, L.E. Hood, S.G. Davare, K.C. Robbins, S.A. Aaronson, and H.N. Antoniades, "Simian Sarcoma Virus onc Gene, v-sis, Is Derived from the Gene (or Genes) Encoding a Platelet-derived Growth Factor," *Science* 221(4607):275-277, 1983.

4.4.8.2 A Contemporary Example: Protein Family Classification and Data Integration for Functional Analysis of Proteins

New bioinformatics methods allow inference of protein function using associative analysis ("guilt by association") of functional properties to complement the traditional sequence homology-based methods.[131] Associative properties that have been used to infer function not evident from sequence homology include co-occurrence of proteins in operons or genome context; proteins sharing common domains in fusion proteins; proteins in the same pathway, subcellular network, or complex; proteins with correlated gene or protein expression patterns; and protein families with correlated taxonomic distribution (common phylogenetic or phyletic patterns).

Coupling protein classification and data integration allows associative studies of protein family, function, and structure.[132] An example is provided in Figure 4.4, which illustrates how the collective use of protein family, pathway, and genome context in bacteria helped researchers to identify a long-sought human gene associated with the methylmalonic aciduria disorder.

Domain-based or structural classification-based searches allow identification of protein families sharing domains or structural fold classes. Functional convergence (unrelated proteins with the same activity) and functional divergence are revealed by the relationships between the enzyme classification and protein family classification. With the underlying taxonomic information, protein families that occur in given lineages can be identified. Combining phylogenetic pattern and biochemical pathway information for protein families allows identification of alternative pathways to the same end product in different taxonomic groups, which may present attractive potential drug targets. The systematic approach for protein family curation using integrative data leads to novel prediction and functional inference for uncharacterized "hypothetical" proteins, and to detection and correction of genome annotation errors (a few examples are listed in Table 4.2). Such studies may serve as a basis for further analysis of protein functional evolution, and its relationship to the coevolution of metabolic pathways, cellular networks, and organisms.

Underlying this approach is the availability of resources that provide analytical tools and data. For example, the Protein Information Resource (PIR) is a public bioinformatics resource that provides an advanced framework for comparative analysis and functional annotation of proteins. PIR recently joined the European Bioinformatics Institute and Swiss Institute of Bioinformatics to establish UniProt,[133] an international resource of protein knowledge that unifies the PIR, Swiss-Prot, and TrEMBL databases. Central to the PIR-UniProt functional annotation of proteins is the PIRSF (SuperFamily) classification system[134] that provides classification of whole proteins into a network structure to reflect their evolutionary relationships. This framework is supported by the iProClass integrated database of protein family, function, and structure,[135] which provides value-added descriptions of all UniProt proteins with rich links to more than 50 other databases of protein family, function, pathway, interaction, modification, structure, genome, ontology, literature, and taxonomy. As a core resource, the PIR environment is widely used by researchers to develop other bioinformatics infrastructures and algorithms and to enable basic and applied scientific research, as shown by examples in Table 4.3.

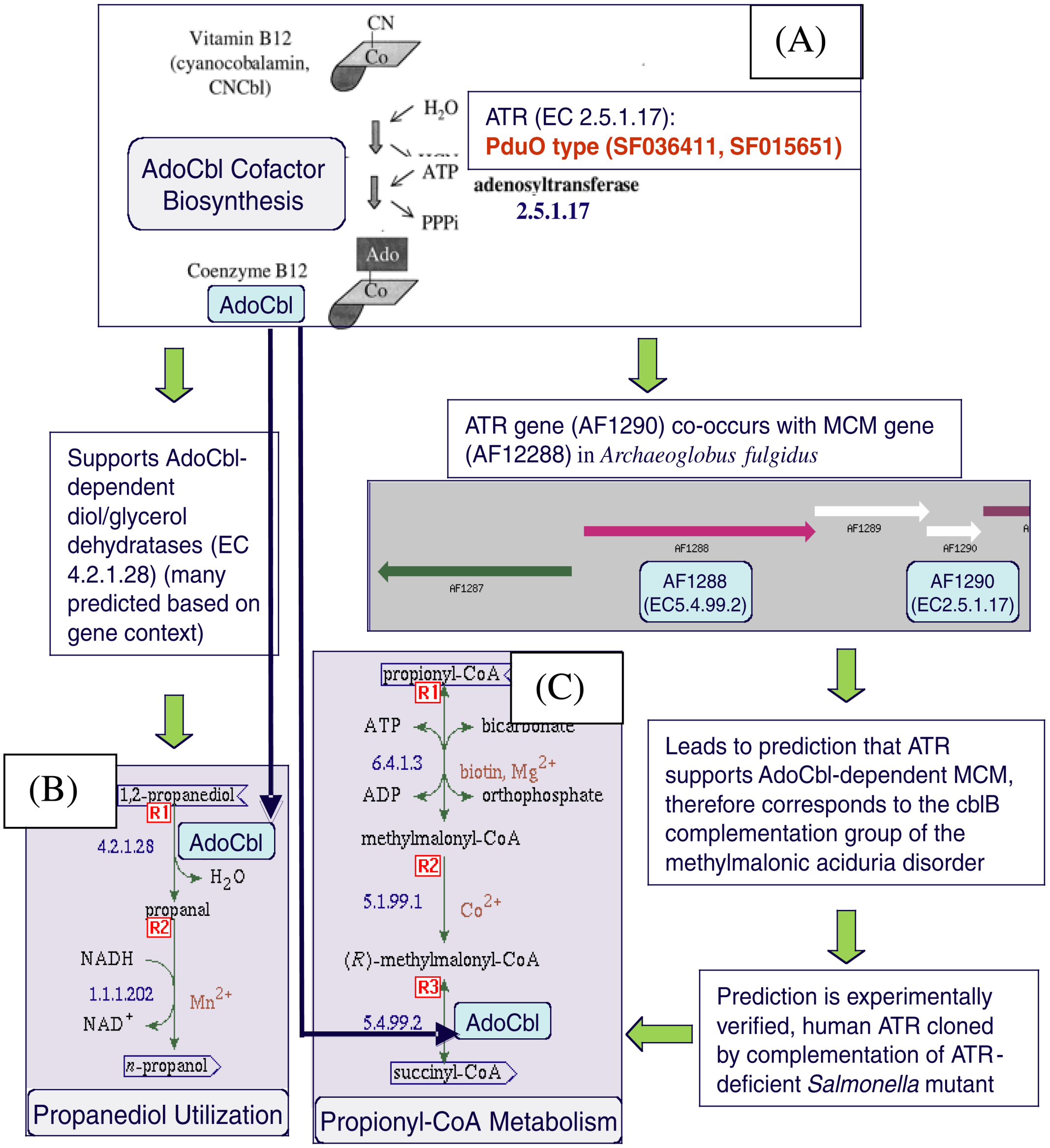
[131] E.M. Marcotte, M. Pellegrini, M.J. Thompson, T.O. Yeates, and D. Eisenberg, "Combined Algorithm for Genome-wide Prediction of Protein Function," *Nature* 402(6757):83-86, 1999.

[132] C.H. Wu, H. Huang, A. Nikolskaya, Z. Hu, and W.C. Barker, "The iProClass Integrated Database for Protein Functional Analysis," *Computational Biology and Chemistry* 28(1):87-96, 2004.

[133] R. Apweiler, A. Bairoch, C.H. Wu, W.C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, et al., "UniProt: Universal Protein Knowledgebase," *Nucleic Acids Research* 32(Database issue):D115-D119, 2004.

[134] C.H. Wu, A. Nikolskaya A, H. Huang, L.S. Yeh, D.A. Natale, C.R. Vinayaka, Z.Z. Hu, et al., "PIRSF Family Classification System at the Protein Information Resource," *Nucleic Acids Research* 32(Database issue):D112-D114, 2004.

[135] C.H. Wu, H. Huang, A. Nikolskaya, Z. Hu, and W.C. Barker, "The iProClass Integrated Database for Protein Functional Analysis," *Computational Biology and Chemistry* 28(1):87-96, 2004.



ATR gene (AF1290) co-occurs with MCM gene (AF12288) in *Archaeoglobus fulgidus*

Supports AdoCbl-dependent diol/glycerol dehydratases (EC 4.2.1.28) (many predicted based on gene context)

Leads to prediction that ATR supports AdoCbl-dependent MCM, therefore corresponds to the cblB complementation group of the methylmalonic aciduria disorder

Prediction is experimentally verified, human ATR cloned by complementation of ATR-deficient *Salmonella* mutant

FIGURE 4.4 Integration of protein family, pathway, and genome context data for disease gene identification.

The ATR enzyme (EC 2.5.1.17) converts inactive cobalamins to AdoCbl (A), a cofactor for enzymes in several pathways, including diol/glycerol dehydratase (EC 4.2.1.28) (B) and methylmalonyl-CoA mutase (MCM) (EC 5.4.99.2) (C). Many prokaryotic ATRs are predicted to be required for EC 4.2.1.28 based on the genome context of the corresponding genes. However, in at least one organism (*Archaeoglobus fulgidus*), the ATR gene is adjacent to the MCM gene, which provided a clue for cloning the human and bovine ATRs.

SOURCE: Courtesy of Cathy Wu, Georgetown University.

TABLE 4.2 Protein Family Classification and Integrative Associative Analysis for Functional Annotation

Superfamily Classification	Description
A. Functional inference of uncharacterized hypothetical proteins	
SF034452	TIM-barrel signal transduction protein
SF004961	Metal-dependent hydrolase
SF005928	Nucleotidyltransferase
SF005933	ATPase with chaperone activity and inactive LON protease domain
SF005211	alpha/beta hydrolase
SF014673	Lipid carrier protein
SF005019	[Ni,Fe]-Hydrogenase-3-type complex, membrane protein EhaA
B. Correction or improvement of genome annotations	
SF025624	Ligand-binding protein with an ACT domain
SF005003	Inactive homologue of metal-dependent protease
SF000378	Glycyl radical cofactor protein YfiD
SF000876	Chemotaxis response regulator methylesterase CheB
SF000881	Thioesterase, type II
SF002845	Bifunctional tetrapyrrole methylase and MazG NTPase
C. Enhanced understanding of structure, function, evolutionary relationships	
SF005965	Chorismate mutase, AroH class
SF001501	Chorismate mutase, AroQ class, prokaryotic type

NOTE: PIRSF protein family reports detail supporting evidence for both experimentally validated and computationally predicted annotations.

4.4.9 Determination of Three-dimensional Protein Structure

One central problem of proteomics is that of protein folding. Protein folding is one of the most important cellular processes because it produces the final conformation required for a protein to attain biological activity. Diseases such as Alzheimer's disease or bovine spongiform encephalopathy (BSE, or "Mad Cow" disease) are associated with the improper folding of proteins. For example, in BSE the protein (called the scrapie prion), which is soluble when it folds properly, becomes insoluble when one of the intermediates along its folding pathway misfolds and forms an aggregation that damages nerve cells.[136]

Due to the importance of the functional conformation of proteins, many efforts have been attempted to predict computationally a three-dimensional structure of a protein from its amino acid sequence. Although experimental determination of protein structure based on X-ray crystallography and nuclear magnetic resonance yields protein structures in high resolution, it is slow, labor-intensive, and expensive and thus not appropriate for large-scale determination. Also, it can apply only to already-synthesized or isolated proteins, while an algorithm could be used to predict the structure of a great number of potential proteins.

[136] See, for example, C.M. Dobson, "Protein Misfolding, Evolution and Disease," *Trends in Biochemical Science* 24(9):329-332, 1999; C.M. Dobson, "Protein Folding and Its Links with Human Disease." *Biochemical Society Symposia* 68:1-26, 2001; C.M. Dobson, "Protein Folding and Misfolding," *Nature* 426(6968):884-890, 2003.

TABLE 4.3 Algorithms, Databases, Analytical Systems, and Scientific Research Enabled by the PIR Resource

Resource	Topic	Reference
Algorithm	Benchmarking for sequence similarity search statistics	Pearson, *J. Mol. Biol.* 276:71-84, 1998
	PANDORA keyword-based analysis of proteins	Kaplan, *Nucleic Acids Research* 31:5617-5626, 2003
	Computing motif correlations for structure prediction	Horng et al., *J. Comp. Chem.* 24(16):2032-2043, 2003
Database	NESbase database of nuclear export signals	la Cour et al., *Nucleic Acids Research* 31(l):393-396, 2003
	TMPDB database of transmembrane topologies	Ikeda et al., *Nucleic Acids Research* 31:406-409, 2003
	SDAP database and tools for allergenic proteins	Ivanciuc et al., *Nucleic Acids Research* 31:359-362, 2003
System	SPINE 2 system for collaborative structural proteomics	Goh et al., *Nucleic Acids Research* 31:2833-2838, 2003
	ERGOTM genome analysis and discovery system	Overbeek et al., *Nucleic Acids Research* 31(l):164-171, 2003
	Automated annotation pipeline and cDNA annotation system	Kasukawa et al., *Genome Res.* 13(6B):1542-1551, 2003
	Systers, GeneNest, SpliceNest from genome to protein	Krause et al., *Nucleic Acids Research* 30(l):299-300, 2002
Research	Intermediate filament proteins during carcinogenesis or apoptosis	Prasad et al., *Int. J. Oncol.* 14(3):563-570, 1999
	Conserved pathway by global protein network alignment	Kelley et al., *PNAS* 100(20):11394-11399, 2003
	Membrane targeting of phospholipase C pleckstrin	Singh and Murray, *Protein Sci.* 12:1934-1953, 2003
	Analysis of human and mouse cDNA sequences	Strausberg et al., *PNAS* 99(26):16899-16903, 2002
	A novel *Schistosoma mansoni* G protein-coupled receptor	Hamdan et al., *Mol. Biochem. Parasitol.* 119(l):75-86, 2002
	Proteomics reveals open reading frames (ORFs) in *Mycobacterium tuberculosis*	Jungblut et al., *Infect. Immunol.* 69(9):5905-5907, 2001

Protein structures predicted in high resolution can help characterize the biological functions of proteins. Biotechnology companies are hoping to accelerate their efforts to discover new drugs that interact with proteins by using structure-based drug design technologies. By combining computational and combinatorial chemistry, researchers expect to find more viable leads. Algorithms create molecular structure built de novo to optimize interactions within the protein's active sites. The use of so-called virtual screening in combination with studies of co-crystallized drugs and proteins could be a powerful tool for drug development.

A number of tools for protein structure prediction have been developed, and progress in prediction by these methods has been evaluated by the Critical Assessment of Protein Structure Prediction (CASP) experiment held every two years since 1994.[137] In a CASP experiment, the amino acid sequences of proteins whose experimentally determined structures have not yet been released are published, and computational research groups are then invited to predict structures of these target sequences using their methods and any other publicly available information (e.g., known structures that exist in the Protein Data Bank (PDB), the data repository for protein structures). The methods used by the groups

[137] See http://predictioncenter.llnl.gov/.

can be divided into three areas depending on the similarity of the target protein to proteins of known structure: comparative (also known as homology) modeling, fold recognition (also known as threading), and de novo/new fold methods (also known as ab initio). This traditional division of prediction methods has become blurred as the methods in each category incorporate detailed information used by methods in the other categories.

In comparative (or homology) modeling, one or more template proteins of known structure with high sequence homology (greater than 25 percent sequence identity) to the target sequence are identified. The target and template sequences are aligned through multiple sequence alignment (similar to comparative genomics), and a three-dimensional structure of the target protein is generated from the coordinates of the aligned residues of the template proteins. Finally, the model is evaluated using a variety of criteria, and if necessary, the alignment and the three-dimensional model are refined until a satisfactory model is obtained.

If no reliable template protein can be identified from sequence homology alone, the prediction problem is denoted as a fold recognition (or threading) problem. The primary goal is to identify one or more folds in the template proteins that are consistent with the target sequence. In the classical threading methods, known as "rigid body assembly," a model is constructed from a library of known core regions, loops, side chains, and folds, and the target sequence is then threaded onto the known folds. After evaluating how well the model fits the known folds, the best fit is chosen. The assumption in fold recognition is that only a finite number of folds exist and most existing folds can be identified from known structures in the PDB. Indeed, as new sequences are deposited and more protein structures are solved, there appear to be fewer and fewer unique folds. When two sequences share more than 25 percent similarity (or sequence identity), their structures are expected to have similar folds. However, there are still remaining issues such as the high rate of false positives in fold recognition, and therefore, the resulting alignment with the fold structure is poor. At 30 percent sequence identity, the fraction of incorrectly aligned residues is about 20 percent, and the number rises sharply with further decreases in sequence similarity. This limits the usefulness of comparative modeling.[138]

If no template structure (or fold) can be identified with confidence by sequence homology methods, the target sequence may be modeled using new fold prediction methods. The goal in this prediction method rests on the biological assumption that proteins adopt their lowest free energy conformation as their functional state. Thus, computational methods to predict structure ab initio comprise three elements: (1) protein geometry, (2) potential energy functions, and (3) an energy space search method (energy minimization method). First, setting protein geometry involves determining the number of particles to be used to represent the protein structure (for example, all-atom, united-atom, or virtual-atom model) and the nature of the space where atoms can be allocated (e.g., continuous (off-lattice) or discrete (lattice) model). In a simple ab initio folding such as a virtual-atom lattice model, one virtual atom represents a number of atoms in a protein (i.e., the backbone is represented as a sequence of alpha carbons) and an optimization method searches only the predetermined lattice points for positions of the virtual atoms to minimize the energy functions. Second, the potential energy functions in ab initio models include covalent terms, such as bond stretching, bond angle stretching, improper dihedrals, and torsional angles, and noncovalent terms, such as electrostatic and van der Waals forces. The use of molecular mechanics for refinement in comparative modeling is equivalent to ab initio calculation using all atoms in an off-lattice model. Third, many optimizations tools, such as genetic algorithms, Monte Carlo, simulated annealing, branch and bound, and successive quadratic programming (SQP), have been used to search for the global minimum in the energy (or structure) spaces with a number of local minima. These approaches have provided encouraging results, although the performance of each method may be limited by the shape of the energy space.

[138] T. Head-Gordon and J. Wooley, "Computational Challenges in Structural and Functional Genomics," *IBM Systems Journal* 40(2):265-296, 2001.

Beyond studies of protein structure is the problem of describing a solvent environment (such as water) and its influence on a protein's conformational behavior. The importance of hydration in protein stability and folding is widely accepted. Models are needed to incorporate the effects of solvents in protein three-dimensional structure.

4.4.10 Protein Identification and Quantification from Mass Spectrometry

A second important problem in proteomics is protein identification and quantification. That is, given a particular biological sample, what specific proteins are present and in what quantities? This problem is at the heart of studying protein–protein interactions at proteomic scale, mapping various organelles, and generating quantitative protein profiles from diverse species. Making inferences about protein identification and abundance in biological samples is often challenging, because cellular proteomes are highly complex and because the proteome generally involves many proteins at relatively low abundances. Thus, highly sensitive analytical techniques are necessary.

Today, techniques based on mass spectrometry increasingly fill this need. The mass spectrometer works on a biological sample in ionized gaseous form. A mass analyzer measures the mass-to-charge ratio (m/z) of the ionized analytes, and a detector measures the number of ions at each m/z value. In the simplest case, a procedure known as peptide mass fingerprinting (PMF) is used. PMF is based on the fact that a protein is composed of multiple peptide groups, and identification of the complete set of peptides will with high probability characterize the protein in question. After enzymatically breaking up the protein into its constituent peptides, the mass spectrometer is used to identify individual peptides, each of which has a known mass. The premise of PMF is that only a very few (one in the ideal case) proteins will correspond to any particular set of peptides, and protein identification is effected by finding the best fit of the observed peptide masses to the calculated masses derived from, say, a sequence database. Of course, the "best fit" is an algorithmic issue, and a variety of approaches have been taken to determine the most appropriate algorithms.

The applicability of PMF is limited when samples are complex (that is, when they involve large numbers of proteins at low abundances). The reason is that only a small fraction of the constituent peptides are typically ionized, and those that are observed are usually from the dominant proteins in the mixture. Thus, for complex samples, multiple (tandem) stages of mass spectrometry may be necessary. In a typical procedure, peptides from a database are scored on the likelihood of their generating a tandem mass spectrum, and the top scoring peptide is chosen. This computational approach has shown great success, and contributed to the industrialization of proteomics.

However, much remains to be done. First, the generation of the spectrum is a stochastic process governed by the peptide composition, and the mass spectrometer. By mining data to understand these fragmentation propensities, scoring and identification can be further improved. Second, if the peptide is not in the database, de novo or homology-based methods must be developed for identification. Many proteins are post-translationally modified, with the modifications changing the mass composition. Enumeration and scoring of all modifications leads to a combinatorial explosion that must be addressed using novel computational techniques. It is fair to say that computation will play an important role in the success of mass spectrometry as the tool of choice for proteomics.

Mass spectrometry is also coming into its own for protein expression studies. The major problem here is that the intensity of a peak depends not only on the peptide abundance, but also on the physicochemical properties of the peptide. This makes it difficult to measure expression levels directly. However, relative abundance can be measured using the proven technique of stable-isotope dilution. This method makes use of the facts that pairs of chemically identical analytes of different stable-isotope composition can be differentiated in a mass spectrometer owing to their mass difference, and that the ratio of signal intensities for such analyte pairs accurately indicates the abundance ratio for the two analytes.

This approach shows great promise. However, computational methods are needed to correlate data across different experiments. If the data were produced using liquid chromatography coupled with mass spectrometry, a peptide pair could be approximately labeled by its retention time in the column, and its mass-to-charge ratio. Such pairs can be matched across experiments using geometric matching. Combining the relative abundance levels from different experiments using statistical methods will greatly help in improving the reliability of this approach.

4.4.11 Pharmacological Screening of Potential Drug Compounds[139]

The National Cancer Institute (NCI) has screened more than 60,000 compounds against a panel of 60 human cancer cell lines. The extent to which any single compound inhibits growth in any given cell line is simply one data point relevant to that compound-cell line combination—namely the concentration associated with a 50 percent inhibition in the growth of that cell line. However, the pattern of such values across all 60 cell lines can provide insight into the mechanisms of drug action and drug resistance. Combined with molecular structure data, these activity patterns can be used to explore the NCI database of 460,000 compounds for growth-inhibiting effects in these cell lines, and can also provide insight into potential target molecules and modulators of activity in the 60 cell lines. Based on this approach, five compounds have been screened in this manner and selected for entry into clinical trials.

This approach to drug discovery and molecular pharmacology serves a number of useful functions. According to Weinstein et al.,

> (i) It suggests novel targets and mechanisms of action or modulation.
> (ii) It detects inhibition of integrated biochemical pathways not adequately represented by any single molecule or molecular interaction. (This feature of cell-based assays is likely to be more important in the development of therapies for cancer than it is for most other diseases; in the case of cancer, one is fighting the plasticity of a poorly controlled genome and the selective evolutionary pressures for development of drug resistance.)
> (iii) It provides candidate molecules for secondary testing in biochemical assays; conversely, it provides a well-characterized biological assay in vitro for compounds emerging from biochemical screens.
> (iv) It ''fingerprints'' tested compounds with respect to a large number of possible targets and modulators of activity.
> (v) It provides such fingerprints for all previously tested compounds whenever a new target is assessed in many or all of the 60 cell lines. (In contrast, if a battery of assays for different biochemical targets were applied to, for example, 60,000 compounds, it would be necessary to retest all of the compounds for any new target or assay.)
> (vi) It links the molecular pharmacology with emerging databases on molecular markers in microdissected human tumors—which, under the rubric of this article, constitute clinical (C) databases.
> (vii) It provides the basis for pharmacophore development and searches of an S [structure] database for additional candidates. If an agent with a desired action is already known, its fingerprint patterns of activity can be used by . . . [various] pattern-recognition technologies to find similar compounds.

Box 4.6 provides an example of this approach.

4.4.12 Algorithms Related to Imaging

Biological science is rich in images. Most familiar are images taken through optical microscopes, but there are many other imaging modalities—electron microscopes, computed tomography scans, X-rays, magnetic resonance imaging, and so on. For most of the history of life science research, images have

[139] Section 4.4.11 is based heavily on J.N. Weinstein, T.G. Myers, P.M. O'Connor, S.H. Friend, A.J. Fornace, Jr., K.W. Kohn, T. Fojo, et al., "An Information-Intensive Approach to the Molecular Pharmacology of Cancer," *Science* 275(5298):343-349, 1997.

Box 4.6
An Information-intensive Approach to Cancer Drug Discovery

Given one compound as a "seed," [an algorithm known as] COMPARE searches the database of screened agents for those most similar to the seed in their patterns of activity against the panel of 60 cell lines. Similarity in pattern often indicates similarity in mechanism of action, mode of resistance, and molecular structure. . . .

A formulation of this approach in terms of three databases [includes databases for] the activity patterns [A], . . . molecular structural features of the tested compounds [S], and . . . possible targets or modulators of activity in the cells [T]. . . . The (S) database can be coded in terms of any set of two-dimensional (2D) or 3D molecular structure descriptors. The NCI's Drug Information System (DIS) contains chemical connectivity tables for approximately 460,000 molecules, including the 60,000 tested to date. 3-D structures have been obtained for 97% of the DIS compounds, and a set of 588 bitwise descriptors has been calculated for each structure by use of the Chem-X computational chemistry package. This data set provides the basis for pharmacophoric searches; if a tested compound, or set of compounds, is found to have an interesting pattern of activity, its structure can be used to search for similar molecules in the DIS database.

In the target (T) database, each row defines the pattern (across 60 cell lines) of a measured cell characteristic that may mediate, modulate, or otherwise correlate with the activity of a tested compound. When the term is used in this general shorthand sense, a "target" may be the site of action or part of a pathway involved in a cellular response. Among the potential targets assessed to date are oncogenes, tumor-suppressor genes, drug resistance-mediating transporters, heat shock proteins, telomerase, cytokine receptors, molecules of the cell cycle and apoptotic pathways, DNA repair enzymes, components of the cytoarchitecture, intracellular signaling molecules, and metabolic enzymes.

In addition to the targets assessed one at a time, others have been measured en masse as part of a protein expression database generated for the 60 cell lines by 2D polyacrylamide gel electrophoresis.

Each compound displays a unique "fingerprint" pattern, defined by a point in the 60D space (one dimension for each cell line) of possible patterns. In information theoretic terms, the transmission capacity of this communication channel is very large, even after one allows for experimental noise and for biological realities that constrain the compounds to particular regions of the 60D space. Although the activity data have been accumulated over a 6-year period, the experiments have been reproducible enough to generate . . . patterns of coherence.

SOURCE: Reprinted by permission from J.N. Weinstein, T.G. Myers, P.M. O'Connor, S.H. Friend, A.J. Fornace, Jr., K.W. Kohn, T. Fojo, et al., "An Information-intensive Approach to the Molecular Pharmacology of Cancer," *Science* 275(5298):343-349, 1997. Copyright 1997 AAAS.

been a source of qualitative insight.[140] While this is still true, there is growing interest in using image data more quantitatively.

Consider the following applications:

- Automated identification of fungal spores in microscopic digital images and automated estimation of spore density;[141]
- Automated analysis of liver MRI images from patients with putative hemochromatosis to determine the extent of iron overload, avoiding the need for an uncomfortable liver biopsy;[142]

[140] Note also that biological imaging itself is a subset of the intersection between biology and visual techniques. In particular, other biological insight can be found in techniques that consider spectral information, e.g., intensity as a function of frequency and perhaps a function of time. Processing microarray data (discussed further in Section 7.2.1) ultimately depends on the ability to extract interesting signals from patterns of fluorescing dots, as does quantitative comparison of patterns obtained in two-dimensional polyacrylamide gel electrophoresis. (See S. Veeser, M.J. Dunn, and G.Z. Yang, "Multiresolution Image Registration for Two-dimensional Gel Electrophoresis," *Proteomics* 1(7):856-870, 2001, available at http://vip.doc.ic.ac.uk/2d-gel/2D-gel-final-revision.pdf.)

[141] T. Bernier and J.A. Landry, "Algorithmic Recognition of Biological Objects," *Canadian Agricultural Engineering* 42(2):101-109, 2000.

[142] George Reeke, Rockefeller University, personal communication to John Wooley, October 8, 2004.

Box 4.7
The Open Microscopy Environment[1]

Responding to the need to manage a large number of multispectral movies of mitotic cells in the late 1990s, Sorger and Swedlow began work on the open microscopy environment (OME). The OME is designed as infrastructure that manages optical microscopy images, storing both the primary image data and appropriate metadata on those images, including data on the optics of the microscope, the experimental setup and sample, and information derived by analysis of the images. OME also permits data federation that allows information from multiple sources (e.g., genomic or chemical databases) to be linked to image records.

In addition, the OME provides an extensible environment that enables users to write their own applications for image analysis. Consider, for example, the task of tracking labeled vesicles in a time-lapse movie. As noted by Swedlow et al., this problem requires the following: a segmentation algorithm to find the vesicles and to produce a list of centroids, volumes, signal intensities, and so on; a tracker to define trajectories by linking centroids at different time points according to a predetermined set of rules; and a viewer to display the analytic results overlaid on the original movie.[2]

OME provides a mechanism for linking together various analytical modules by specifying data semantics that enable the output of one module to be accepted as input to another. These semantic data types of OME describe analytic results such as "centroid," "trajectory," and "maximum signa," and allow users, rather than a predefined standard, to define such concepts operationally, including in the machine-readable definition and the processing steps that produce it (e.g., the algorithm and the various parameter settings used).

[1] See www.openmicroscopy.org.

[2] J.R. Swedlow, I. Goldberg, E. Brauner, and P.K. Sorger, "Informatics and Quantitative Analysis in Biological Imaging," *Science* 300(5616):100-102, 2003.

SOURCE: Based largely on the paper by Swedlow et al. cited in Footnote145 and on the OME Web page at www.openmicroscopy.org.

- Fluorescent speckle microscopy, a technique for quantitatively tracking the movement, assembly, and disassembly of macromolecules in vivo and in vitro, such as those involved in cytoskeleton dynamics;[143] and
- Establishing metrics of similarity between brain images taken at different times.[144]

These applications are only an infinitesimal fraction of those that are possible. Several research areas associated with increasing the utility of biological images are discussed below. Box 4.7 describes the open microscopy environment, an effort intended to automate image analysis, modeling, and mining of large sets of biological images obtained from optical microscopy.[145]

As a general rule, biologists need to develop better imaging methods that are applicable across the entire spatial scale of interest, from the subcellular to the organismal. (In this context, "better" means imaging that occurs in real time (or nearly so) with the highest possible spatial and temporal resolution.) These methods will require new technologies (such as the multiphoton microscope) and also new protein and nonprotein reporter molecules that can be expressed or introduced into cells or organisms.

[143] C.M. Waterman-Storer and G. Danuser, "New Directions for Fluorescent Speckle Microscopy," *Current Biology* 12(18):R633-R640, 2002.

[144] M.I. Miller, A. Trouve, and L. Younes, "On the Metrics and Euler-Lagrange Equations of Computational Anatomy," *Annual Review of Biomedical Engineering* 4:375-405, 2002, available at http://www.cis.jhu.edu/publications/papers_in_database/EulerLagrangeEqnsCompuAnatomy.pdf.

[145] J.R. Swedlow, I. Goldberg, E. Brauner, and P.K. Sorger, "Informatics and Quantitative Analysis in Biological Imaging," *Science* 300(5616):100-102, 2003.

The discussion below focuses only on a narrow slice of the very general problem of biological imaging, as a broader discussion would go beyond the scope of this report.

4.4.12.1 Image Rendering[146]

Images have been central to the study of biological phenomena ever since the invention of the microscope. Today, images can be obtained from many sources, including tomography, MRI, X-rays, and ultrasound. In many instances, biologists are interested in the spatial and geometric properties of components within a biological entity. These properties are often most easily understood when viewed through an interactive visual representation that allows the user to view the entity from different angles and perspectives. Moreover, a single analysis or visualization session is often not sufficient, and processing across many image volumes is often required.

The requirement that a visual representation be interactive places enormous demands on the computational speed of the imaging equipment in use. Today, the data produced by imaging equipment are quickly outpacing the capabilities offered by the image processing and analysis software currently available. For example, the GE EVS-RS9 CT scanner is able to generate image volumes with resolutions in the 20-90 mm range, which results in a dataset size of multiple gigabytes. Datasets of such size require software tools specifically designed for the imaging datasets of today and tomorrow (see Figure 4.5) so that researchers can identify subtle features that can otherwise be missed or misrepresented. Also with increasing dataset resolution comes increasing dataset size, which translates directly to lengthening dataset transfer, processing, and visualization times.

New algorithms that take advantage of state-of-the-art hardware in both relatively inexpensive workstations and multiprocessor supercomputers must be developed and moved into easy-to-access software systems for the clinician and researcher. An example is ray-tracing, a method commonly used in computer graphics that supports highly efficient implementations on multiple processors for interactive visualization. The resulting volume rendition permits direct inspection of internal structures, without a precomputed segmentation or surface extraction step, through the use of multidimensional transfer functions. As seen in the visualizations in Figure 4.6, the resolution of the CT scan allows subtleties such as the definition of the cochlea, the modiolus, the implanted electrode array, and the lead wires that connect the array to a head-mounted connector. The co-linear alignment of the path of the cochlear nerve with the location of the electrode shanks and tips is the necessary visual confirmation of the correct surgical placement of the electrode array.

In both of the studies described in Figure 4.5 and Figure 4.6, determination of three-dimensional structure and configuration played a central role in biological inquiry. Volume visualization created detailed renderings of changes in bone morphology due to a Pax3 mutation in mice, and it provided visual confirmation of the precise location of an electrode array implanted in the feline skull. The scientific utility of volume visualization will benefit from further improvements in its interactivity and flexibility, as well as simultaneous advances in high-resolution image acquisition and the development of volumetric image-processing techniques for better feature extraction and enhancement.

4.4.12.2 Image Segmentation[147]

An important problem in automated image analysis is image segmentation. Digital images are recorded as a set of pixels in a two- or three-dimensional array. Images that represent natural scenes usually contain different objects, so that, for example, a picture of a park may depict people, trees, and

[146] Section 4.4.12.1 is based on material provided by Chris Johnson, University of Utah.

[147] Section 4.4.11.2 is adapted from and includes excerpts from National Research Council, *Mathematics and Physics of Emerging Biomedical Imaging*, National Academy Press, Washington, DC, 1996.

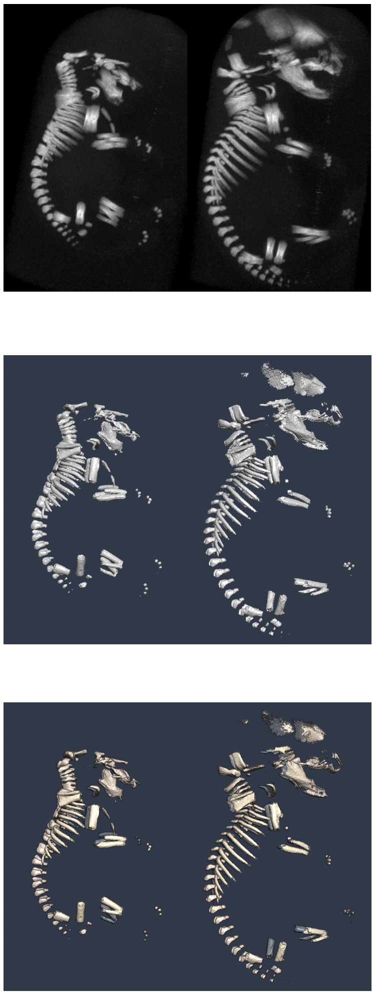


FIGURE 4.5 Visualizations of mutant (*left*) and normal (*right*) mice embryos.

CT values are inspected by maximum intensity projection in (a) and with standard isosurface rendering in (b). Volume rendering (c) using multidimensional opacity functions allows more accurate bone emphasis, depth cueing, and curvature-based transfer functions to enhance bone contours in image space. In this case, Drs. Keller and Capecchi are investigating the birth defects caused by a mutation in the Pax3 gene, which controls musculoskeletal development in mammalian embryos. In their model, they have activated a dominantly acting mutant Pax3 gene and have uncovered two of its effects: (1) abnormal formation of the bones of the thoracolumbar spine and cartilaginous rib cage and (2) cranioschisis, a more drastic effect in which the dermal and skeletal covering of the brain is missing. Imaging of mutant and normal mouse embryos was performed at the University of Utah Small Animal Imaging Facility, producing two 1.2 GB 16-bit volumes of 769 × 689 × 1173 samples, with resolution of 21 × 21 × 21 microns.

SOURCE: Courtesy of Chris Johnson, University of Utah; see also http://www.sci.utah.edu/stories/2004/spr_imaging.html.

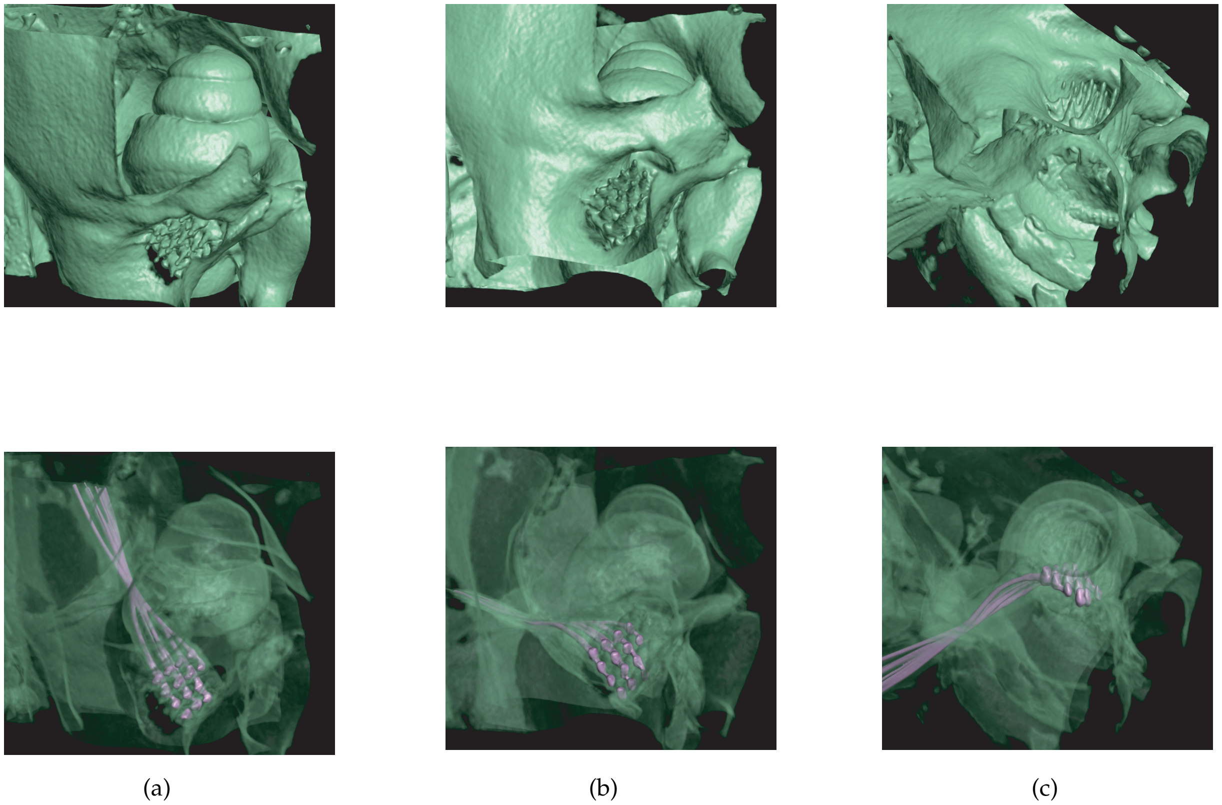


FIGURE 4.6 Volume renderings of electrode array implanted in feline skull.

In this example, scanning produced a 131 MB 16-bit volume of 425 × 420 × 385 samples, with resolution of 21 × 21 × 21 microns. Renderings of the volume were generated using a ray-tracing algorithm across multiple processors allowing interactive viewing of this relatively large dataset. The resolution of the scan allows definition of the shanks and tips of the implanted electrode array. Volumetric image processing was used to isolate the electrode array from the surrounding tissue, highlighting the structural relationship between the implant and the bone. There are distinct CT values for air, soft tissue, bone, and the electrode array, enabling the use of a combination of ray tracing and volume rendering to visualize the array in the context of the surrounding structures, specifically the bone surface. The volume is rotated gradually upward in columns (a), (b), and (c), from seeing the side of the cochlea exterior in (a), to looking down the path of the cochlear nerve in (c). From top to bottom, each row uses different rendering styles: (1), summation projections of CT values (green) and gradients (magenta); (2), volume renderings with translucent bone, showing the electrode leads in magenta.

SOURCE: Courtesy of Chris Johnson, University of Utah; see also http://www.sci.utah.edu/stories/2004/spr_imaging.html.

benches. Similarly, a scanned image of a magazine page may contain text and graphics (e.g., a picture of a park). Segmentation refers to the process by which an object (or characteristics of the object) in an image is extracted from image data for purposes of visualization and measurement. (Extraction means that the pixels associated with the object of interest are isolated.) In a biological context, a typical problem in image segmentation might involve extracting different organs in a CT scan of the body. Segmentation research involves the development of automatic, computer-executable rules that can isolate enough of these pixels to produce an acceptably accurate segmentation. Segmentation is a central problem of image analysis because segmentation must be accomplished before many other interesting problems in image analysis can be solved, including image registration, shape analysis, and volume and area estimation. A specific laboratory example would be the segmentation of spots on two-dimensional electrophoresis gels.

There is no common method or class of methods applicable to even the majority of images. Segmentation is easiest when the objects of interest have intensity or edge characteristics that allow them to be separated from the background and noise, as well as from each other. For example, an MRI image of the human body would be relatively easy to segment for bones: all pixels with intensity below a given threshold would be eliminated, leaving mostly the pixels associated with high-signal-intensity bone.

Generally, edge detection depends on a search for intensity gradients. However, it is difficult to find gradients when, as is usually the case in biomedical images, intensities change only gradually between the structure of interest and the surrounding structure(s) from which it is to be extracted. Continuity and connectivity are important criteria for separating objects from noise and have been exploited quite widely.

A number of different approaches to image segmentation are described in more detail by Pham et al.[148]

4.4.12.3 Image Registration[149]

Different modes of imaging instrumentation may be used on the same object because they are sensitive to different object characteristics. For example, an X-ray of an individual will produce different information than a CT scan. For various purposes, and especially for planning surgical and radiation treatment, it can be important for these images to be aligned with each other, that is, for information from different imaging modes to be displayed in the same locations. This process is known as image registration.

There are a variety of techniques for image registration, but in general they can be classified based on the features that are being matched. For example, such features may be external markers that are fixed (e.g., on a patient's body), internal anatomic markers that are identifiable on all images, the center of gravity for one or more objects in the images, crestlines of objects in the images, or gradients of intensity. Another technique is minimization of the distance between corresponding surface points of a predefined object. Image registration often depends on the identification of similar structures in the images to be registered. In the ideal case, this identification can be performed through an automated segmentation process.

Image registration is well defined for rigid objects but is more complicated for deformable objects or for objects imaged from different angles. When soft tissue deforms (e.g., because a patient is lying on his side rather than on his back), elastic warping is required to transform one dataset into the other. The difficulty lies in defining enough common features in the images to enable specifying appropriate local deformations.

An example of an application in which image registration is important is the Cell-Centered Database (CCDB).[150] Launched in 2002, the CCDB contains structural and protein distribution information derived from confocal, multiphoton, and electron microscopy for use by the structural biology and neuroscience communities. In the case of neurological images, most of the imaging data are referenced to a higher level of brain organization by registering their location in the coordinate system of a standard brain atlas. Placing data into an atlas-based coordinate system provides one method by which data taken across scales

[148] D.L. Pham, C. Xu, and J.L. Prince, "Current Methods in Medical Image Segmentation," *Annual Review of Biomedical Engineering* 2:315-338, 2000.

[149] Section 4.4.12.3 is adapted from National Research Council, *Mathematics and Physics of Emerging Biomedical Imaging,* National Academy Press, Washington, DC, 1996.

[150] See M.E. Martone, S.T. Peltier, and M.H. Ellisman, "Building Grid Based Resources for Neurosciences," unpublished paper 2003, National Center for Microscopy and Imaging Research, Department of Neurosciences, University of California, San Diego, San Diego, CA, and http://ccdb.ucsd.edu/CCDB/about.shtml.

and distributed across multiple resources can be compared reliably. Through the use of atlases and tools for surface warping and image registration, it is possible to express the location of anatomical features or signals in terms of a standardized and quantitative coordinate system, rather by using terms that describe objects in the field of view. The expression of brain data in terms of atlas coordinates also allows it to be transformed spatially to provide alternative views that may offer additional information (e.g., flat maps or additional parcellation schemes). Finally, a standard coordinate system allows the same brain region to be sampled repeatedly to allow data to be accumulated over time.

4.4.12.4 Image Classification

Image classification is the process through which a set of images can be sorted into meaningful categories. Categories can be defined through low-level features such as color mix and texture patterns or through high-level features such as objects depicted. As a rule, low-level features can be computed with little difficulty, and a number of systems have been developed that take advantage of such features.[151]

However, users are generally much more interested in semantic content that is not easily represented in such low-level features. The easiest method to identify interesting semantic content is simply to annotate an image manually with text, although this process is quite tedious and is unlikely to capture the full range of content in an image. Thus, automated techniques hold considerable interest.

The general problem of automatic identification of such image content has not been solved. One approach described by Huang et al. relies on supervised learning to classify images hierarchically.[152] This approach relies on using good low-level features and then performing feature-space reconfiguration using singular value decomposition to reduce noise and dimensionality. A hierarchical classification tree can be generated from training data and subsequently used to sort new images into categories.

A second approach is based on the fact that biological images often contain branching structures. (For example, both muscle and neural tissue contain blood vessels and dendrites that are found in branching structures.) The fractal dimensionality of such structures can then be used as a measure of similarity, and images that contain structures of similar fractal dimension can be grouped into categories.[153]

4.5 DEVELOPING COMPUTATIONAL TOOLS

The computational tools described above were once gleams in the eye of some researcher. Despite the joy and satisfaction felt when a prototype program supplies the first useful results to its developer, it is a long, long way to converting that program into a genuine product that is general, robust, and useful to others. Indeed, in his classic text *The Mythical Man-Month* (Addison-Wesley, Reading, MA, 1995), Frederick P. Brooks, Jr., estimates the difference in effort necessary to create a programming systems product from a program as an order of magnitude.

Some of the software engineering considerations necessary to turn a program into a product include the following:

- *Quality.* The program, of course, must be as free of defects as possible, not only in the sense of running without faults, but also of precisely implementing the stated algorithm. It must be tested for all

[151] See, for example, M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker, "Query by Image and Video Content: The QBIC System," *IEEE Computer* 28(9):23-32, 1995, available at http://wwwqbic.almaden.ibm.com/.

[152] J. Huang, S.R. Kumar, and R. Zabih, "An Automatic Hierarchical Image Classification Scheme," ACM Conference on Multimedia, Bristol, England, September 1998. A revised version appears in *EURASIP Journal on Applied Signal Processing*, 2003, available at http://www.cs.cornell.edu/rdz/Papers/Archive/mm98.pdf.

[153] D. Cornforth, H. Jelinek, and L. Peich, "Fractop: A Tool for Automated Biological Image Classification," available at http://csu.edu.au/~dcornfor/Fractop_v7.pdf.

potential inputs, and combinations of factors, and must be robust even in the face of invalid usage. The program should have well-understood and bounded resource demands, including memory, input-output, and processing time.

- *Maintenance.* When bugs are discovered, they must be tracked, patched, and provided to users. This often means that the code should be structured for maintainability; for example, Perl, which is extremely powerful, is often written in a way that is incomprehensible to programmers other than the author (and often even to the author). Differences in functionality between versions must be documented carefully.
- *Documentation.* If the program is to be usable by others, all of the functionality must be clearly documented, including data file formats, configuration options, output formats, and of course program usage. If the source code of the program is made available (as is often the case with scientific tools), the code must be documented in such a way that users can check the validity of the implementation as well as alter it to meet their needs.
- *User interface.* The program must have a user interface, although not necessarily graphical, that is unambiguous and able to access the full range of functions of the program. It should be easy to use, difficult to make mistakes, and clear in its instructions and display of state.
- *System integration and portability.* The program must be distributed to users in a convenient way, and be able to run on different platforms and operating systems in a way that does not interfere with existing software or system settings. It should be easily configurable and customizable for particular requirements, and should install easily without access to specialized software, such as nonstandard compilers.
- *General.* The program should accept a wide selection of data types, including common formats, units, precisions, ranges, and file sizes. The internal coding interfaces should have precisely defined syntax and semantics, so that users can easily extend the functionality or integrate it into other tools.

Tool developers address these considerations to varying degrees, and users may initially be more tolerant of something that is more program than product if the functionality it confers is essential and unique. Over time, however; such programs will eventually become more product-like because users will not tolerate significant inconvenience.

Finally, there is an issue of development methodology. A proprietary approach to development can be adopted for a number of competitive reasons, ranging from the ultimate desire to reap financial benefit to staying ahead of competing laboratories. Under a proprietary approach, source code for the tools would be kept private, so that potential competitors would be unable to exploit the code easily for their own purposes. (Source code is needed to make changes to a program.) An open approach to development calls for the source code to be publicly available, on the theory that broad community input strengthens the utility of the tools being made available and better enables one team to build on another team's work.

5

Computational Modeling and Simulation as Enablers for Biological Discovery

While the previous chapter deals with the ways in which computers and algorithms could support existing practices of biological research, this chapter introduces a different type of opportunity. The quantities and scopes of data being collected are now far beyond the capability of any human, or team of humans, to analyze. And as the sizes of the datasets continue to increase exponentially, even existing techniques such as statistical analysis begin to suffer. In this data-rich environment, the discovery of large-scale patterns and correlations is potentially of enormous significance. Indeed, such discoveries can be regarded as hypotheses asserting that the pattern or correlation may be important—a mode of "discovery science" that complements the traditional mode of science in which a hypothesis is generated by human beings and then tested empirically.

For exploring this data-rich environment, simulations and computer-driven models of biological systems are proving to be essential.

5.1 ON MODELS IN BIOLOGY

In all sciences, models are used to represent, usually in an abbreviated form, a more complex and detailed reality. Models are used because in some way, they are more accessible, convenient, or familiar to practitioners than the subject of study. Models can serve as explanatory or pedagogical tools, represent more explicitly the state of knowledge, predict results, or act as the objects of further experiments. Most importantly, a model is a representation of some reality that embodies some essential and interesting aspects of that reality, but not all of it.

Because all models are by definition incomplete, the central intellectual issue is whether the essential aspects of the system or phenomenon are well represented (the term "essential" has multiple meanings depending on what aspects of the phenomenon are of interest). In biological phenomena, what is interesting and significant is usually a set of relationships—from the interaction of two molecules to the behavior of a population in its environment. Human comprehension of biological systems is limited, among other things, by that very complexity and by the problems that arise when attempting to dissect a given system into simpler, more easily understood components. This challenge is compounded by our current inability to understand relationships between the components as they occur in reality, that is, in the presence of multiple, competing influences and in the broader context of time and space.

Different fields of science have traditionally used models for different purposes; thus, the nature of the models, the criteria for selecting good or appropriate models, and the nature of the abbreviation or simplification have varied dramatically. For example, biologists are quite familiar with the notion of model organisms.[1] A model organism is a species selected for genetic experimental analysis on the basis of experimental convenience, homology to other species (especially to humans), relative simplicity, or other attractive attributes. The fruit fly *Drosophila melanogaster* is a model organism attractive at least in part because of its short generational time span, allowing many generations in the course of an experiment.

At the most basic level, any abstraction of some biological phenomenon counts as a model. Indeed, the cartoons and block diagrams used by most biologists to represent metabolic, signaling, or regulatory pathways are models—qualitative models that lay out the connectivity of elements important to the phenomenon. Such models throw away details (e.g., about kinetics) implicitly asserting that omission of such details does not render the model irrelevant.

A second example of implicit modeling is the use of statistical tests by many biologists. All statistical tests are based on a null hypothesis, and all null hypotheses are based on some kind of underlying model from which the probability distribution of the null hypothesis is derived. Even those biologists who have never thought of themselves as modelers are using models whenever they use statistical tests.

Mathematical modeling has been an important component of several biological disciplines for many decades. One of the earliest quantitative biological models involved ecology: the Lotka-Volterra model of species competition and predator-prey relationships described in Section 5.2.4. In the context of cell biology, models and simulations are used to examine the structure and dynamics of a cell or organism's function, rather than the characteristics of isolated parts of a cell or organism.[2] Such models must consider stochastic and deterministic processes, complex pleiotropy, robustness through redundancy, modular design, alternative pathways, and emergent behavior in biological hierarchy.

In a cellular context, one goal of biology is to gain insight into the interactions, molecular or otherwise, that are responsible for the behavior of the cell. To do so, a quantitative model of the cell must be developed to integrate global organism-wide measurements taken at many different levels of detail.

The development of such a model is iterative. It begins with a rough model of the cell, based on some knowledge of the components of the cell and possible interactions among them, as well as prior biochemical and genetic knowledge. Although the assumptions underlying the model are insufficient and may even be inappropriate for the system being investigated, this rough model then provides a zeroth-order hypothesis about the structure of the interactions that govern the cell's behavior.

Implicit in the model are predictions about the cell's response under different kinds of perturbation. Perturbations may be genetic (e.g., gene deletions, gene overexpressions, undirected mutations) or environmental (e.g., changes in temperature, stimulation by hormones or drugs). Perturbations are introduced into the cell, and the cell's response is measured with tools that capture changes at the relevant levels of biological information (e.g., mRNA expression, protein expression, protein activation state, overall pathway function). Box 5.1 provides some additional detail on cellular perturbations.

The next step is comparison of the model's predictions to the measurements taken. This comparison indicates where and how the model must be refined in order to match the measurements more closely. If the initial model is highly incomplete, measurements can be used to suggest the particular components required for cellular function and those that are most likely to interact. If the initial model is relatively well defined, its predictions may already be in good qualitative agreement with measurement, differing only in minor quantitative ways. When model and measurement disagree, it is often

[1] See, for example, http://www.nih.gov/science/models for more information on model organisms.

[2] Section 5.1 draws heavily on excerpts from T. Ideker, T. Galitski, and L. Hood, "A New Approach to Decoding Life: Systems Biology," *Annual Review of Genomics and Human Genetics* 2:343-372, 2001; and H. Kitano, "Systems Biology: A Brief Overview," *Science* 295(5560):1662-1664, 2002.

Box 5.1
Perturbation of Biological Systems

Perturbation of biological systems can be accomplished through a number of genetic mechanisms, such as the following:

- *High-throughput genomic manipulation*. Increasingly inexpensive and highly standardized tools are available that enable the disruption, replacement, or modification of essentially any genomic sequence. Furthermore, these tools can operate simultaneously on many different genomic sequences.
- *Systematic gene mutations*. Although random gene mutations provide a possible set of perturbations, the random nature of the process often results in nonuniform coverage of possible genotypes—some genes are targeted multiple times, others not at all. A systematic approach can cover all possible genotypes and the coverage of the genome is unambiguous.
- *Gene disruption*. While techniques of genomic manipulation and systematic gene mutation are often useful in analyzing the behavior of model organisms such as yeast, they are not practical for application to organisms of greater complexity (i.e., higher eukaryotes). On the other hand, it is often possible to induce disruptions in the function of different genes, effectively silencing (or deleting) them to produce a biologically significant perturbation.

SOURCE: Adapted from T. Ideker, T. Galitski, and L. Hood, "A New Approach to Decoding Life: Systems Biology," *Annual Review of Genomics and Human Genetics* 2:343-372, 2001.

necessary to create a number of more refined models, each incorporating a different mechanism underlying the discrepancies in measurement.

With the refined model(s) in hand, a new set of perturbations can be applied to the cell. Note that new perturbations are informative only if they elicit different responses between models, and they are most useful when the predictions of the different models are very different from one another. Nevertheless, a new set of perturbations is required because the predictions of the refined model(s) will generally fit well with the old set of measurements.

The refined model that best accounts for the new set of measurements can then be regarded as the initial model for the next iteration. Through this process, model and measurement are intended to converge in such a way that the model's predictions mirror biological responses to perturbation. Modeling must be connected to experimental efforts so that experimentalists will know what needs to be determined in order to construct a comprehensive description and, ultimately, a theoretical framework for the behavior of a biological system. Feedback is very important, and it is this feedback, along with the global—or, loosely speaking, genomic-scale—nature of the inquiry that characterizes much of 21st century biology.

5.2 WHY BIOLOGICAL MODELS CAN BE USEFUL

In the last decade, mathematical modeling has gained stature and wider recognition as a useful tool in the life sciences. Most of this revolution has occurred since the era of the genome, in which biologists were confronted with massive challenges to which mathematical expertise could successfully be brought to bear. Some of the success, though, rests on the fact that computational power has allowed scientists to explore ever more complex models in finer detail. This means that the mathematician's talent for abstraction and simplification can be complemented with realistic simulations in which details not amenable to analysis can be explored. The visual real-time simulations of modeled phenomena give more compelling and more accessible interpretations of what the models predict.[3] This has made it easier to earn the recognition of biologists.

On the other hand, modeling—especially computational modeling—should not be regarded as an intellectual panacea, and models may prove more hindrance than help under certain circumstances. In models with many parameters, the state space to be explored may grow combinatorially fast so that no amount of data and brute force computation can yield much of value (although it may be the case that some algorithm or problem-related insight can reduce the volume of state space that must be explored to a reasonable size). In addition, the behavior of interest in many biological systems is not characterized as equilibrium or quasi-steady-state behavior, and thus convergence of a putative solution may never be reached. Finally, modeling presumes that the researcher can both identify the important state variables and obtain the quantitative data relevant to those variables.[4]

Computational models apply to specific biological phenomena (e.g., organisms, processes) and are used for a number of purposes as described below.

5.2.1 Models Provide a Coherent Framework for Interpreting Data

A biologist surveys the number of birds nesting on offshore islands and notices that the number depends on the size (e.g., diameter) of the island: the larger the diameter d, the greater is the number of nests N. A graph of this relationship for islands of various sizes reveals a trend. Here the mathematically informed and uninformed part ways: simple linear least-squares fit of the data misses a central point. A trivial "null model" based on an equal subdivision of area between nesting individuals predicts that $N \sim d^2$, (i.e., the number of nests should be roughly proportional to the square of island area). This simple geometric property relating area to population size gives a strong indication of the trend researchers should expect to see. Departures from this trend would indicate that something else may be important. (For example, different parts of islands are uninhabitable, predators prefer some islands to others, and so forth.)

Although the above example is elementary, it illustrates the idea that data are best interpreted within a context that shapes one's expectations regarding what the data "ought" to look like; often a mathematical (or geometric) model helps to create that context.

5.2.2 Models Highlight Basic Concepts of Wide Applicability

Among the earliest applications of mathematical ideas to biology are those in which population levels were tracked over time and attempts were made to understand the observed trends. Malthus proposed in 1798 the fitting of population data to exponential growth curves following his simple model for geometric growth of a population.[5] The idea that simple reproductive processes produce

[3] As one example, Ramon Felciano studied the use of "domain graphics" by biologists. Felciano argued that certain visual representations (known as domain graphics) become so ingrained in the discourse of certain subdisciplines of biology that they become good targets for user interfaces to biological data resources. Based on this notion, Felciano constructed a reusable interface based on the standard two-dimensional layout of RNA secondary structure. See R. Felciano, R. Chen, and R. Altman, "RNA Secondary Structure as a Reusable Interface to Biological Information Resources," *Gene* 190:59-70, 1997.

[4] In some cases, obtaining the quantitative data is a matter of better instrumentation and higher accuracy. In other cases, the data are not available in any meaningful sense of practice. For example, Richard Lewontin notes that the probability of survival P_s of a particular genotype is an ensemble property, rather than the property of a single individual who either will or will not survive. But if what is of interest is P_s as a function of the alternative genotypes deriving from a single locus, the effects of the impacts deriving from other loci must be randomized. However, in sexually reproducing organisms, there is no way known to produce an ensemble of individuals that are all identical with respect to a single locus but randomized over other loci. Thus, a quantitative characterization of P_s is in practice not possible, and no alternative measurement technologies will be of much value in solving this problem. See R. Lewontin, *The Genetic Basis of Evolutionary Change*, Columbia University Press, New York, 1974.

[5] T.R. Malthus, *An Essay on the Principle of Population*, First Edition, E.A. Wrigley and D. Souden, eds., Penguin Books, Harmondsworth, England, 1798.

exponential growth (if birth rates exceed mortality rates) or extinction (in the opposite case) is a fundamental principle: its applicability in biology, physics, chemistry, as well as simple finance, is central.

An important refinement of the Malthus model was proposed in 1838 to explain why most populations do not experience exponential growth indefinitely. The refinement was the idea of the density-dependent growth law, now known as the logistic growth model.[6] Though simple, the Verhulst model is still used widely to represent population growth in many biological examples. Both Malthus and Verhulst models relate observed trends to simple underlying mechanisms; neither model is fully accurate for real populations, but deviations from model predictions are, in themselves, informative, because they lead to questions about what features of the real systems are worthy of investigation.

More recent examples of this sort abound. Nonlinear dynamics has elucidated the tendency of excitable systems (cardiac tissue, nerve cells, and networks of neurons) to exhibit oscillatory, burst, and wave-like phenomena. The understanding of the spread of disease in populations and its sensitive dependence on population density arose from simple mathematical models. The same is true of the discovery of chaos in the discrete logistic equation (in the 1970s). This simple model and its mathematical properties led to exploration of new types of dynamic behavior ubiquitous in natural phenomena. Such biologically motivated models often cross-fertilize other disciplines: in this case, the phenomenon of chaos was then found in numerous real physical, chemical, and mechanical systems.

5.2.3 Models Uncover New Phenomena or Concepts to Explore

Simple conceptual models can be used to uncover new mechanisms that experimental science has not yet encountered. The discovery of chaos mentioned above is one of the clearest examples of this kind. A second example of this sort is Turing's discovery that two chemicals that interact chemically in a particular way (activate and inhibit one another) and diffuse at unequal rates could give rise to "peaks and valleys" of concentration. His analysis of reaction-diffusion (RD) systems showed precisely what ranges of reaction rates and rates of diffusion would result in these effects, and how properties of the pattern (e.g., distance between peaks and valleys) would depend on those microscopic rates. Later research in the mathematical community also uncovered how other interesting phenomena (traveling waves, oscillations) were generated in such systems and how further details of patterns (spots, stripes, etc.) could be affected by geometry, boundary conditions, types of chemical reactions, and so on.

Turing's theory was later given physical manifestation in artificial chemical systems, manipulated to satisfy the theoretical criteria of pattern formation regimes. And, although biological systems did not produce simple examples of RD pattern formation, the theoretical framework originating in this work motivated later more realistic and biologically based modeling research.

5.2.4 Models Identify Key Factors or Components of a System

Simple conceptual models can be used to gain insight, develop intuition, and understand "how something works." For example, the Lotka-Volterra model of species competition and predator-prey[7] is largely conceptual and is recognized as not being very realistic. Nevertheless, this and similar models have played a strong role in organizing several themes within the discipline: for example, competitive exclusion, the tendency for a species with a slight advantage to outcompete, dominate, and take over from less advantageous species; the cycling behavior in predator-prey interactions; and the effect of

[6] P.F. Verhulst, "Notice sur la loi que la population suit dans son accroissement," *Correspondence Mathématique et Physique*, 1838.

[7] A.J. Lotka, *Elements of Physical Biology*, Williams & Wilkins Co., Baltimore, MD, 1925; V. Volterra, "Variazioni e fluttuazioni del numero d'individui in specie animali conviventi," *Mem. R. Accad. Naz. dei Lincei.*, Ser. VI, Vol. 2, 1926. The Lotka-Volterra model is a set of coupled differential equations that relate the densities of prey and predator given parameters involving the predator-free rate of prey population increase, the normalized rate at which predators can successfully remove prey from the population, the normalized rate at which predators reproduce, and the rate at which predators die.

resource limitations on stabilizing a population that would otherwise grow explosively. All of these concepts arose from mathematical models that highlighted and explained dynamic behavior within the context of simple models. Indeed, such models are useful for helping scientists to recognize patterns and predict system behavior, at least in gross terms and sometimes in detail.

5.2.5 Models Can Link Levels of Detail (Individual to Population)

Biological observations are made at many distinct hierarchies and levels of detail. However, the links between such levels are notoriously difficult to understand. For example, the behavior of single neurons and their response to inputs and signaling from synaptic connections might be well known. The behavior of a large assembly of such neurons in some part of the central nervous system can be observed macroscopically by imaging or electrode recording techniques. However, how the two levels are interconnected remains a massive challenge to scientific understanding. Similar examples occur in countless settings in the life sciences: due to the complexity of nonlinear interactions, it is nearly impossible to grasp intuitively how collections of individuals behave, what emergent properties of these groups arise, or the significance of any sensitivity to initial conditions that might be magnified at higher levels of abstraction. Some mathematical techniques (averaging methods, homogenization, stochastic methods) allow the derivation of macroscopic statements based on assumptions at the microscopic, or individual, level. Both modeling and simulation are important tools for bridging this gap.

5.2.6 Models Enable the Formalization of Intuitive Understandings

Models are useful for formalizing intuitive understandings, even if those understandings are partial and incomplete. What appears to be a solid verbal argument about cause and effect can be clarified and put to a rigorous test as soon as an attempt is made to formulate the verbal arguments into a mathematical model. This process forces a clarity of expression and consistency (of units, dimensions, force balance, or other guiding principles) that is not available in natural language. As importantly, it can generate predictions against which intuition can be tested.

Because they run on a computer, simulation models force the researcher to represent explicitly important components and connections in a system. Thus, simulations can only complement, but never replace, the underlying formulation of a model in terms of biological, physical, and mathematical principles. That said, a simulation model often can be used to indicate gaps in one's knowledge of some phenomenon, at which point substantial intellectual work involving these principles is needed to fill the gaps in the simulation.

5.2.7 Models Can Be Used as a Tool for Helping to Screen Unpromising Hypotheses

In a given setting, quantitative or descriptive hypotheses can be tested by exploring the predictions of models that specify precisely what is to be expected given one or another hypothesis. In some cases, although it may be impossible to observe a sequence of biological events (e.g., how a receptor-ligand complex undergoes sequential modification before internalization by the cell), downstream effects may be observable. A model can explore the consequences of each of a variety of possible sequences can and help scientists to identify the most likely candidate for the correct sequence. Further experimental observations can then refine one's understanding.

5.2.8 Models Inform Experimental Design

Modeling properly applied can accelerate experimental efforts at understanding. Theory embedded in the model is an enabler for focused experimentation. Specifically, models can be used alongside experiments to help optimize experimental design, thereby saving time and resources. Simple models give a framework for observations (as noted in Section 5.2.1) and thereby suggest what needs to be measured experimentally and, indeed, what need not be measured—that is how to refine the set of observations so as to extract optimal knowledge about the system. This is particularly true when models and experiments go hand-in-hand. As a rule, several rounds of modeling and experimentation are necessary to lead to informative results.

Carrying these general observations further, Selinger et al.[8] have developed a framework for understanding the relationship between the properties of certain kinds of models and the experimental sampling required for "completeness" of the model. They define a model as a set of rules that maps a set of inputs (e.g., possible descriptions of a cell's environment) to a set of outputs (e.g., the resulting concentrations of all of the cell's RNAs and proteins). From these basic properties, Selinger et al. are able to determine the order of magnitude of the number of measurements needed to populate the space of all possible inputs (e.g., environmental conditions) with enough measured outputs (e.g., transcriptomes, proteomes) to make prediction feasible, thereby establishing how many measurements are needed to adequately sample input space to allow the rule parameters to be determined.

Using this framework, Salinger et al. estimate the experimental requirements for the completeness of a discrete transcriptional network model that maps all *N* genes as inputs to all *N* genes as outputs in which the genes can take on three levels of expression (low, medium, and high) and each gene has, at most, *K* direct regulators. Applying this model to three organisms—*Mycoplasma pneumoniae*, *Escherichia coli*, and *Homo sapiens*—they find that 80, 40,000, and 700,000 transcriptome experiments, respectively, are necessary to fill out this model. They further note that the upper-bound estimate of experimental requirements grows exponentially with the maximum number of regulatory connections *K* per gene, although genes tend to have a low *K*, and that the upper-bound estimate grows only logarithmically with the number of genes *N*, making completeness feasible even for large genetic networks.

5.2.9 Models Can Predict Variables Inaccessible to Measurement

Technological innovation in scientific instrumentation has revolutionized experimental biology. However, many mysteries of the cell, of physiology, of individual or collective animal behavior, and of population-level or ecosystem-level dynamics remain unobservable. Models can help link observations to quantities that are not experimentally accessible. At the scale of a few millimeters, Marée and Hogeweg recently developed[9] a computational model based on a cellular automaton for the behavior of the social amoeba *Dictyostelium discoideum.* Their model is based on differential adhesion between cells, cyclic adenosine monophosphate (cAMP) signaling, cell differentiation, and cell motion. Using detailed two- and three-dimensional simulations of an aggregate of thousands of cells, the authors showed how a relatively small set of assumptions and "rules" leads to a fully accurate developmental pathway. Using the simulation as a tool, they were able to explore which assumptions were blatantly inappropriate (leading to incorrect outcomes). In its final synthesis, the Marée-Hogeweg model predicts dynamic distributions of chemicals and of mechanical pressure in a fully dynamic simulation of the culminating *Dictyostelium* slug. Some, but not all, of these variables can be measured experimentally: those that are measurable are well reproduced by the model. Those that cannot (yet) be measured are predicted inside the evolving shape. What is even more impressive: the model demonstrates that the system has self-correcting properties and accounts for many experimental observations that previously could not be explained.

[8] D.W. Selinger, M.A. Wright, and G.M. Church, "On the Complete Determination of Biological Systems," *Trends in Biotechnology* 21(6):251-254, 2003.

[9] A.F.M. Marée and P. Hogeweg, "How Amoeboids Self-organize into a Fruiting Body: Multicellular Coordination in *Dictyostelium discoideum*," *Proceedings of the National Academy of Sciences* 98(7):3879-3883, 2001.

5.2.10 Models Can Link What Is Known to What Is Yet Unknown

In the words of Pollard, "Any cellular process involving more than a few types of molecules is too complicated to understand without a mathematical model to expose assumptions and to frame the reactions in a rigorous setting."[10] Reviewing the state of the field in cell motility and the cytoskeleton, he observes that even with many details of the mechanism as yet controversial or unknown, modeling plays an important role. Referring to a system (of actin and its interacting proteins) modeled by Mogilner and Edelstein-Keshet,[11] he points to advantages gained by the mathematical framework: "A mathematical model incorporating molecular reactions and physical forces correctly predicts the steady-state rate of cellular locomotion." The model, he notes, correctly identifies what limits the motion of the cell, predicts what manipulations would change the rate of motion, and thus suggests experiments to perform. While details of some steps are still emerging, the model also distinguishes quantitatively between distinct hypotheses for how actin filaments are broken down for purposes of recycling their components.

5.2.11 Models Can Be Used to Generate Accurate Quantitative Predictions

Where detailed quantitative information exists about components of a system, about underlying rules or interactions, and about how these components are assembled into the system as a whole, modeling may be valuable as an accurate and rigorous tool for generating quantitative predictions. Weather prediction is one example of a complex model used on a daily basis to predict the future. On the other hand, the notorious difficulties of making accurate weather predictions point to the need for caution in adopting the conclusions even of classical models, especially for more than short-term predictions, as one might expect from mathematically chaotic systems.

5.2.12 Models Expand the Range of Questions That Can Meaningfully Be Asked[12]

For much of life science research, questions of purpose arise about biological phenomena. For instance, the question, Why does the eye have a lens? most often calls for the purpose of the lens—to focus light rays—and only rarely for a description of the biological mechanism that creates the lens. That such an answer is meaningful is the result of evolutionary processes that shape biological entities by enhancing their ability to carry out fitness-enhancing functions. (Put differently, biological entities are the result of nature's engineering of devices to perform the function of survival; this perspective is explored further in Chapter 6.)

Lander points out that molecular biologists traditionally have shied away from teleological matters, and that geneticists generally define function not in terms of the useful things a gene does, but by what happens when the gene is altered. However, as the complexity of biological mechanism is increasingly revealed, the identification of a purpose or a function of that mechanism has enormous explanatory power. That is, what purpose does all this complexity serve?

As the examples in Section 5.4 illustrate, computational modeling is an approach to exploring the implications of the complex interactions that are known from empirical and experimental work. Lander notes that one general approach to modeling is to create models in which networks are specified in terms of elements and interactions (the network "topology"), but the numerical values that quantify those interactions (the parameters) are deliberately varied over wide ranges to explore the functionality of the network—whether it acts as a "switch," "filter," "oscillator," "dynamic range adjuster," "producer of stripes," and so on.

[10] T.D. Pollard, "The Cytoskeleton, Cellular Motility and the Reductionist Agenda," *Nature* 422(6933):741-745, 2003.

[11] A. Mogilner and L. Edelstein-Keshet, "Regulation of Actin Dynamics in Rapidly Moving Cells: A Quantitative Analysis," *Biophysical Journal* 83(3):1237-1258, 2002.

[12] Section 5.2.12 is based largely on A.D. Lander, "A Calculus of Purpose," *PLoS Biology* 2(6):e164, 2004.

Lander explains the intellectual paradigm for determining function as follows:

> By investigating how such behaviors change for different parameter sets—an exercise referred to as "exploring the parameter space"—one starts to assemble a comprehensive picture of all the kinds of behaviors a network can produce. If one such behavior seems useful (to the organism), it becomes a candidate for explaining why the network itself was selected; i.e., it is seen as a potential purpose for the network. If experiments subsequently support assignments of actual parameter values to the range of parameter space that produces such behavior, then the potential purpose becomes a likely one.

5.3 TYPES OF MODELS[13]

5.3.1 From Qualitative Model to Computational Simulation

Biology makes use of many different types of models. In some cases, biological models are qualitative or semiquantitative. For example, graphical models show directional connections between components, with the directionality indicating influence. Such models generally summarize a great deal of known information about a pathway and facilitate the formation of hypotheses about network function. Moreover, the use of graphical models allows researchers to circumvent data deficiencies that might be encountered in the development of more quantitative (and thus data-intensive) models. (It has also been argued that probabilistic graphical models provide a coherent, statistically sound framework that can be applied to many problems, and that certain models used by biologists, such as hidden Markov models or Bayesian Networks), can be regarded as special cases of graphical models.[14])

On the other hand, the forms and structures of graphical models are generally inadequate to express much detail, which might well be necessary for mechanistic models. In general, qualitative models do not account for mechanisms, but they can sometimes be developed or analyzed in an automated manner. Some attempts have been made to develop formal schemes for annotating graphical models (Box 5.2).[15]

Qualitative models can be logical or statistical as well. For example, statistical properties of a graph of protein-protein interaction have been used to infer the stability of a network's function against most "deletions" in the graph.[16] Logical models can be used when data regarding mechanism are unavailable and have been developed as Boolean, fuzzy logical, or rule-based systems that model complex networks[17] or genetic and developmental systems.

In some cases, greater availability of data (specifically, perturbation response or time-series data) enables the use of statistical influence models. Linear,[18] neural network-like,[19] and Bayesian[20] models have all been used to deduce both the topology of gene expression networks and their dynamics. On the

[13] Section 5.3 is adapted from A.P. Arkin, "Synthetic Cell Biology," *Current Opinion in Biotechnology* 12(6):638-644, 2001.

[14] See, for example, Y. Moreau, P. Antal, G. Fannes, and B. De Moor, "Probabilistic Graphical Models for Computational Biomedicine, *Methods of Information in Medicine* 42(2):161-168, 2003.

[15] K.W. Kohn, "Molecular Interaction Map of the Mammalian Cell Cycle: Control and DNA Repair Systems," *Molecular Biology of the Cell* 10(8):2703-2734, 1999; I. Pirson, N. Fortemaison, C. Jacobs, S. Dremier, J.E. Dumont, and C. Maenhaut, "The Visual Display of Regulatory Information and Networks," *Trends in Cell Biology* 10(10):404-408, 2000. (Both cited in Arkin, 2001.)

[16] H. Jeong, S.P. Mason, A.L. Barabasi, and Z.N. Oltvai, "Lethality and Centrality in Protein Networks," *Nature* 411(6833):41-42, 2001; H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi, "The Largescale Organization of Metabolic Networks," *Nature* 407(6804):651-654, 2000. (Cited in Arkin, 2001.)

[17] D. Thieffry and R. Thomas, "Qualitative Analysis of Gene Networks," pp. 77-88 in *Pacific Symposium on Biocomputing*, 1998. (Cited in Arkin, 2001.)

[18] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear Modeling of mRNA Expression Levels During CNS Development and Injury," pp. 41-52 in *Pacific Symposium on Biocomputing*, 1999. (Cited in Arkin, 2001.)

[19] E. Mjolsness, D.H. Sharp, and J. Reinitz, "A Connectionist Model of Development," *Journal of Theoretical Biology* 152(4):429-453, 1999. (Cited in Arkin, 2001.)

[20] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian Networks to Analyze Expression Data," *Journal of Computational Biology* 7(3-4):601-620, 2000. (Cited in Arkin, 2001.)

Box 5.2
On Graphical Models

A large fraction of today's knowledge of biochemical or genetic regulatory networks is represented either as text or as cartoon-like diagrams. However, text has the disadvantage of being inherently ambiguous, and every reader must reinterpret the text of a journal article. Diagrams are usually informal, often confusing, and thus fail to present all of the information that is available to the presenter of the research. For example, the meanings of nodes and arcs within a diagram are inconsistent—one arrow may mean activation, but another arrow in the same diagram may mean transition of the state or translocation of materials.

To remedy this state of affairs, a system of graphical representation should be powerful enough to express sufficient information in a clearly visible and unambiguous way and should be supported by software tools. There are several criteria for a graphical notation system, including the following:

1. *Expressiveness.* The notation system should be able to describe every possible relationship among the entities in a system—for example, those between genes and proteins in a biological model.
2. *Semantical unambiguity*. Notation should be unambiguous. Different semantics should be assigned to different symbols that are clearly distinguishable.
3. *Visual unambiguity*. Each symbol should be identified clearly and not be mistaken with other symbols. This feature should be maintained with low-resolution displays, using only black and white.
4. *Extension capability*. The notation system should be flexible enough to add new symbols and relationships in a consistent manner. This may include the use of color coding to enhance expressiveness and readability, but information should not be lost even with black-and-white displays.
5. *Mathematical translation*. The notation should be able to convert itself into mathematical formalisms, such as differential equations, so that it can be applied directly for numerical analysis.
6. *Software support.* The notation should be supported by software for its drawing, viewing, editing, and translation into mathematical formalisms.

No current graphical notation system satisfies all of these criteria fully, although a number of systems satisfy some of them.[1]

SOURCE: Adapted by permission from H. Kitano, "A Graphical Notation for Biochemical Networks," *Biosilico* 1(5):159-176. Copyright 2003 Elsevier.

[1] See, for example, K.W. Kohn, "Molecular Interaction Map of the Mammalian Cell Cycle Control and DNA Repair Systems," *Molecular Biology of the Cell* 10(8):2703-2734, 1999; K. Kohn, "Molecular Interaction Maps as Information Organizers and Simulation Guides," *Chaos* 11(1):84-97, 2001.

other hand, statistical influence models are not causal and may not lead to a better understanding of underlying mechanisms.

Quantitative models make detailed statements about biological processes and hence are easier to falsify than more qualitative models. These models are intended to be predictive and are useful for understanding points of control in cellular networks and for designing new functions within them.

Some models are based on power law formalisms.[21] In such cases, the data are shown to fit generic power laws, and the general theory of power law scaling (for example) is used to infer some degree of causal structure. They do not provide detailed insight into mechanism, although power law models form the basis for a large class of metabolic control analyses and dynamic simulations.

Computational models—simulations—represent the other end of the modeling spectrum. Simulation is often necessary to explore the implications of a model, especially its dynamical behavior, because

[21] E.O. Voit and T. Radivoyevitch, "Biochemical Systems Analysis of Genomewide Expression Data," *Bioinformatics* 16(11):1023-1037, 2000. (Cited in Arkin, 2001.)

human intuition about complex nonlinear systems is often inadequate.[22] Lander cites two examples. The first is that "intuitive thinking about MAP [mitogen-activated protein] kinase pathways led to the long-held view that the obligatory cascade of three sequential kinases serves to provide signal amplification. In contrast, computational studies have suggested that the purpose of such a network is to achieve extreme positive cooperativity, so that the pathway behaves in a switch-like, rather than a graded, fashion."[23] The second example is that while intuitive interpretations of experiments in the study of morphogen gradient formation in animal development led to the conclusion that simple diffusion is not adequate to transport most morphogens, computational analysis of the same experimental data led the opposite conclusion.[24]

Simulation, which traces functional biological processes through some period of time, generates results that can be checked for consistency with existing data ("retrodiction" of data) and can also predict new phenomena not explicitly represented in but nevertheless consistent with existing datasets. Note also that when a simulation seeks to capture essential elements in some oversimplified and idealized fashion, it is unrealistic to expect the simulation to make detailed predictions about specific biological phenomena. Such simulations may instead serve to make qualitative predictions about tendencies and trends that become apparent only when averaged over a large number of simulation runs. Alternatively, they may demonstrate that certain biological behaviors or responses are robust and do not depend on particular details of the parameters involved within a very wide range.

Simulations can also be regarded as a nontraditional form of scientific communication. Traditionally, scientific communications have been carried by journal articles or conference presentations. Though articles and presentations will continue to be important, simulations—in the form of computer programs—can be easily shared among members of the research community, and the explicit knowledge embedded in them can become powerful points of departure for the work of other researchers.

With the availability of cheap and powerful computers, modeling and simulation have become nearly synonymous. Yet, a number of subtle differences should be mentioned. Simulation can be used as a tool on its own or as a companion to mathematical analysis.

In the case of relatively simple models meant to provide insight or reveal a concept, analytical and mathematical methods are of primary utility. With simple strokes and pen-and-paper computations, the dependence of behavior on underlying parameters (such as rate constants), conditions for specific dynamical behavior, and approximate connections between macroscopic quantities (e.g., the velocity of a cell) and underlying microscopic quantities (such the number of actin filaments causing the membrane to protrude) can be revealed. Simulations are not as easily harnessed to making such connections.

Simulations can be used hand-in-hand with analysis for simple models: exploring slight changes in equations, assumptions, or rates and gaining familiarity can guide the best directions to explore with simple models as well. For example, G. Bard Ermentrout at the University of Pittsburgh developed XPP software as an evolving and publicly available experimental modeling tool for mathematical biologists.[25] XPP has been the foundation of computational investigations in many challenging problems in neurophysiology, coupled oscillators, and other realms.

Mathematical analysis of models, at any level of complexity, is often restricted to special cases that have simple properties: rectangular boundaries, specific symmetries, or behavior in a special class. Simulations can expand the repertoire and allow the modeler to understand how analysis of the special cases

[22] A.D. Lander, "A Calculus of Purpose," *PLoS Biology* 2 (6):e164, 2004.

[23] C.Y. Huang and J.E. Ferrell, "Ultrasensitivity in the Mitogen Activated Protein Kinase Cascade," *Proceedings of the National Academy of Sciences* 93(19):10078-10083, 1996. (Cited in Lander, "A Calculus of Purpose," 2004.)

[24] A.D. Lander, Q. Nie, and F.Y. Wan, "Do Morphogen Gradients Arise by Diffusion?" *Developmental Cell* 2(6):785-796, 2002. (Cited in Lander, 2004.)

[25] See http://www.math.pitt.edu/~bard/xpp/xpp.html.

relates to more realistic situations. In this case, simulation takes over where analysis ends.[26] Some systems are simply too large or elaborate to be understood using analytical techniques. In this case, simulation is a primary tool. Forecasts requiring heavy "number-crunching" (e.g., weather prediction, prediction of climate change), as well as those involving huge systems of diverse interacting components (e.g., cellular networks of signal transduction cascades), are only amenable to exploration using simulation methods.

More detailed models require a detailed consideration of chemical or physical mechanisms involved (i.e., these models are mechanistic[27]). Such models require extensive details of known biology and have the largest data requirements. They are, in principle, the most predictive. In the extreme, one can imagine a simulation of a complete cell—an "in silico" cell or cybercell—that provides an experimental framework in which to investigate many possible interventions. Getting the right format, and ensuring that the in silico cell is a reasonable representation of reality, has been and continues to be an enormous challenge.

No reasonable model is based entirely on a bottom-up analysis. Consider, for example, that solving Schrödinger's equation for the millions of atoms in a complex molecule in solution would be a futile exercise, even if future supercomputers could handle this task. The question to ask is how and why such work would be contemplated: finding the correct level of representation is one of the key steps to good scientific work. Thus, some level of abstraction is necessary to render any model both interesting scientifically and feasible computationally. Done properly, abstractions can clarify the sources of control in a network and indicate where more data are necessary. At the same time, it may be necessary to construct models at higher degrees of biophysical realism and detail in any event, either because abstracted models often do not capture the essential behavior of interest or to show that indeed the addition of detail does not affect the conclusions drawn from the abstracted model.[28]

It is also helpful to note the difference between a computational artifact that reproduces some biological behavior (a task) and a simulation. In the former case, the relevant question is: "How well does the artifact accomplish the task?" In the latter case, the relevant question is: "How closely does the simulation match the essential features of the system in question?"

Most computer scientists would tend to assign higher priority to performance than to simulation. The computer scientist would be most interested in a biologically inspired approach to a computer science problem when some biological behavior is useful in a computational or computer systems context and when the biologically inspired artifact can demonstrate better performance than is possible through some other way of developing or inspiring the artifact. A model of a biological system then becomes useful to the computer scientist only to the extent that high-fidelity mimicking of how nature accomplishes a task will result in better performance of that task.

By contrast, biologists would put greater emphasis on simulation. Empirically tested and validated simulations with predictive capabilities would increase their confidence that they understood in some fundamental sense the biological phenomenon in question. However, it is important to note that because a simulation is judged on the basis of how closely it represents the *essential* features of a biological system, the question "What counts as essential?" is central (Box 5.3). More generally, one fundamental focus of biological research is a determination of what the "essential" features of a biological system are,

[26] At times, it is also desirable to employ a mix of analysis and simulation. Analysis would be used to generate the basic equations underlying a complex phenomenon. Solutions to these equations would then be explored and with luck, considerably simplified. The simplified models can then be simulated. See, for example, E.A. Ezrachi, R. Levi, J.M. Camhi, and H. Parnas, "Right-Left Discrimination in a Biologically Oriented Model of the Cockroach Escape System," *Biological Cybernetics* 81(2):89-99, 1999.

[27] Note that mechanistic models can be stochastic—the term "mechanistic" should not be taken to mean deterministic.

[28] Tensions between these perspectives were apparent even in reviews of the draft of this report. In commenting on neuroscience topics in this report, advocates of the first point of view argued that ultrarealistic simulations accomplish little to further our understanding about how neurons work. Advocates of the second point of view argued that simple neural models could not capture the implications of the complex dynamics of each neuron and its synapses and that these models would have to be supplemented by more physiological ideas. From the committee's perspective, both points of view have merit, and the scientific challenge is to find an appropriate simplification or abstraction that does capture the interesting behavior at reasonable fidelity.

Box 5.3
An Illustration of "Essential"

Consider the following modeling task. The phenomenon of interest is a monkey learning to fetch a banana from behind a transparent conductive screen. The first time, the monkey sees the banana, goes straight ahead, bumps into the screen, and then goes around the screen to the banana. The second time, the monkey, having discovered the existence of the screen that blocks his way, goes directly around the screen to the banana.

To model this phenomenon, a system is constructed, consisting of a charged ball and a metal sheet. The charged metal ball is hung from a string above the banana and then held at an angle so the screen separates the ball and the banana. The first time the ball is released, the ball swings toward the screen, and then touches it, transferring part of its charge to the screen. The similar charges on the screen and the ball now repel each other, and the ball swings around the screen. The second time the ball is released, the ball sees a similarly charged screen and goes around the screen directly.

This model reproduces the behavior of the monkey in the first instance. However, no one would claim that it is an accurate model of the learning that takes place in the monkey's brain, even though the model replicates the most salient feature of the monkey's learning consistently: both the ball and the monkey dodge the screen on the second attempt. In other words, even though it demonstrates the same behavior, the model does not represent the essential features of the biological system in question.

recognizing that what is "essential" cannot be determined once and for all, but rather depends on the class of questions under consideration.

5.3.2 Hybrid Models

Hybrid models are models composed of objects with different mathematical representations. These allow a model builder the flexibility to mix modeling paradigms to describe different portions of a complex system. For example, in a hybrid model, a signal transduction pathway might be described by a set of differential equations, and this pathway could be linked to a graphical model of the genetic regulatory network that it influences. An advantage of hybrid models is that model components can evolve from high-level abstract descriptions to low-level detailed descriptions as the components are better characterized and understood.

An example of hybrid model use is offered by McAdams and Shapiro,[29] who point out that genetic networks involving large numbers of genes (more than tens) are difficult to analyze. Noting the "many parallels in the function of these biochemically based genetic circuits and electrical circuits," they propose "a hybrid modeling approach that integrates conventional biochemical kinetic modeling within the framework of a circuit simulation. The circuit diagram of the bacteriophage lambda lysislysogeny decision circuit represents connectivity in signal paths of the biochemical components. A key feature of the lambda genetic circuit is that operons function as active integrated logic components and introduce signal time delays essential for the in vivo behavior of phage lambda."

There are good numerical methods for simulating systems that are formulated in terms of ordinary differential equations or algebraic equations, although good methods for analysis of such models are still lacking. Other systems, such as those that mix continuous with discrete time or Markov processes with partial differential equations, are sometimes hard to solve even by numerical methods. Further, a particular model object may change mathematical representation during the course of the analysis. For example, at the beginning of a biosynthetic process there may be very small amounts of product so its

[29] See H.H. McAdams and L. Shapiro, "Circuit Simulation of Genetic Networks," *Science* 269(5224):650-656, 1994.

concentration would have to be modeled discretely. As more of it is synthesized, the concentration becomes high enough that a continuous approximation is justified and is then more efficient for simulation and analysis.

The point at which this switch is made is dependent not just on copy number but also on where in the dynamical state space the system resides. If the system is near a bifurcation point, small fluctuations may be significant. Theories of how to accomplish this dynamic switching are lacking. As models grow more complex, different parts of the system will have to be modeled with different mathematical representations. Also, as models from different sources begin to be joined, it is clear that different representations will be used. It is critical that the theory and applied mathematics of hybrid dynamical systems be developed.

5.3.3 Multiscale Models

Multiscale models describe processes occurring at many time and length scales. Depending on the biological system of interest, the data needed to provide the basis for a greater understanding of the system will cut across several scales of space and time. The length dimensions of biological interest range from small organic molecules to multiprotein complexes at 100 angstroms to cellular processes at 1,000 angstroms to tissues at 1-10 microns, and the interaction of human populations with the environment at the kilometer scale. The temporal domain includes the femtosecond chemistry of molecular interactions to the millions of years of evolutionary time, with protein folding in seconds and cell and developmental processes in minutes, hours, and days. In turn, the scale of the process involved (e.g., from the molecular scale to the ecosystem scale) affects both the complexity of the representation (e.g., molecule base, concentration based, at equilibrium or fully dynamic) and the modality of the representation (e.g., biochemical, genetic, genomic, electrophysiological, etc.).

Consider the heart as an example. The macroscopic unit of interest is the heartbeat, which lasts about a second and involves the whole heart of 10 cm scale. But the cardiac action potential (the electrical signal that initiates myocellular contractions) can change significantly on time scales of milliseconds as reflected in the appropriate kinetic equations. In turn, the molecular interactions that underlie kinetic flows occur on time scales on the order of femtoseconds. Across such variation in time scales, it is not feasible to model 10^{15} molecular interactions in order to model a complete heartbeat. Fortunately, in many situations the response with the shorter time scale will converge quickly to equilibrium or quasi-steady-state behavior, obviating the need for a complete lower-level simulation.[30]

For most biological problems, the scale at which data could provide a central insight into the operation of the whole system is not known, so multiple scales are of interest. Thus, biological models have to allow for transition among different levels of resolution. A biologist might describe a protein as a simple ellipsoid and then in the next breath explain the effect of a point mutation by the atomic-level structural changes it causes in the active site.[31]

Identifying the appropriate ranges of parameters (e.g., rate constants that govern the pace of chemical reactions) remains one of the difficulties that every modeler faces sooner or later. As modelers know well, even qualitative analysis of simple models depends on knowing which “leading-order terms” are to be kept on which time scales. When the relative rates are entirely unknown—true of many biochemical steps in living cells—it is hard to know where to start and how to assemble a relevant model, a point that underscores the importance of close dialogue between the laboratory biologist and the mathematical or computational modeler.

Finally, data obtained at a particular scale must be sufficient to summarize the essential biological activity at that scale in order to be evaluated in the context of interactions at greater scales of complexity. The challenge, therefore, is one of understanding not only the relationship of multiple variables operating at one scale of detail, but also the relationship of multivariable datasets collected at different scales.

[30] A.D. McCulloch and G. Huber, “Integrative Biological Modelling in Silico,” pp. 4-25 in *‘In Silico’ Simulation of Biological Processes No. 247,* Novartis Foundation Symposium, G. Bock and J.A. Goode, eds., John Wiley & Sons Ltd., Chichester, UK, 2002.

[31] D. Endy and R. Brent, “Modeling Cellular Behavior,” *Nature* 409(6818):391-395, 2001.

5.3.4 Model Comparison and Evaluation

Models are ultimately judged by their ability to make predictions. Qualitative models predict trends or types of dynamics that can occur, as well as thresholds and bifurcations that delineate one type of behavior from another. Quantitative models predict values that can be compared to actual experimental data. Therefore, the selection of experiments to be performed can be determined, at least in part, by their usefulness in constraining a model or selecting one model from a set of competing models.

The first step in model evaluation is to replicate and test a computational model of biological systems that has been published. However, most papers contain typographical errors and do not provide a complete specification of the biological properties that were represented in the model. One should be able to extract the specification from the model's source code, but for a whole host of reasons it is not always possible to obtain the actual files that were used for the published work.

In the neuroscience field, ModelDB (http://senselab.med.yale.edu/senselab/modeldb/) is being developed to answer the need for a database of published models used in neuroscience research.[32] It is part of the SenseLab project (http://senselab.med.yale.edu/), which is supported through the Human Brain Project by the National Institute of Mental Health (NIMH), the National Institute of Neurologist disorders and Stroke (NINDS), and the National Cancer Institute (NCI).

ModelDB is a curated database that is designed for convenient entry, search, and retrieval of models written for any programming language or simulation environment. As of December 10, 2004, it contained 141 downloadable models. Most of these are for NEURON, but 40 of them are for MATLAB, GENESIS, SNNAP, or XPP, and there are also some models in C/C++ and FORTRAN. Database entries are linked to the published literature so that users can more easily determine the "scientific context" of any given model.

Although ModelDB is still in a developmental or research stage, it has already begun to have a positive effect on computational modeling in neuroscience. Database logs indicate that it is seeing heavy usage, and from personal communications the committee has learned that even experienced programmers who write their own code in C/C++ are regularly examining models written for NEURON and other domain-specific simulators, in order to determine key parameter values and other important details. Recently published papers are beginning appear that cite ModelDB and the models it contains as sources of code, equations, or parameters. Furthermore, a leading journal has adopted a policy that requires authors to make their source code available as a condition of publication and encourages them to use ModelDB for this purpose.

As for model comparison, it is not possible to ascertain in isolation whether a given model is correct since contradictory data may become available later, and indeed even "incorrect" models may make correct predictions. Suitably complex models can be made to fit to any dataset, and one must guard against "overfitting" a model. Thus, the predictions of a model must be viewed in the context of the number of degrees of freedom of the model, and one measure that one model is better than another is a judgment about which model best explains experimental data with the least model complexity. In some cases, measures of the statistical significance of a model can be computed using a likelihood distribution over predicated state variables taking into account the number of degrees of freedom present in the model.

At the same time, lessons learned over many centuries of scientific investigation regarding the use of Occam's Razor may have limited applicability in this context. Because biological phenomena are the result of an evolutionary process that simply uses what is available, many biological phenomena are simply cobbled together and in no sense can be regarded as the "simplest" way to accomplish something.

As noted in Footnote 28, there is a tension between the need to capture details faithfully in a model and the desire to simplify those details so as to arrive at a representation that can be analyzed, understood fully, and converted into scientific "knowledge." There are numerous ways of reducing models that are well known in applied mathematics communities. These include dimensional analysis and multiple timescale analysis (i.e., dissecting a system into parts that evolve rapidly versus those that change on a slower

[32] M.L. Hines, T. Morse, M. Migliore, N.T. Carnevale, and G.M. Shepherd, "ModelDB: A Database to Support Computational Neuroscience," *Journal of Computational Neuroscience* 17(1):7-11, 2004; B.J. Richmond, "Editorial Commentary," *Journal of Computational Neuroscience* 17(1):5, 2004.

time scale). In some cases, leaving out some of the interacting components (e.g., those whose interactions are weakest or least significant) may be a workable method. In other cases, lumping together families or groups of substances to form aggregate components or compartments works best. Sensitivity analysis of alternative model structures and parameters can be performed using likelihood and significance measures. Sensitivity analysis is important to inform a model builder of the essential components of the model and to attempt to reduce model complexity without loss of explanatory power.

Model evaluation can be complicated by the robustness of the biological organism being represented. Robustness generally means that the organism will endure and even prosper under a wide range of conditions—which means that its behavior and responses are relatively insensitive to variations in detail.[33] That is, such differences are unlikely to matter much for survival. (For example, the modeling of genetic regulatory networks can be complicated by the fact that although the data may show that a certain gene is expressed under certain circumstances, the biological function being served may not depend on the expression of that gene.) On the other hand, this robustness may also mean that a flawed understanding of detailed processes incorporated into a model that does explain survival responses and behavior will not be reflected in the model's output.[34]

Simulation models are essentially computer programs and hence suffer from all of the problems that plague software development. Normal practice in software development calls for extensive testing to see that a program returns the correct results when given test data for which the appropriate results are known independently of the program as well as for independent code reviews. In principle, simulation models of biological systems could be subject to such practices. Yet the fact that a given simulation model returns results that are at variance with experimental data may be attributable to an inadequacy of the underlying model or to an error in programming.[35] Note also that public code reviews are impossible if the simulation models are proprietary, as they often are when they are created by firms seeking to obtain competitive advantage in the marketplace.

These points suggest a number of key questions in the development of a model.

- How much is given up by looking at simplified versions?
- How much poorer, and in what ways poorer, is a simplified model in its ability to describe the system?
- Are there other, new ways of simplifying and extracting salient features?
- Once the simplified representation is understood, how can the details originally left out be reincorporated into a model of higher fidelity?

Finally, another approach to model evaluation is based on notions of logical consistency. This approach uses program verification tools originally developed by computer scientists to determine whether a given program is consistent with a given formal specification or property. In the biological context, these tools are used to check the consistency and completeness of a model's description of the biological system's processes. These descriptions are dynamic and thus permit "running" a model to observe developments in time. Specifically, Kam et al. have demonstrated this approach using the languages, methods, and tools of scenario-based reactive system design and applied it to modeling the well-characterized process of cell fate acquisition during *Caenorhabditis elegans* vulval development. (Box 5.4 describes the intellectual approach in more detail.[36])

[33] L.A. Segel, "Computing an Organism," *Proceedings of the National Academy of Sciences* 98(7):3639-3640, 2001.

[34] On the basis of other work, Segel argues that a biological model enjoys robustness only if it is "correct" in certain essential features.

[35] Note also the well-known psychological phenomenon in programming—being a captive of one's test data. Programming errors that prevent the model from accounting for the data tend to be hunted down and fixed. However, if the model does account for the data, there is a tendency to assume that the program is correct.

[36] N. Kam, D. Harel, H. Kugler, R. Marelly, A. Penueli, J. Hubbard, et al., "Formal Modeling of *C. elegans* Development: A Scenario-based Approach," pp. 4-20 in *Proceedings of the First International Workshop on Computational Methods in Systems Biology* (CMSB03; Rovereto, Italy, February 2003), Vol. 2602, Lecture Notes in Computer Science, Springer-Verlag, Berlin, Heidelberg, 2003. This material is scheduled to appear in the following book: G. Ciobanu, ed., *Modeling in Molecular Biology,* Natural Computing Series, Springer, available at http://www.wisdom.weizmann.ac.il/~kam/CelegansModel/Publications/MMB_Celegans.pdf.

Box 5.4
Formal Modeling of *Caenorhabditis elegans* Development

Our understanding of biology has become sufficiently complex that it is increasingly difficult to integrate all the relevant facts using abstract reasoning alone. [Formal modeling presents] a novel approach to modeling biological phenomena. It utilizes in a direct and powerful way the mechanisms by which raw biological data are amassed, and smoothly captures that data within tools designed by computer scientists for the design and analysis of complex reactive systems.

A considerable quantity of biological data is collected and reported in a form that can be called "condition-result" data. The gathering is usually carried out by initializing an experiment that is triggered by a certain set of circumstances (conditions), following which an observation is made and the results recorded. The condition is most often a perturbation, such as mutating genes or exposing cells to an altered environment. . . . [and] a large proportion of biological data is reported as stories, or "scenarios," that document the results of experiments conducted under specific conditions.

The challenge of modeling these aspects of biology is to be able to translate such "condition-result" phenomena from the "scenario"-based natural language format into a meaningful and rigorous mathematical language. Such a translation process will allow these data to be integrated more comprehensively by the application of high-level computer-assisted analysis. In order for it to be useful, the model must be rigorous and formal, and thus amenable to verification and testing.

We have found that modeling methodologies originating in computer science and software engineering, and created for the purpose of designing complex *reactive systems*, are conceptually well suited to model this type of condition-result biological data. Reactive systems are those whose complexity stems not necessarily from complicated computation but from complicated reactivity over time. They are most often highly concurrent and time-intensive, and exhibit hybrid behavior that is predominantly discrete in nature but has continuous aspects as well. The structure of a reactive system consists of many interacting components, in which control of the behavior of the system is highly distributed amongst the components. Very often the structure itself is dynamic, with its components being repeatedly created and destroyed during the system's life span.

The most widely used frameworks for developing models of such systems feature *visual formalisms*, which are both graphically intuitive and mathematically rigorous. These are supported by powerful tools that enable full model executability and analysis, and are linkable to graphical user interfaces (GUIs) of the system. This enables realistic simulation prior to actual implementation. At present, such languages and tools—often based on the *object-oriented* paradigm—are being strengthened by verification modules, making it possible not only to execute and simulate the system models (test and observe) but also to verify dynamic properties thereof (prove). . . .

[M]any kinds of biological systems exhibit characteristics that are remarkably similar to those of reactive systems. The similarities apply to many different levels of biological analysis, including those dealing with molecular, cellular, organ-based, whole organism, or even population biology phenomena. Once viewed in this light, the dramatic concurrency of events, the chain-reactions, the time-dependent patterns, and the event-driven discrete nature of their behaviors, are readily apparent. Consequently, we believe that biological systems can be productively modeled as reactive systems, using languages and tools developed for the construction of man-made systems. . . .

SOURCE: N. Kam et al., "Formal Modeling of *C. elegans* Development: A Scenario-based Approach," pp. 4-20 in *Proceedings of the First International Workshop on Computational Methods in Systems Biology* (CMSB03; Rovereto, Italy, February 2003), Vol. 2602, Lecture Notes in Computer Science, Springer-Verlag, Berlin, Heidelberg, 2003, available at http://www.wisdom.weizmann.ac.il/~kam/CelegansModel/Publications/MMB_Celegans.pdf. Reprinted with permission from Springer-Verlag.

5.4 MODELING AND SIMULATION IN ACTION

The preceding discussion has been highly abstract. This section provides some illustrations of how modeling and simulation have value across a variety of subfields in biology. No claim is made to comprehensiveness, but the committee wishes to illustrate the utility of modeling and simulations at levels of organization from gene to ecosystem.

5.4.1 Molecular and Structural Biology

5.4.1.1 Predicting Complex Protein Structures

Interactions between proteins are crucial to the functioning of all cells. While there is much experimental information being gathered regarding protein structures, many interactions are not fully understood and have to be modeled computationally. The topic of computational prediction of protein-protein structure remains to be solved and is one of the most active areas of research in bioinformatics and structural biology.

ZDOCK and RDOCK are two computer programs that address this problem, also known as protein docking.[37] ZDOCK is an initial stage protein docking program that performs a full search of the relative orientations of two molecules (referred to by convention as the ligand and receptor) to determine their best fit based on surface complementarity, electrostatics and desolvation. The efficiency of the algorithm is enhanced by discretizing the molecules onto a grid and performing a fast Fourier transform (FFT) to quickly explore the translational degrees of freedom.

RDOCK takes as input the ZDOCK predictions and improves them using two steps. The first step is to improve the energetics of the prediction and remove clashes by performing small movements of the predicted complex, using a program known as CHARMM. The second step is to rescore these minimized predictions with more detailed scoring functions for electrostatics and desolvation.

The combination of these two algorithms has been tested and verified with a benchmark set of proteins collected for use in testing docking algorithms. Now at version 2.0, this benchmark is publicly available and contains 87 test cases. These test cases cover a breadth of interactions, such as antibody-antigen, and cases involving significant conformational changes.

The ZDOCK-RDOCK programs have consistently performed well in the international docking competition CAPRI (Figure 5.1). Some notable predictions were for the *Rotavirus* VP6/Fab (50 of 52 contacting residues correctly predicted), and SAG-1/Fab complex (61 of 70 contacts correct), and the cellulosome cohesion-dockerin structure (50 of 55 contacts correct). In the first two cases, the number of contacts in the ZDOCK-RDOCK predictions were the highest among all participating groups.

5.4.1.2 A Method to Discern a Functional Class of Proteins

The DNA-binding helix-turn-helix structural motif plays an essential role in a variety of cellular pathways that include transcription, DNA recombination and repair, and DNA replication. Current methods for identifying the motif rely on amino acid sequence, but since members of the motif belong to different sequence families that have no sequence homology to each other, these methods have been unable to identify all motif members.

A new method based on three-dimensional structure was created that involved the following steps:[38] (1) choosing a conserved component of the motif, (2) measuring structural features relative

[37] For more information, see http://zlab.bu.edu.

[38] W.A. McLaughlin and H.M. Berman, "Statistical Models for Discerning Protein Structures Containing the DNA-binding Helix-Turn-Helix Motif," *Journal of Molecular Biology* 330(1):43-55, 2003.

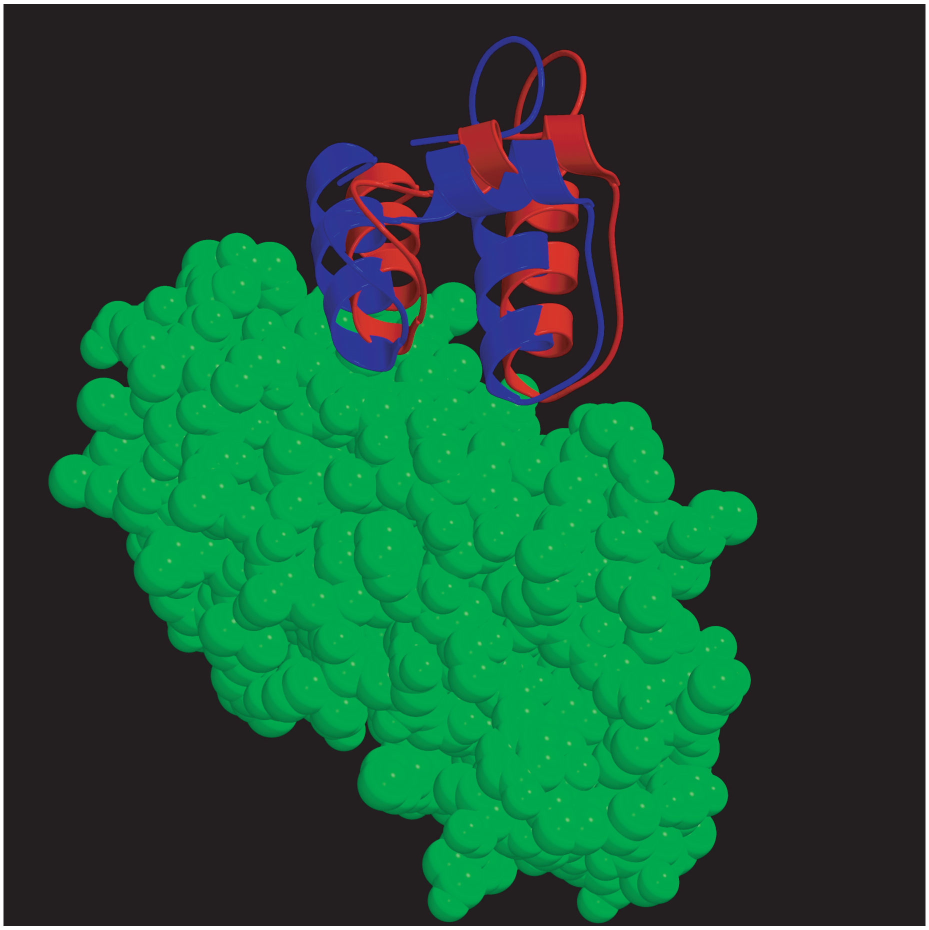


FIGURE 5.1 The ZDOCK/RDOCK prediction for dockerin (in red) superposed on the crystal structure for CAPRI Target 13, cohesin/dockerin. SOURCE: Courtesy of Brian Pierce and Zhiping Weng, Boston University.

to that component, and (3) creating classification models by comparing measurements of structures known to contain the motif to measurements of structures known not to contain the motif. In this case, the conserved component chosen was the recognition helix (i.e., the alpha helix that makes sequence-specific contact with DNA), and two types of relevant measurements were the hydrophobic area of interaction between secondary structure elements (SSEs) and the relative solvent accessibility of SSEs.

With a classification model created, the entire Protein Data Bank of experimentally measured structures was searched and new examples of the motif were found that have no detected sequence homology with previously known examples. Two such examples are Esa1 histone acetyltransferase and isoflavone 4-O-methyltransferase. The result emphasizes an important utility of the approach: sequence-based methods used to discern a functional class of proteins may be supplemented through the use of a classification model based on three-dimensional structural information.

5.4.1.3 Molecular Docking

Using a simple, uniform representation of molecular surfaces that requires minimal parameterization, Jain[39] has constructed functions that are effective for scoring protein-ligand interactions, quantitatively comparing small molecules, and making comparisons of proteins in a manner that does not depend on protein backbone. These methods rely on computational approaches that are rooted in understanding the physics of molecular interactions, but whose functional forms *do not* resemble those used in physics-based approaches. That is, this problem can be treated as a pure computer science problem that can be solved using combinations of scoring and search or optimization techniques parameterized with the use of domain knowledge. The approach is as follows:

- Molecules are approximated as collections of spheres with fixed radii: H = 1.2; C = 1.6; N = 1.5; O = 1.4; S = 1.95; P = 1.9; F = 1.35; Cl = 1.8; Br = 1.95; I = 2.15.
- A labeling of the features of polar atoms is superimposed on the molecular representation: polarity, charge, and directional preference (Figure 5.2, subfigures A and B).
- A scoring function is derived that, given a protein and a ligand in some relative alignment, yields a prediction of the energy of interaction.
- The function is parameterized in terms of the pairwise distances between molecular surfaces.
- The dominant terms are a hydrophobic term that characterizes interactions between nonpolar atoms and a polar term that captures complementary polar contacts with proper directionality.
- The parameters of the function were derived from empirical binding data and 34 protein-ligand complexes that were experimentally determined.
- The scoring function is described in Figure 5.2, Subfigure C. The hydrophobic term peaks at approximately 0.1 unit with a slight surface interpenetration. The hydrophobic term for an ideal hydrogen bond peaks at 1.25 units, and a charged interaction (tertiary amine proton (+1.0) to a charged carboxylate (–0.5)) peaks at about 2.3 units. Note that this scoring function looks nothing like a force field derived from molecular mechanics.
- Figure 5.2, Subfigure D compares eight docking methods on screening efficiency using thymidine kinase as a docking target. For the test, 10 known ligands and 990 random ligands were used. Particularly at low false-positive rates (low database coverage), the scoring function approach shows substantial improvements over the other methods.

5.4.1.4 Computational Analysis and Recognition of Functional and Structural Sites in Protein Structures[40]

Structural genomics initiatives are producing a great increase in protein three-dimensional structures determined by X-ray and nuclear magnetic resonance technologies as well as those predicted by computational methods. A critical next step is to study the relationships between protein structures and functions. Studying structures individually entails the danger of identifying idiosyncratic rather than conserved features and the risk of missing important relationships that would be revealed by statisti-

[39] See A.N. Jain, "Scoring Noncovalent Protein Ligand Interactions: A Continuous Differentiable Function Tuned to Compute Binding Affinities," *Journal of Computer-Aided Molecular Design* 10(5):427-440, 1996; W. Welch, J. Ruppert, and A.N. Jain, "Hammerhead: Fast, Fully Automated Docking of Flexible Ligands to Protein Binding Sites," *Chemistry & Biology* 3(6):449-462, 1996; J. Ruppert, W. Welch, and A.N. Jain, "Automatic Identification and Representation of Protein Binding Sites for Molecular Docking," *Protein Science* 6(3):524-533, 1997; A.N. Jain, "Surflex: Fully Automatic Flexible Molecular Docking Using a Molecular Similarity-based Search Engine," *Journal of Medicinal Chemistry* 46(4):499-511, 2003; A.N. Jain, "Ligand-Based Structural Hypotheses for Virtual Screening." *Journal of Medicinal Chemistry* 47(4):947-961, 2004.

[40] Section 5.4.1.4 is based on material provided by Liping Wei, Nexus Genomics, Inc., and Russ Altman, Stanford University, personal communication, December 4, 2003.

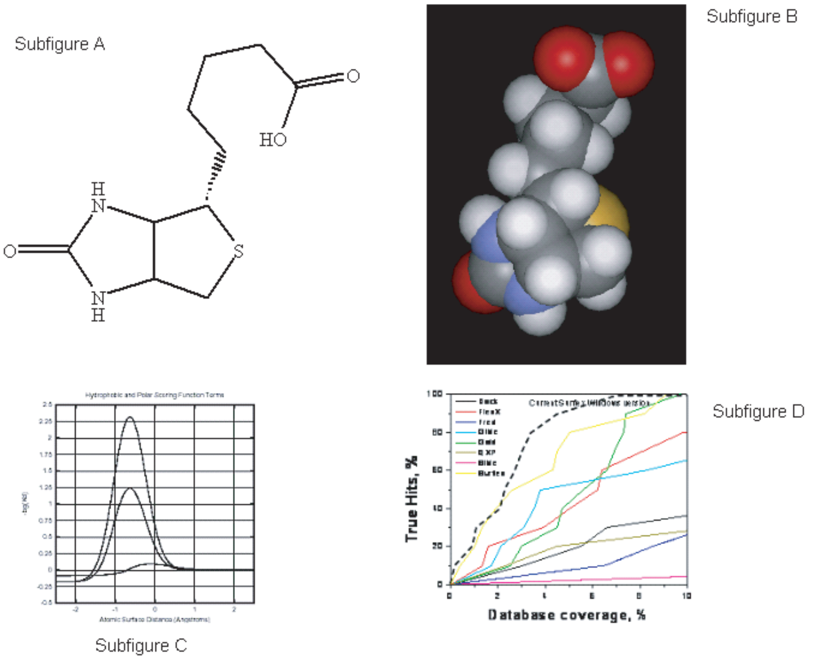


FIGURE 5.2 A Computational Approach to Molecular Docking. SOURCE: Courtesy of A.N. Jain, University of California, San Francisco.

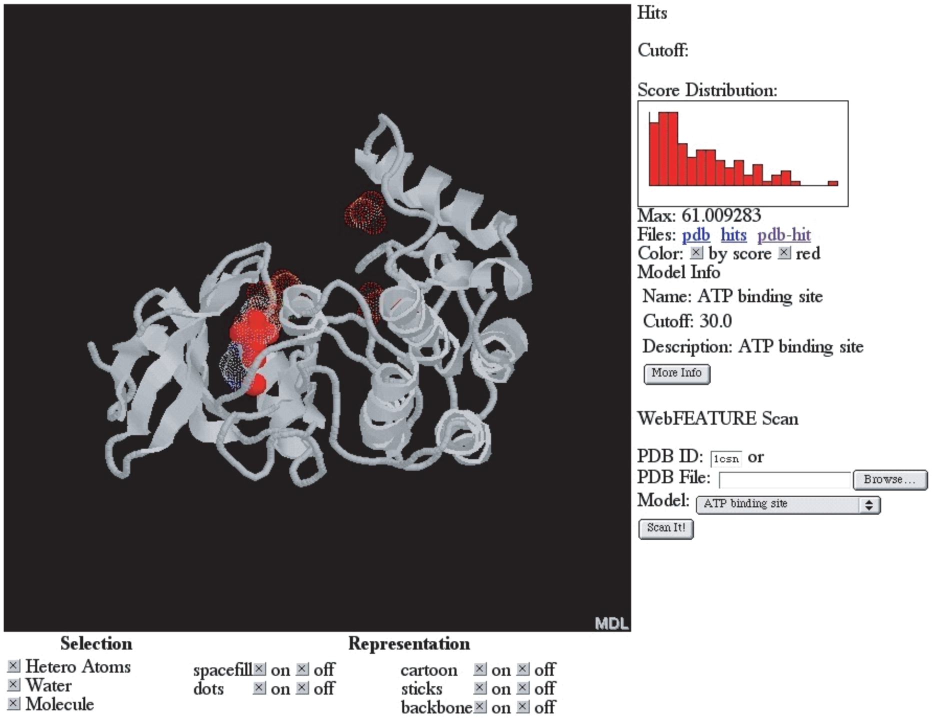
cally pooling relevant data. The expected surfeit of protein structures provides an opportunity to develop computational methods for collectively examining multiple biological structures and extracting key biophysical and biochemical features, as well as methods for automatically recognizing these features in new protein structures.

Wei and Altman have developed an automated system known as FEATURE that statistically studies the important functional and structural sites in protein structures such as active sites, binding sites, disulfide bonding sites, and so forth. FEATURE collects all known examples of a type of site from the Protein Data Bank (PDB) as well as a number of control "nonsite" examples. For each of them, FEATURE computes the spatial distributions of a large set of defined biophysical and biochemical properties spanning multiple levels of details in order to capture conserved features beyond basic amino acid sequence similarity. It then uses a nonparametric statistical test, the Wilcoxin Rank Sum Test, to find the features that are characteristic of the sites, in the context of control nonsites. Figure 5.3 shows the statistical features of calcium binding sites.

By using a Bayesian scoring function that recognizes whether a local region within a three-dimensional structure is likely to be any of the sites and a scanning procedure that searches the whole structure for the sites, FEATURE can also provide an initial annotation of new protein structures. FEATURE has been shown to have good sensitivity and specificity in recognizing a diverse set of site types, including active sites, binding sites, and structural sites and is especially useful when the sites do not have conserved residues or residue geometry. Figure 5.4 shows the result of searching for ATP (adenosine triphosphate) binding sites in a protein structure.

Calcium Model	VOLUME					
	0	1	2	3	4	5
ATOM-NAME-IS-ANY	<	>		>	>	
ATOM-NAME-IS-C	<	<	>	>	>	
ATOM-NAME-IS-N		<	<	>		
ATOM-NAME-IS-O	<	>	>	>	>	>
AMIDE		<	<	>		
AMINE			<			
CARBONYL	<	>		>	>	>
RING-SYSTEM			<			
PEPTIDE	<			>	>	>
VDW-VOLUME	<	<		>	>	
CHARGE		>	>	>		
NEG-CHARGE		>	>	>		
CHARGE-WITH-HIS		>	>	>		
HYDROPHOBICITY		<	<	<		<
MOBILITY	<	>	>			
SOLVENT-ACCESSIBILITY	<			>		
RESIDUE_NAME_IS_ASN		>	>	>	>	
RESIDUE_NAME_IS_ASP		>	>	>	>	>
RESIDUE_NAME_IS_GLU		>	>	>	>	>
RESIDUE_NAME_IS_GLY		>		>	>	
RESIDUE_NAME_IS_ILE				>		
RESIDUE_NAME_IS_LEU			>			
RESIDUE_NAME_IS_LYS			>			
RESIDUE_NAME_IS_SER					>	>
RESIDUE_NAME_IS_VAL			<		<	
RESIDUE_NAME_IS_HOH		>				
RESIDUE_CLASS1_IS_HYDROPHOBIC	<		<			
RESIDUE_CLASS1_IS_CHARGED		>	>	>	>	>
RESIDUE_CLASS1_IS_POLAR	<			>	>	>
RESIDUE_CLASS1_IS_UNKNOWN		>		>		
RESIDUE_CLASS2_IS_NONPOLAR	<		<			
RESIDUE_CLASS2_IS_POLAR	<			>	>	
RESIDUE_CLASS2_IS_BASIC			<			
RESIDUE_CLASS2_IS_ACIDIC		>	>	>	>	>
RESIDUE_CLASS2_IS_UNKNOWN		>		>		
SECONDARY_STRUCTURE1_IS_TURN		>				
SECONDARY_STRUCTURE1_IS_BEND		>	>	>	>	>
SECONDARY_STRUCTURE1_IS_COIL		>	>	>	>	>
SECONDARY_STRUCTURE1_IS_HET		>		>		
SECONDARY_STRUCTURE2_IS_BETA	<			>	>	
SECONDARY_STRUCTURE2_IS_COIL		>	>	>	>	>
SECONDARY_STRUCTURE2_IS_HET		>		>		

FIGURE 5.3 Statistical features of calcium binding sites determined by FEATURE. The volumes in this case correspond to concentrate radial shells 1 Å in thickness around the calcium ion or a control nonsite location. The column shows properties that are statistically significantly different (at *p*-value cutoff of 0.01) in at least one volume between known examples of calcium binding sites and those of control nonsites. A ">" (greater than sign) indicates that the calcium binding sites have significantly higher value for that property at that volume compared to control nonsites. A "<" (less than sign) indicates the opposite. An empty box indicates the lack of statistically significant difference. SOURCE: Courtesy of Liping Wei, Nexus Genomics, Inc., and Russ Altman, Stanford University, personal communication, December 4, 2003.



Hits

Cutoff:

Score Distribution:

Max: 61.009283
Files: pdb hits pdb-hit
Color: ☒ by score ☒ red
Model Info
Name: ATP binding site
Cutoff: 30.0
Description: ATP binding site

WebFEATURE Scan

PDB ID: 1csn or
PDB File:
Model: ATP binding site

Selection
☒ Hetero Atoms
☒ Water
☒ Molecule

Representation
spacefill ☒ on ☒ off
dots ☒ on ☒ off
cartoon ☒ on ☒ off
sticks ☒ on ☒ off
backbone ☒ on ☒ off

FIGURE 5.4 Results of automatic scanning for ATP binding sites in the structure of casein kinase (PDB ID 1csn) using WebFEATURE, a freely available, Web-based server of FEATURE. The solid red dots show the prediction of FEATURE, they correspond correctly with the true location of the ATP binding site, shown as white cloud. SOURCE: Courtesy of Liping Wei, Nexus Genomics, Inc., and Russ Altman, Stanford University, personal communication, December 4, 2003.

5.4.2 Cell Biology and Physiology

5.4.2.1 Cellular Modeling and Simulation Efforts

Cellular simulation requires a theoretical framework for analyzing the interactions of molecular components, of modules made up of those components, and of systems in which such modules are linked to carry out a variety of functions. The theoretical goal is to quantitatively organize, analyze, and interpret complex data on cell biological processes, and experiments provide images, biochemical and electrophysiological data on the initial concentrations, kinetic rates, and transport properties of the molecules and cellular structures that are presumed to be the key components of a cellular event.[41] A simulation embeds the relevant rate laws and rate constants for the biochemical transformations being modeled. Based on these laws and parameters, the model accepts as initial conditions the initial concentrations, diffusion coefficients, and locations of all molecules implicated in the transformation, and generates predictions for the concentration of all molecular species as a function of time and space. These predictions are compared against experiment, and the differences between prediction and experiment are used to further refine the model. If the system is perturbed by the addition of a ligand, electrical stimulus, or other experimental intervention, the model should be capable of predicting changes as well in the relevant spatiotemporal distributions of the molecules involved.

[41] A brief introduction to the rationale underlying cellular modeling can be found at the National Resource for Cell Analysis and Modeling (http://www.nrcam.uchc.edu/applications/applications.html).

TABLE 5.1 Sample Simulation Programs

Name	Descriptors[a]	Web Site
Gepasi/Copasi	fkFW	http://gepasi.dbs.aber.ac.uk/softw/gepasi.html
BioSim	qWMU	http://www.molgen.mpg.de/~biosim/BioSim/BioSimHome.html
Jarnac	krfbFWS	http://members.tripod.co.uk/sauro/Jarnac.htm
MCELL	rsU	http://www.mcell.cnl.salk.edu/
Virtual Cell	ksDFWMU	http://www.nrcam.uchc.edu/
E-Cell	kWUS	http://www.e-cell.org/
Neuron	ksFWMUS	http://neuron.duke.edu/
Genesis	ksUS	http://www.bbb.caltech.edu/GENESIS/genesis.html
Plas	kfbFW	http://correio.cc.fc.ul.pt/~aenf/plas.html
Ingeneue	qkFMWUS	http://www.ingeneue.org/
DynaFit	kfW	http://www.biokin.com/dynafit/
Stochsim	rS	http://www.zoo.cam.ac.uk/comp-cell/StochSim.html
T7 Simulator	kUS	http://virus.molsci.org/t7/
Molecularizer/Stochastirator	krUS	http://opnsrcbio.molsci.org/alpha/comps/sim.html

NOTE: All packages have facilities for chemical kinetic simulation of one sort or another. Some are better designed for metabolic systems, others for electrochemical systems, and still others for genetic systems.

[a] The descriptors are as follows: b, bifurcation analyses and steady-state calculation; f, flux balance or metabolic control and related analyses; k, deterministic kinetic simulation; q, qualitative simulation; r, stochastic process models; s, spatial processes; D, database connectivity; F, fitting, sensitivity, and optimization code; M, runs on Macintosh; S, source code available; U, runs on Linux or Unix; W, runs on windows.

There are many different tools for simulating and analyzing models of cellular systems (Table 5.1). More general tools, such as Mathematica and MATLAB or other systems that can be used for solving systems of differential or stochastic-differential equations, can be used to develop simulations, and because these tools are commonly used by many researchers, their use facilitates the transfer of models among different researchers. Another approach is to link data gathering and biological information systems to software that can integrate and predict behavior of interacting components (currently, researchers are far from this goal, but see Box 5.5 and Box 5.6). Finally, several platform-independent model specification languages are under development that will facilitate greater sharing and interoperability. For example, SBML,[42] Gepasi,[43] and CellML[44] are specialized systems for biological and biochemical modeling. Madonna[45] is a general-purpose system for solving a variety of equations (differential equations, integral equations, and so on).

Rice and Stolovitzky describe the task of inferring signaling, metabolic, or gene regulatory pathways from experimental data as one of reverse engineering.[46] They note that automated, high-throughput methods that collect species- and tissue-specific datasets in large volume can help to deal with the risks in generalizing signaling pathways from one organism to another. At the same time, fully detailed kinetic models of intracellular processes are not generally feasible. Thus, one step is to consider models that describe network topology (i.e., that identify the interactions between nodes in the system—genes, proteins, metabolites, and so on). A model with more detail would describe network topology that is causally directional (i.e., that specifies which entities serve as input to others). Box 5.7 provides more detail.

[42] See http://www.cds.caltech.edu/erato/sbml/.

[43] See http://www.gepasi.org/.

[44] See http://www.cellml.org/.

[45] See http://www.berkeleymadonna.com/.

[46] J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2:70-77, 2004.

Box 5.5
BioSPICE

BioSPICE, the Biological Simulation Program for Intra-Cellular Evaluation, is in essence a modeling framework that provides users with model components, tools, databases, and infrastructure to develop predictive dynamical models of cellular function. BioSPICE seeks to promote a synergy between experiment and model, in which model predictions drive experiment and experimental results identify areas in which a given model needs to be improved, and the intent is that researchers go from data to models to analysis and hypothesis generation, iteratively refining their understanding of the biological processes.

An important component of BioSPICE is a library of experimentally validated (and hence trusted) model components that can be used as starting points in larger-scale simulations, as elements from this library are composed in new ways or adapted to investigate other biological systems. Many biological parts and processes are represented as components, including phosphorylization events, chemotaxis, and conserved elements of various pathways. Also, because BioSPICE is designed as an open-source environment, it is hoped that the user community itself will make available a repertoire of model components that span a wide range of spatial, temporal, and functional scales, including those that simulate a single chemical reaction with high fidelity, those that simulate entire pathways, and those that simulate more abstract higher-order motifs.

BioSPICE tools are intended to enable researchers to use public databases and local resources to formulate a qualitative description of the cellular process of interest (e.g., models of networks or pathways), to annotate the links between entities with biochemical interactions, and finally to convert this annotated qualitative description to a set of equations that can be analyzed and simulated. In addition, BioSPICE provides a number of simulation engines with the capability to simulate ordinary, stochastic, and partial differential equations and other tools that support stability and bifurcation analysis and qualitative reasoning that combines probabilistic and temporal logic.

SOURCE: Sri Kumar, Defense Advanced Research Projects Agency, June 30, 2003.

An example of a cellular simulation environment is E-CELL, an open-source system for modeling biochemical and genetic processes. Organizationally, E-CELL is an international research project aimed at developing theoretical and functioning technologies to allow precise "whole cell" simulation; it is supported by the New Energy and Industrial Technology Development Organization (NEDO) of Japan.

E-CELL simulations allow a user to model hypothetical virtual cells by defining functions of proteins, protein-protein interactions, protein-DNA interactions, regulation of gene expression, and other features of cellular metabolism.[47] Based on reaction rules that are known through experiment and assumed concentrations of various molecules in various locations, E-CELL numerically integrates differential equations implicitly described in these reaction rules, resulting in changes over time in the concentrations of proteins, protein complexes, and other chemical compounds in the cell.

Developers hope E-CELL will ultimately allow investigators a cheap, fast way to screen drug candidates, study the effects of mutations or toxins, or simply probe the networks that govern cell behavior. One application of E-CELL has been to construct a model of a hypothetical cell capable of

[47] See http://www.e-cell.org/project/. For a view of the computer science challenges, see also K. Takahashi, K. Yugi, K. Hashimoto, Y. Yamada, C.J.F. Pickett, and M. Tomita, "Computational Challenges in Cell Simulation: A Software Engineering Approach," *IEEE Intelligent Systems* 17(5):64-71, 2002.

Box 5.6 Cytoscape

A variety of computer-aided models has been developed to simulate biological networks, typically focusing on specific cellular processes or single pathways.[1] Cytoscape is a modeling environment particularly suited to the analysis of global data on network interactions (from high-throughput screens for protein-protein, protein-DNA, and gene interactions) and on network states (including data on gene expression, protein abundance, and metabolite concentrations.) The Java-based, open-source software uses plug-ins to incorporate analyses of individual processes and pathways.[2]

A model in Cytoscape is organized as a network graph, with molecular species represented as nodes and interactions represented as edges between nodes. Nodes and edges are mapped to specific data values called *attributes* that can be text strings, discrete or continuous numbers, URLs, or lists, either loaded from a data repository or generated dynamically. Layered onto attributes are *annotations*, which represent a hierarchical classification of progressively more specific descriptions (such as functions) of groups of nodes and edges. It is possible to have many levels of annotation active simultaneously, each displayed as a different attribute of a node or edge. To visualize the network, Cytoscape supports several layout algorithms that fix the relative locations of specific nodes and edges in the graphical window. An *attribute-to-visual mapping* facility allows attributes to determine the appearance (color, shape, size) of their associated nodes and edges. *Graph selection and filtering* reduces the complexity of the network by selectively displaying subsets of nodes and edges according to a variety of criteria.

Cytoscape's plug-in extensibility addresses the challenge of bridging high-level information (relationships among network components) with lower-level information (reaction rates, binding constants) of specific processes. A plug-in that organizes the network layout according to putative functional attributes of genes was used to study energy transduction pathways in *Halobacterium*.[3] Another plug-in allows Cytoscape to simulate stochastic SBML-biochemical models.[4] The authors hope a community will further develop and enhance Cytoscape.

[1] A. Gilman and A.P. Arkin, "Genetic 'Code': Representations and Dynamical Models of Genetic Components and Networks," *Annual Review of Genomics and Human Genetics* 3:341-369, 2002

[2] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, et al., "Integrated Models of Biomolecular Interaction Networks," *Genome Research* 13:2498-2504, 2003.

[3] N.S. Baliga, M. Pan, Y.A. Goo, E.C. Yi, D.R. Goodlett, K. Dimitrov, P. Shannon, et al., "Coordinate Regulation of Eenergy Transduction Modules in *Halobacterium* species Analyzed by a Global Systems Approach," *Proceedings of the National Academy of Sciences* 99(23):14913-14918, 2002.

[4] M. Hucka, A. Finney, H.M. Sauro, H. Bolouri, J. Doyle, and H. Kitano, "The ERATO Systems Biology Workbench: Enabling Interaction and Exchange Between Software Tools for Computational Biology," *Pacific Symposium in Biocomputing,* 450-461, 2002.

SOURCE: Adapted from P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin et al., "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Research* 13(11):2498-2504, 2003.

transcription, translation, energy production, and phospholipid synthesis with only 127 genes. Most of these genes were taken from *Mycoplasma genitalium*, the organism with the smallest known chromosome (the complete genome sequence is 580 kilobases).[48] E-CELL has also been used to construct a computer model of the human erythrocyte,[49] to estimate a gene regulatory network and signaling

[48] M. Tomita, K. Hashimoto, K. Takahashi, Y. Matsuzaki, R. Matsushima, K. Saito, K. Yugi, et al., "E-CELL Project Overview: Towards Integrative Simulation of Cellular Processes," *Genome Informatics* 9:242-243, 1998, available at http://giw.ims.u-tokyo.ac.jp/giw98/cdrom/Poster-pdf/poster02.pdf.

[49] M. Tomita et al., "In Silico Analysis of Human Erythrocyte Using E-Cell System," poster session, The Future of Biology in the 21st Century: 2nd International Conference on Systems Biology, California Institute of Technology, Pasadena, November 4-7, 2001, available at http://www.icsb2001.org/Posters/032_kinoshita.pdf.

Box 5.7
Pathway Reconstruction: A Systems Approach

On Topology.

In this level, we are only concerned with identifying the interaction between nodes (genes, proteins, metabolites, etc.) in the system. The goal is the generation of a diagram of non-directional connections between all interacting nodes. For example, many have sought to develop large-scale maps of protein–protein interactions derived from various sources. Two-hybrid studies have produced genome-wide interaction maps for *E. coli* bacteriophage T7, yeast, *Drosophila,* and *C. elegans*. Although this approach can be comprehensive in regard to being genome wide, many interactions are not reproducible (a potential source of false negatives) and putative interactions occur between unlikely protein combinations (a potential source of false positives). . . . Another approach to constructing large-scale connection maps is by mining databases. Specific databases of protein interactions are being developed, the largest of which are DIP and BIND. These databases combine data from many high-throughput experiments along with data from other sources, such as published literature. . . . Along other lines, investigators have attempted to identify topological links by analyzing the dynamic behavior of networks. Pioneering work in this area shows that metabolic network topologies can be derived from correlation of time-series measurements of species concentrations. The method is further refined to better identify connections in non-linear systems using mutual information instead of correlation. In another method, pair-wise correlation of gene expression data is used to predict functional connections that could then be combined into "relevance networks" of linked genes. Other methods may seek to use some combination of data sources, although this may not be completely straightforward.

On Inferring Qualitative Connections.

In this level, we include not only associations between cellular entities but also the causal relations of such associations, such as which entities serve as input to others. . . . Researchers have proposed methods that infer connectivities from the estimations of the Jacobian matrix for metabolic, signaling, and genetic networks. Ross and co-workers have proposed a method based on propagated perturbations of chemical species that can reconstruct causal sequences of reactions from synthetic and experimental data. To reconstruct gene regulatory systems, methods include fuzzy logic analysis of facilitator/repressor groups in the yeast cell cycle and reconstruction of binary networks. However, the wide application of such methods is often limited because the continuous nature of many biological systems prevents easy abstractions into coarser signals. Recently, there has been considerable work using Bayesian network inference. Examples include inferring gene regulation using gene expression data from the yeast cell cycle or using data from synthetic gene networks.

SOURCE: Reprinted by permission from J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2(2):70-77. Copyright 2004 Elsevier. (References omitted.)

pathway involved in the circadian rhythm of *Synechococcus* sp. PCC 7942,[50] and to model mitochondrial energy metabolism and metabolic pathways in rice.[51]

Another cellular simulation environment is the Virtual Cell, developed at the University of Connecticut Health Center.[52] The Virtual Cell is a tool for experimentalists and theoreticians for computationally testing hypotheses and models. To address a particular question, these mechanisms (chemical kinetics, membrane fluxes and reactions, ionic currents, and diffusion) are combined with a specific set of experimental conditions (geometry, spatial scale, time scale, stimuli) and applicable conservation laws to specify

[50] F. Miyoshi et al., "Estimation of Genetic Networks of the Circadian Rhythm in Cyanobacterium Using the E-CELL system," poster session, presented at US-Japan Joint Workshop on Systems Biology of Useful Microorganisms, September 6-18, 2002, Keio University, Yamagata, Japan, available at http://nedo-doe.jtbcom.co.jp/abstracts/35.pdf.

[51] E. Wang et al., "e-Rice Project: Reconstructing Plant Cell Metabolism Using E-CELL System," poster session presented at Systems Biology: The Logic of Life—3rd International Conference on Systems Biology, December 13-15, 2002, Karolinska Institutet, Stockholm, available at http://www.ki.se/icsb2002/pdf/ICSB_222.pdf.

[52] L.M. Loew and J.C. Schaff, "The Virtual Cell: A Software Environment for Computational Cell Biology," *Trends in Biotechnology* 19(10):401-406, 2001.

a concrete system of differential and algebraic equations. This experimental geometry may assume well-mixed compartments or a one-, two-, or three-dimensional spatial representation (e.g., experimental images from a microscope). Models are constructed from biochemical and electrophysiological data mapped to appropriate subcellular locations in images obtained from a microscope. A variety of modeling approximations are available including pseudo-steady state in time (infinite kinetic rates) or space (infinite diffusion or conductivity). In the case of spatial simulations, the results are mapped back to experimental images and can be analyzed by applying the arsenal of image-processing tools that is familiar to a cell biologist. Section 5.4.2.4 describes a study undertaken within the Virtual Cell framework.

Simulation models can be useful for many purposes. One important use is to facilitate an understanding of what design properties of an intracellular network are necessary for its function. For example, von Dassow et al.[53] used a simulation model of the gap and pair-rule gene network in *Drosophila melanogaster* to show that the structure of the network is sufficient to explain a great deal of the observed cellular patterning. In addition, they showed that the network behavior was robust to parameter variation upon the addition of hypothetical (but reasonable) elements to the known network. Thus, simulations can also be used to formally propose and justify new hypothetical mechanisms and predict new network elements.

Another use of simulation models is in exploring the nature of control in networks. An example of exploring network control with simulation is the work of Chen et al.·[54] in elucidating the control of different phases of mitosis and explaining the impact of 50 different mutants on cellular decisions related to mitosis.

Simulations have also been used to model metabolic pathways. For example, Edwards and Palsson developed a constraint-based genome-scale simulation of *Escherichia coli* metabolism (Box 5.8). By applying successive constraints (stoichiometric, thermodynamic, and enzyme capacity constraints) to the metabolic network, it is possible to impose limits on cellular, biochemical, and systemic functions, thereby identifying all allowable solutions (i.e., those that do not violate the applicable constraints). Compared to the detailed theory-based models, such an approach has the major advantage that it does not require knowledge of the kinetics involved (since it is concerned only with steady-state function). (On the other hand, it is impossible to implement without genome-scale knowledge, because only genome-scale knowledge can bound the system in question.) Within the space of allowable solutions, a particular solution corresponds to the maximization of some selected function, such as cellular growth or a response to some environmental change. A more robust model accounting for a larger number of pathways is also described in Box 5.8.

The Edwards and Palsson model has been used to predict the evolution of *E. coli* metabolism under a variety of environmental conditions. In the words of Ibarra et al., "When placed under growth selection pressure, the growth rate of *E. coli* on glycerol reproducibly evolved over 40 days, or about 700 generations, from a sub-optimal value to the optimal growth rate predicted from a whole-cell in silico model. These results open the possibility of using adaptive evolution of entire metabolic networks to realize metabolic states that have been determined a priori based on in silico analysis."[55]

Simulation models can also be used to test design ideas for engineering networks in cells. For example, very simple models have been used to provide insight into a genetic oscillator and a switch in *E. coli*.[56] Models have also been used to test designs for the control of cellular networks, as illustrated by

[53] G. Von Dassow, E. Meir, E.M. Munro, and G.M. Odell, "The Segment Polarity Network Is a Robust Developmental Module," *Nature* 406(6792):188-192, 2000.

[54] K.C. Chen, A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novak, and J.J. Tyson, "Kinetic Analysis of a Molecular Model of the Budding Yeast Cell Cycle," *Molecular Biology of the Cell* 11(1):369-391, 2000.

[55] R.U. Ibarra, J.S. Edwards, and B.O. Palsson, "*Escherichia coli* K-12 Undergoes Adaptive Evolution to Achieve in Silico Predicted Optimal Growth," *Nature* 420(6912):186-189, 2002.

[56] M.B. Elowitz and S. Leibler, "A Synthetic Oscillatory Network of Transcriptional Regulators," *Nature* 403(6767):335-338, 2000; T.S. Gardner, C.R. Cantor, and J.J. Collins, "Construction of a Genetic Toggle Switch in Escherichia coli," *Nature* 403(6767):339-342, 2000.

Box 5.8
Escherichia coli Constraint-based Models

A. In Silico Model[1]

The *Escherichia coli* MG1655 genome has been completely sequenced. The annotated sequence, biochemical information, and other information were used to reconstruct the *E. coli* metabolic map. The stoichiometric coefficients for each metabolic enzyme in the *E. coli* metabolic map were assembled to construct a genome-specific stoichiometric matrix. The *E. col*i stoichiometric matrix was used to define the system's characteristics and the capabilities of *E. coli* metabolism. The effects of gene deletions in the central metabolic pathways on the ability of the in silico metabolic network to support growth were assessed, and the in silico predictions were compared with experimental observations. It was shown that based on stoichiometric and capacity constraints the in-silico analysis was able to qualitatively predict the growth potential of mutant strains in 86% of the cases examined. Herein, it is demonstrated that the synthesis of in silico metabolic genotypes based on genomic, biochemical, and strain-specific information is possible, and that systems analysis methods are available to analyze and interpret the metabolic phenotype.

B. Genome-scale Model[2]

An expanded genome-scale metabolic model of *E. coli* (iJR904 GSM/GPR) has been reconstructed which includes 904 genes and 931 unique biochemical reactions. The reactions in the expanded model are both elementally and charge balanced. Network gap analysis led to putative assignments for 55 open reading frames (ORFs). Gene to protein to reaction associations (GPR) are now directly included in the model. Comparisons between predictions made by iJR904 and iJE660a models show that they are generally similar but differ under certain circumstances. Analysis of genome-scale proton balancing shows how the flux of protons into and out of the medium is important for maximizing cellular growth. . . . *E. coli* iJR904 has improved capabilities over iJE660a [a model that accounted for 660 genes and 627 unique biochemical reactions and was itself a slight modification of the original model described in the above paragraph]. iJR904 is a more complete and chemically accurate description of *E. coli* metabolism than iJE660a. Perhaps most importantly, iJR904 can be used for analyzing and integrating the diverse datasets. iJR904 will help to outline the genotype-phenotype relationship for *E. coli* K-12, as it can account for genomic, transcriptomic, proteomic and fluxomic data simultaneously.

[1] Reprinted from J.S. Edwards and B.O. Palsson, "The *Escherichia coli* MG1655 in Silico Metabolic Genotype: Its Definition, Characteristics, and Capabilities," *Proceedings of the National Academy of Sciences* 97(10): 5528-5533, 2000. Copyright 2000 National Academy of Sciences.

[2] J.L. Reed, T.D. Vo, C.H. Schilling, and B.O. Palsson, "An Expanded Genome-scale Model of *Escherichia coli* K-12 (iJR904 GSM/GPR)," *Genome Biology* 4(9): Article R54, 2003, available at http://genomebiology.com/2003/4/9/R54. Reprinted by permission of the authors.

Endy and Yin in using their T7 model to propose a pharmaceutical strategy for preventing both T7 propagation and the development of drug resistance through mutation.[57]

Given observed cell behavior, simulation models can be used to suggest the necessity of a given regulatory motif or the sufficiency of known interactions to produce the phenomenon. For example, Qi et al. demonstrate the sufficiency of membrane energetics, protein diffusion, and receptor-binding kinetics to generate a particular dynamic pattern of protein location at the synapse between two immune cells.[58]

The following sections describe several simulation studies in more detail.

[57] D. Endy and J. Yin, "Toward Antiviral Strategies That Resist Viral Escape," *Antimicrobial Agents and Chemotherapy* 44(4):1097-1099, 2000.

[58] S.Y. Qi, J.T. Groves, and A.K. Chakraborty, "Synaptic Pattern Formation During Cellular Recognition," *Proceedings of the National Academy of Sciences* 98(12):6548-6553, 2001.

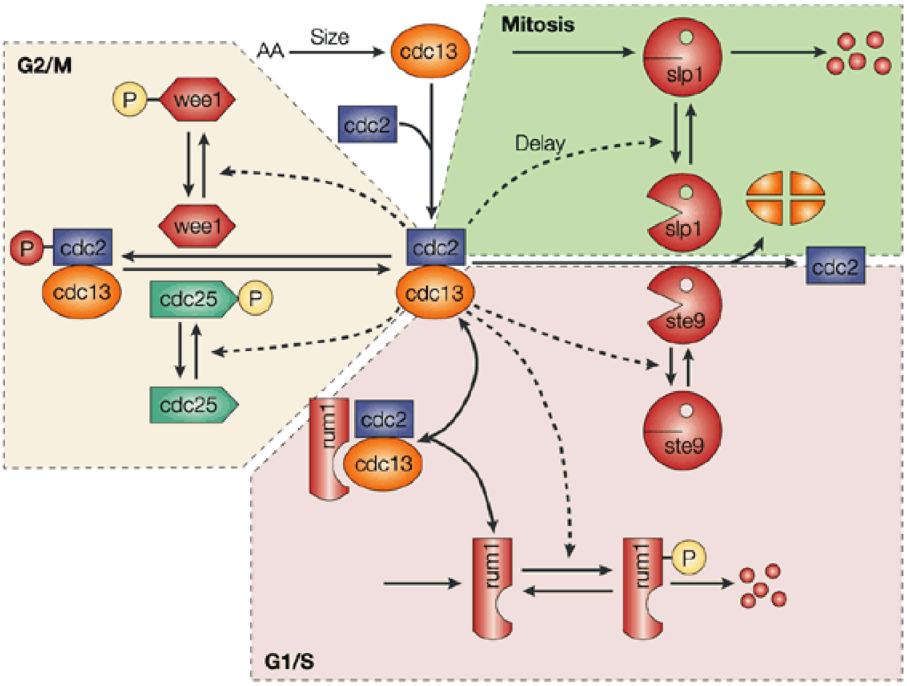


FIGURE 5.5 The cell-cycle control system in fission yeast. This system can be divided into three modules, which regulate the transitions from G1 into S phase, from G2 into M phase, and exit from mitosis. SOURCE: J.J. Tyson, K. Chen, and B. Novak, "Network Dynamics and Cell Physiology," *Nature Reviews of Molecular Cell Biology* 2(12):908-916, 2001. Figure and caption reproduced with permission from *Nature Reviews of Molecular Cell Biology*. Copyright 2001 Macmillan Magazines Ltd.

5.4.2.2 Cell Cycle Regulation

Biological growth and reproduction depend ultimately on the cycle of DNA synthesis and physical separation of the replicate DNA molecules within individual cells. In eukaryotes, these processes are triggered by cyclin-dependent protein kinases (CDKs). In fission yeast, CDK activity (cdc2 = kinase subunit, cdc13 = cyclin subunit) is regulated by a network of protein interactions (Figure 5.5), including cyclin synthesis and degradation, phosphorylation of cdc2, and binding to an inhibitor.

A network of such complexity, with multiple feedback loops, cannot be understood thoroughly by casual intuition. Instead, the network is converted into a set of nonlinear differential equations, and the physiological implications of these equations are studied.[59] Numerical simulation of the equations (Figure 5.6) provides complete time courses of every component and can be interpreted in terms of observable events in the cell cycle. Simulations can be run, not only of wild-type cells but also of dozens of mutants constructed by deleting or overexpressing each component singly or in multiple combinations. From the observed phenotypes of these mutants it is possible to reverse-engineer the regulatory network and the set of kinetic constants associated with the component reactions.

[59] J.J. Tyson, K. Chen, and B. Novak, "Network Dynamics and Cell Physiology," *Nature Reviews: Molecular Cell Biology* 2(12):908-916, 2001; J.J. Tyson, A. Csikasz-Nagy, and B. Novak, "The Dynamics of Cell Cycle Regulation," *BioEssays* 24(12):1095-1109, 2002.

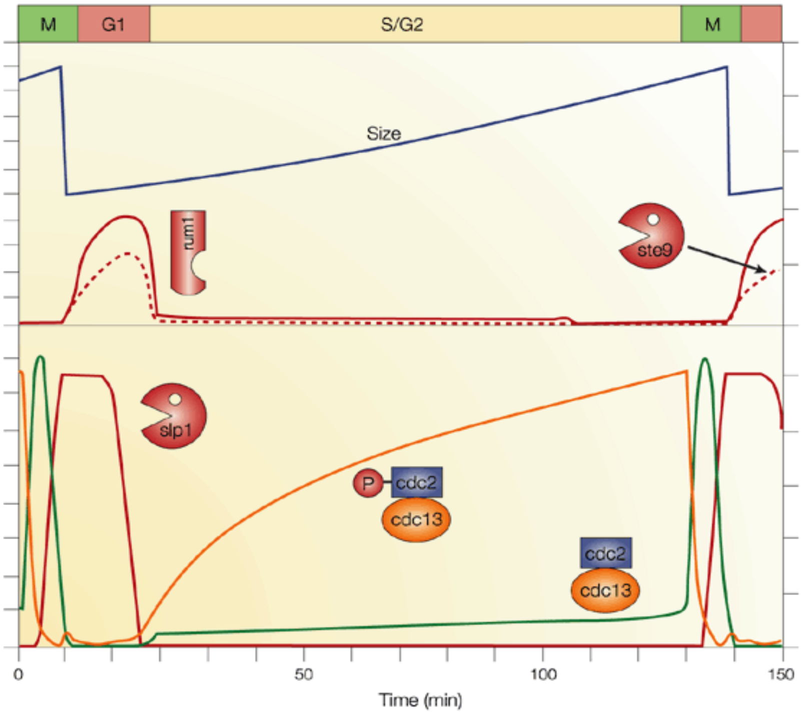


FIGURE 5.6 Simulated time courses of cdc2 and related proteins during the cell cycle of fission yeast. Numerical integration of the full set of differential equations that describe the wiring diagram in Figure 5.5 yields these time courses. Time is expressed in minutes; all other variables are given in arbitrary units. "Size" refers to the number of ribosomes per nucleus. Notice the brief G1 phase, when ste9 is active and rum1 is abundant. After a long S/G2 phase, during which cdc2 is tyrosine phosphorylated, the cell enters M phase, when cdc25 removes the inhibitory phosphate group. After some delay, slp1 activates and degrades cdc13. As cdc2–cdc13 activity falls, the cell exits mitosis. Size decreases twofold at nuclear division. SOURCE: J.J. Tyson, K. Chen, and B. Novak, "Network Dynamics and Cell Physiology," *Nature Reviews of Molecular Cell Biology* 2(12):908-916, 2001. Figure and caption reproduced with permission from *Nature Reviews of Molecular Cell Biology.* Copyright 2001 Macmillan Magazines Ltd.

For understanding the dynamics of molecular regulatory systems, bifurcation theory is a powerful complement to numerical simulation. The bifurcation diagram in Figure 5.7 presents recurrent solutions (steady states and limit cycle oscillations) of the differential equations as functions of cell size. The control system has three characteristic steady states: low cdc2 activity (G1 = pre-replication), medium cdc2 activity (S/G2 = replication and post-replication), and high cdc2 activity (M = separation of replicated DNA molecules). G1 and S/G2 are stable steady states; M is unstable because of a negative feedback loop as shown in Figure 5.5 (cdc2-cdc13 activates Slp1, which degrades cdc13).

When the time courses of size and cdc2 activity from Figure 5.6 are superimposed on the bifurcation diagram (curve labeled "size"), one sees how progress through the cell cycle is governed by the bifurcations that turn stable steady states into unstable steady states and/or stable oscillations. A mutation changes a specific rate constant, which changes the locations of the bifurcation points in Figure 5.7, which changes how cells progress through (or halt in) the cell cycle. By this route one can trace the dynamical consequences of genetic information all the way to observable cell behavior.

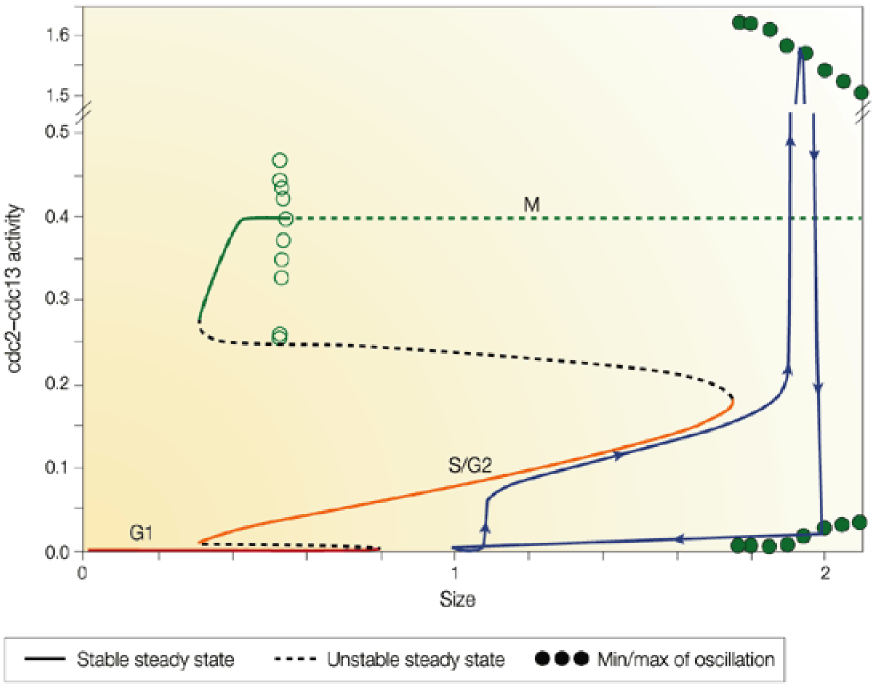


FIGURE 5.7 Bifurcation diagram for the full cell-cycle control network. . . . [T]he full diagram is not a simple sum of the bifurcation diagrams of its modules. In particular, oscillations around the M state are greatly modified in the composite control system. Superimposed on the bifurcation diagram is a "cell-cycle orbit" (line on the right with arrows): from the time courses in Figure 5.6, we plot size on the abscissa and cdc2–cdc13 activity on the ordinate for representative times between birth and division. Notice that, at small cell size, all three modules support stable steady states. Notice how the cell-cycle orbit follows the attractors of the control system. SOURCE: J.J. Tyson, K. Chen and B. Novak, "Network Dynamics and Cell Physiology," *Nature Reviews Molecular Cell Biology* 2(12):908-916, 2001. Figure and caption, reproduced with permission from *Nature Reviews Molecular Cell Biology*. Copyright 2001 Macmillan Magazines Ltd.

5.4.2.3 A Computational Model to Determine the Effects of SNPs in Human Pathophysiology of Red Blood Cells

The completion of the Human Genome Project has led to the construction of single nucleotide polymorphism (SNP) maps. Single nucleotide polymorphisms are common DNA sequence variations among individuals. A result of the construction of SNP maps is to determine the effects of SNPs on the development of disease(s) since sequence variations can lead to altered biological function or disease.

Currently, it is difficult to determine the causal relationship between the variations in sequence, SNPs, and the physiological function. One way to analyze this relationship is to create computational models or simulations of biological processes. Since erythrocyte (red blood cell) metabolism has been studied extensively over the years and many SNPs have been characterized, Jamshidi et al. used this information to build their computational models.[60]

Two important metabolic enzymes, glucose-6-phosphate dehydrogenase (G6PD) and pyruvate kinase (PK), were studied for alterations in their kinetic properties in an in silico model to calculate the overall effect of SNPs on red blood cell function. Defects in these enzymes cause hemolytic anemia.

[60] N. Jamshidi, S.J. Wiback, and B.O. Palsson, "In Silico Model-driven Assessment of the Effects of Single Nucleotide Polymorphisms (SNPs) on Human Red Blood Cell Metabolism," *Genome Research* 12(11):1687-1692, 2002.

Clinical data taken from the published literature were used for the measured values of the kinetic parameters. These values were then used in model simulations to determine whether a direct link could be established between the SNP and the disease (anemia).

The computational modeling revealed two results. For the G6PD and PK variants analyzed, there appeared to be no clear relationship between their kinetic properties as a function of sequence variation or SNP. However, upon assessment of overall biological function, a correlation was found between the sequence variation of G6PD and the severity of the clinical disease. Thus, in silico modeling of biological processes may aid in analysis and prediction of SNPs and pathophysiological conditions.

5.4.2.4 Spatial Inhomogeneities in Cellular Development

Simulation models can be used to provide insight into the significance of spatial inhomogeneities. For example, the interior of living cells does not resemble at all a uniform aqueous solution of dissolved chemicals, and yet this is the implicit assumption underlying many views of the cell. This assumption serves traditional biochemistry and molecular biology reasonably well, but research increasingly demonstrates that the physical locations of specific molecules are crucial. Multiprotein complexes act as machines for internal movements or as integrated circuits in signaling. Messenger RNA molecules are transported in a highly directed fashion to specific regions of the cell (in nerve axons, for example). Cells adopt highly complex shapes and undergo complex movements thanks to the matrix of protein filaments and associated proteins within their cytoplasm.

5.4.2.4.1 Unraveling the Physical Basis of Microtubule Structure and Stability Microtubules are cylindrical polymers found in every eukaryotic cell. Microtbles play a role in cellular architecture and as molecular train tracks used to transport everything from chromosomes to drug molecules. An understanding of microtubule structure and function is key not just to unraveling fundamental mechanisms of the cell, but also to opening the way to the discovery of new antiparasitic and anticancer drugs.

Until now, researchers have known that the microtubules, constructed of units called protofilaments in a hollow, helical arrangement, are rigid but not static, and undergo periods of growth and sudden collapse. Yet the mechanism for this construction-destruction had eluded researchers.

Over the past several years, McCammon and his colleagues have pioneered the use of a combination of an atomically detailed model for a microtubule and large-scale computations using the adaptive Poisson-Boltzmann Solver to create a high-resolution, 1.25-million-atom map of the electrostatic interactions within the microtubule.[61]

More recently, David Sept and Nathan Baker of Washington University and McCammon used the same technique to successfully predict the helicity of the tubule with a striking correspondence to experimental observation.[62] Based on the lateral interactions between protofilaments, they determined that the microtubule prefers to be in a configuration in which the protofilaments assemble with a seam at each turn, rather than spiraling smoothly upward with alpha and beta monomers wrapping the microtubule as if it were a barber's pole. At the end of each turn, a chain of alphas is trailed by a chain of betas, then after that turn, a chain of alphas, and so on. It is as if the red and white stripes on the barber's pole traded places with every twist (Figure 5.8).

[61] N.A. Baker, D. Sept, S. Joseph, M.J. Holst, and J.A. McCammon, "Electrostatics of Nanosystems: Application to Microtubules and the Ribosome," *Proceedings of the National Academy of Sciences* 98(18):10037-10041, 2001.

[62] D. Sept, N.A. Baker, and J.A. McCammon, "The Physical Basis of Microtubule Structure and Stability," *Protein Science* 12(10):2257-2261, 2003.

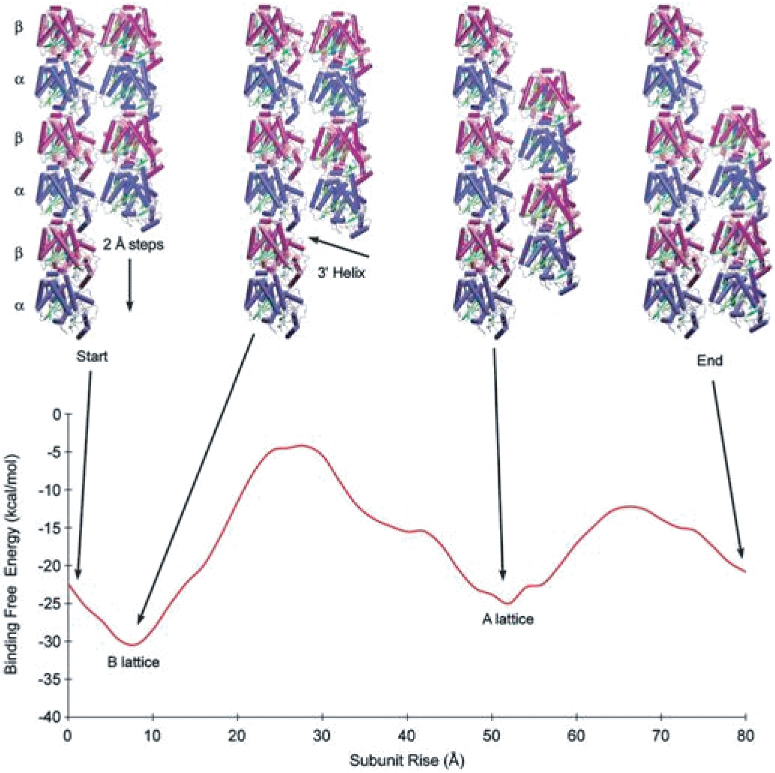


FIGURE 5.8 The binding free energy between two protofilaments as a function of the subunit rise between adjacent dimmers. Sept et al. used electrostatic calculations to determine the binding energy between two protofilaments as a function of the subunit rise between adjacent dimers. Viewed from the growing (+) end of the tubule, the graph demonstrates the most favorable configuration at various points during assembly. SOURCE: Reprinted by permission from D. Sept, N.A. Baker, and J.A. McCammon, "The Physical Basis of Microtubule Structure and Stability," *Protein Science* 12:2257-2261, 2003. Copyright 2003 by the Protein Society.

5.4.2.4.2 The Movement of* Listeria *Bacteria Alberts and Odell have developed a computational model of *Listera monocytogenes* based on an explicit simulation of a large number of monomer-scale biochemical and mechanical interactions,[63] representing all protein-protein binding interactions with on-rate and off-rate kinetic equations. These equations characterize individual actin filaments: the bulk properties of the actin "gel" arise from the contributions of the many individual filaments of the actin network; and the growth of any particular filament depends on that filament's precise location, orientation, and biochemical state, all of which change through time. Mechanical interactions, which resolve collisions and accommodate the stretching of protein-protein linkages, follow Newton's laws.

The model is based on a large set of differential equations that determine how the relevant state variables change with time. These equations are solved numerically, taking into account the fact that discontinuities in time occur frequently as objects suddenly collide and as objects suddenly spring into existence or disappear (due to new filament nucleation and depolymerization). The model accommo-

[63] J.B. Alberts and G.M. Odell, "In Silico Reconstitution of Listeria Propulsion Exhibits Nano-Saltation," *PLoS Biology* 2(12):e412, 2004, available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=532387.

dates arbitrary geometries, explicit stochastic input, and specific small-scale events. Because the model is built from the ground up, it can predict emergent behavior that would not be apparent from intuition or qualitative description of the behavior of individual parts. On the other hand, the simulation requires multiple runs of its stochastic, individual molecule-based model, and parametric relationships emerge not from closed-form equations that demonstrate qualitative functional dependencies but from ensembles of many repeated simulations.

The trajectories generated by this model of *L. monocytogenes* motility display repeated runs and pauses that closely resemble the actual nanoscale measurements of bacterial motion.[64] Further analysis of the simulation state at the beginning and ends of simulated pauses indicate that there is no characteristic step-size or pause duration in these simulated trajectories and that pauses can be caused both by correlated Brownian motion and by synchronously strained sets of ActA-actin filament mechanical links.

5.4.2.4.3 Morphological Control of Spatiotemporal Patterns of Intracellular Signaling Fink and Slepchenko studied calcium waves evoked by activation of the bradykinin receptor in the plasma membrane of a neuronal cell.[65] The neuromodulator bradykinin applied to the cells produced a calcium wave that starts in the neurite and spreads to the soma and growth cones. The calcium wave was monitored with digital microscope imaging of a fluorescent calcium indicator. The hypothesis was that interaction of bradykinin with its receptor on the plasma membrane activated production of inositol-1,4,5-trisphosphate ($InsP_3$) that diffused to its receptor on the endoplasmic reticulum, leading to calcium release.

Using the Virtual Cell software environment, they assembled a simulation model of this phenomenon.[66] The model contained details of the relevant receptor distributions (via immunofluorescence) within the cell geometry, the kinetics of $InsP_3$ production (via biochemical analysis of $InsP_3$ in cell populations and photorelease of caged $InsP_3$ in individual cells), the transport of calcium through the $InsP_3$ receptor calcium channel and the sarcoplasmic/endoplasmic reticulum calcium ATPase (SERCA) pump (from literature studies of single-channel kinetics and radioligand flux), and calcium buffering by both endogenous proteins and the fluorescent indicator (from confocal measurements of indicator concentrations).

The mathematical equations generated by this combination of molecular distributions and reaction and membrane transport kinetics were then solved to produce a simulation of the spatiotemporal pattern of calcium that could be directly compared to the experiment. The characteristic calcium dynamics requires rapid, high-amplitude production of $[InsP_3]_{cyt}$ in the neurite. This requisite $InsP_3$ spatiotemporal profile is provided, in turn, as an intrinsic consequence of the cell's morphology, demonstrating how geometry can locally and dramatically intensify cytosolic signals that originate at the plasma membrane. In addition, the model predicts and experiments confirm that stimulation of just the neurite, but not the soma or growth cone, is sufficient to generate a calcium response throughout the cell.

[64] S.C. Kuo and J.L. McGrath, "Steps and Fluctuations of Listeria Monocytogenes During Actin-based Motility," *Nature* 407(6807):1026-1029, 2000; J. McGrath, N. Eungdamrong, C. Fisher, F. Peng, L. Mahadevan, T.J. Mitchison, and S.C. Kuo, "The Force-Velocity Relationship for the Actin-based Motility of Listeria Moncytogenes," *Current Biology* 13(4):329-332, 2003. (Both cited in Alberts and Odell, 2004.)

[65] C.C. Fink, B. Slepchenko, I.I. Moraru, J. Schaff, J. Watras, and L.M. Loew, "Morphological Control of Inositol-1,4,5-Trisphosphate-dependent Signals." *Journal of Cell Biology* 147(5):929-935, 1999; C.C. Fink, B. Slepchenko, I.I. Moraru, J. Watras, J.C. Schaff, and L.M. Loew, "An Image-based Model of Calcium Waves in Differentiated Neuroblastoma Cells," *Biophysical Journal* 79(1):163-183, 2000.

[66] B.M. Slepchenko, J.C. Schaff, I. Macara, and L.M. Loew, "Quantitative Cell Biology with the Virtual Cell," *Trends in Cell Biology* 13(11):570-576, 2003.

5.4.3 Genetic Regulation

The problem of genetic regulation—how and under what circumstances and the extent to which genes are expressed as proteins—is a central problem of modern biology. The issue originates in an apparent paradox—every cell in a complex organism contains the same DNA sequences, and yet there are many cell types in such organisms (blood cells, skin cells, and so on). In particular, the proteins that comprise any given cell type are different from those of other cell types, even though the genomic information is the same in both. Nor is genomic information the whole story in development—cells also respond to their environment, and external signals coming into a cell from neighboring cells influence which proteins the cell makes.

Genetic regulation is an extraordinarily complex problem. Molecular biologists distinguish between *cis*-regulation and *trans*-regulation. *Cis*-regulatory elements for a given gene are segments of the genome that are located in the vicinity of the structural portion of a gene and regulate the expression of the gene. *Trans*-regulatory elements for a given gene refer to proteins not structurally associated with a gene that nevertheless regulate its expression. The sections below provide examples of several constructs that help shed some light on both kinds of regulation.

5.4.3.1 *Cis*-regulation of Transcription Activity as Process Control Computing

It has been known for some time that the genome contains both genes and *cis*-regulatory elements.[67] The presence or absence of particular combinations of these regulatory elements determines the extent to which specific genes are expressed (i.e., transcribed into specific proteins). In pioneering work undertaken by Davidson et al.,[68] it was shown that *cis*-regulation could—in the case of a specific gene—be viewed as a logical process analogous to a computer program that connected various inputs to a single output determining the precise level of transcription for that gene.

In particular, Davidson and his colleagues developed a high-level computer simulation of the *cis*-regulatory system governing the expression of the *endo16* gene in the sea urchin (*endo16* is a gut-specific gene of the sea urchin embryo). In this context, the term "high-level" means a highly abstracted representation, consisting at its core of 18 lines of code. This simulation enabled them to make predictions about the effect of specific manipulations of the various regulatory factors on *endo16* transcription levels that could be tested against experiment.

Some of the inputs to the simulation were binary values. The value 1 indicated that a binding site was both present and productively occupied by the appropriate *cis*-regulatory factor. A 0 indicated that the site was mutationally destroyed or inactive because its factor was not present or was inactive. The other inputs to the simulation were continuous and varied with time, and represented outputs (protein concentrations) in other parts of the system. The output of this process in some cases was a continuous time-varying variable that regulated the extent to which the specific gene in question was transcribed.

Davidson et al. were able to confirm the predictions made by their computational model, concluding that all of the regulatory functions in question (and the resulting system properties) were encoded in the DNA sequence, and that the regulatory system described is capable of processing complex informational inputs and hence indicates the presence of a multifunctional organization of the *endo16 cis*-regulatory system.[69]

[67] For purposes of the discussion in this subsection (Section 5.4.3.1), regulation refers to *cis*-regulation.

[68] C.H. Yuh, H. Bolouri, and E.H. Davidson, "Genomic *Cis*-regulatory Logic: Experimental and Computational Analysis of a Sea Urchin Gene," *Science* 128(5):617-629, 1998. Some of this discussion is also adapted from commentary on this article: G.A. Wray, "Promoter Logic," *Science* 279(5358):1871-1872, 1998.

[69] In this context, a multifunctional organization of the regulatory system means that the protein associated with the *endo16* gene is differentially expressed in various cells in the sea urchin.

The computational model undoubtedly provides a compact representation of the relationships between different inputs and different outputs.[70] Perhaps a more interesting question, however, is the extent to which it is meaningful to ascribe a computational function to the biochemical substrate underlying the regulatory system. Davidson et al. argue that the DNA sequence in this case specifies "what is essentially a hard-wired, analog computational device," resulting in system properties that are "all explicitly specified in the genomic DNA sequence."[71]

It is highly unlikely that the precise computational structure of *endo16*'s regulatory system will generalize to the regulatory systems of other genes. From the perspective of the biologist, the reason is clear—organisms are not designed as general-purpose devices. Indeed, the evolutionary process virtually guarantees that individualized solutions and architectures will be abundant, because specific adaptations are the rule of the day. Nevertheless, insight into the computational behavior of the *endo16 cis*-regulatory system provides a new way of looking at biological behavior.

Can the regulatory systems of some other genes be cast in similar computational terms? If and when future work demonstrates that such casting is possible, it will become increasingly meaningful to view the genome as thousands of simple computational devices operating in tandem. Davidson's work suggests the possibility that a class of regulatory mechanisms, complex though they might be with respect to their behavior, may be governed by what are in essence hard-wired devices whose essential functionality can be understood in computational terms through a logic of operation that is in fact relatively simple at its core. Prior to Davidson's work and despite extensive research, the literature had not revealed any apparent regularity in the organization of regulatory elements or in the ways in which they interact to regulate gene expression.

Indeed, while many promoters appear either to have a simpler organization or to operate less logically than that of *endo16*, few promoters have been examined with the many precise quantitative assays that were carried out by Davidson et al., and nonquantitative assays would have completely missed most of the functions that the majority of the regulatory system's elements encode.[72] So, it is at this point an open question whether this computational view has applicability beyond the specific case of *endo16*.

5.4.3.2 Genetic Regulatory Networks as Finite-state Automata

Trans-regulation (as contrasted to *cis*-regulation) is based on the notion that some genes can have regulatory effects on others.[73] In reality, the network of connections between genes that regulate and genes that are regulated is highly complex. In an attempt to gain insight into genetic regulatory networks from a gross oversimplification, Kaufmann proposed that actual genetic regulatory networks might be modeled as randomly connected Boolean networks.[74]

Kaufmann's model made several simplifying assumptions:

[70] E.F. Keller, *Making Sense of Life: Explaining Biological Development with Models, Metaphors, and Machines*, Harvard University Press, Cambridge, MA, 2002, p. 241.

[71] This is not to argue that DNA sequence alone is responsible for the specification of system properties. Epigenetic control mechanisms also influence system properties as do environmental conditions and cell state that are not specified in DNA. An analogy might be that although a memory dump of a computer specifies the state of the computer, many contingent activities may affect the actual execution path. For example, the behavior (and timing) of specific input-output activities are likely to be relevant.

[72] G.A. Wray, "Promoter Logic," *Science* 279(5358):1871-1872, 1998.

[73] For purposes of the discussion in this subsection (Section 5.4.3.2), regulation refers to *trans*-regulation.

[74] Much of this work is due to the pioneering work of Stuart Kauffman. See for example, S.A. Kauffman, *The Origins of Order: Self-Organization and Selection in Evolution*, Oxford University Press, New York, 1993. An alternative discussion of this material can be found at http://www.smi.stanford.edu/projects/helix/bmi214/ (May 13); lecture notes of Russell Altman.

- The total number of genes involved is N, a number of order 30,000.
- The number of genes that regulate a given target is a constant (call it K) for all regulated genes; K is a small integer.
- The regulatory signal associated with a connection or the expression of a gene is either on or off. (In fact, almost certainly it is not just the fact of a connection between genes that influences regulation, but rather the nature of that connection as a continuous time-varying value such as a molecular concentration over time.)
- Every gene is governed by the same transition rule (i.e., a Boolean function) that specifies its state (on or off) as a function of the activities of its K inputs at the immediately earlier time.
- The regulatory network operates synchronously (and, by implication, kinetics are unimportant).
- Secondary effects on genetic regulation arising from the nondigital characteristics of DNA (such as methylation) can be neglected.
- The genes that regulate and genes that are regulated (which may overlap) are connected at random.

Box 5.9 provides more details about this model. Because the model treats all genes as identical (i.e., all obey the same transition rule) and assigns connections between genes at random, it obviously lacks fidelity to any specific genome and cannot predict the biological phenomenology of any specific organism. Yet, it may provide insight into biological order that emerges from the structure of the genetic regulatory network itself.

Simulations of the operation of this model yielded interesting behavior, which depends on the values of N and K. For $K = 1$ or $K > 5$, the behavior of the network exhibits little interesting order, where order is defined in terms of fixed cycles known as attractors. If $K = 1$, the networks are static, with the number of attractors exponential in the size of the network and the cycle length approaching unity. If $K > 5$, there are few attractors, and it is the cycle length that is exponential in the size of the network. However, for $K = 2$, the network does exhibit order that has potential biological significance—both the number of attractors and the cycle length are proportional to $N^{1/2}$.[75]

What might be the biological significance of these results?

- The trajectory of an attractor through its successive states would reflect the fact that, over time, different genes are expressed in a biological organism.
- The fact that there are multiple attractors within the same genome suggests that multiple biological structures might exist, even within the same organism, corresponding to the genome being in one of these attractor states. An obvious candidate for such structures would be multiple cell types. That is, this analysis suggests that a cell type corresponds to a given state cycle attractor, and the different attractors to the different cell types of the organism. Another possibility is that different but similar attractors correspond to cells in different states (e.g., disease state, resting state, perturbed state).
- The fact that an attractor is cyclic suggests that it may be related to cyclic behavior in a biological organism. If cell types can be identified with attractors, the cyclic trajectory in phase space of an attractor may correspond to the cell cycle in which a cell divides.
- States that can be moved from one trajectory (for one attractor) to another trajectory (and another attractor) by changing a single state variable are not robust and may represent the phenomenon that small, apparently minor perturbations to a cell's environment may kick it into a different state.
- The square root of the number of genes in the human genome (around 30,000) is 173. Under the assumption of $K = 2$ scaling, this would correspond to the number of cyclic attractors and thus to the number of cell types in the human body. This is not far from the number of cell types actually observed

[75] A. Bhattacharjya and S. Liang, "Power-Law Distributions in Some Random Boolean Networks," *Physical Review Letters* 77(8):1644, 1996.

Box 5.9
Finite-state Automata and a Comparison of Genetic Networks and Boolean Networks

In Kaufmann's Boolean representation of a genetic regulatory network, there are *N* genes, each with two states of activity (expressed or inhibited), and hence 2^N possible states (i.e., sets of activities) in the network. The number of possible connections is combinatorial in *N* and *K*. Starting at time *t*, each gene makes a transition to a new state at time $t + 1$ in accord with the transition rule and the *K* inputs that it receives. Thus, the state of the network at a time $t + 1$ is uniquely determined from its state at time *t*. The trajectory of the network as *t* changes (i.e., the sequence of states that the network assumes) is analogous to the process by which genes are expressed.

This network is an instantiation of a finite-state automaton. Since there are a finite number of states (2^N), the system must eventually find itself in a state previously encountered. Since the system is deterministic, the network then cycles repeatedly through a fixed cycle, called an attractor. Every possible system state either leads to some attractor or is part of an attractor.

Different initial conditions may or may not lead to different attractors. All of the initial conditions that lead to the same attractor constitute what is known as a "basin" for that attractor. Any state within a basin can be exchanged with any other state in the same basin without changing the behavior of the network in the long run. In addition, given a set of attractors, no attractor can intersect with another (i.e., pass through even one state that is contained in another attractor). Thus, attractors are intrinsically stable and are analogous to the genetic expression pattern in a mature cell.

An attractor may be static or dynamic. A static attractor involves a cycle length of one (i.e., the automaton never changes state). A dynamic attractor has a cycle length greater than one (i.e., a sequence of states repeats after some finite number of time increments). Attractors that have extremely long cycle lengths are regarded as chaotic (i.e., they do not repeat in any amount of time that would be biologically interesting).

Two system states differing in only a small number of state variables (i.e., having only a few bits that differ out of the entire set of *N* variables) often lie on dynamical trajectories that converge closer to one another in state space. In other words, their attractors are robust under small perturbations. However, there can be states within a basin of attraction that differ in only one state variable from a trajectory that can lead to a different attractor.

(about 200). Such a result may be numerological coincidence or rooted in the fact that nearly all cells in a given organism (even across eukaryotes) share the same basic housekeeping mechanisms (metabolism, cell-cycle control, cytoskeletal construction and deconstruction, and so on), or it may reflect phenotypic structure driven by the large-scale connectivity in the overall genetic regulatory network. More work will be needed to investigate these possibilities.[76] Box 5.10 provides one view on experimental work that might be relevant.

To illustrate the potential value of Boolean networks as a model for genetic regulatory networks, consider their application to understanding the etiology of cancer.[77] Specifically, cancer is

[76] This point is discussed further in Section 5.4.2.2 and the references therein.

[77] Z. Szallasi and S. Liang, "Modeling the Normal and Neoplastic Cell Cycle with 'Realistic Boolean Genetic Networks': Their Application for Understanding Carcinogenesis and Assessing Therapeutic Strategies," *Pacific Symposium on Biocomputing*, pp. 66-76, 1998.

Box 5.10
Testing the Potential Relevance of the Boolean Network Model

Because of the extreme simplifications embedded in the Boolean network model, detailed predictions (e.g., genes A and B turn on gene C) are unlikely to be possible. Instead, the utility of this approach as a way of looking at genetic regulation will depend on its ability to make qualitative predictions about large-scale structure and trends. Put differently, can Boolean networks behave in biologically plausible ways?

Under certain circumstances, Boolean networks do exhibit certain regularities. Thus, the operative question is whether these features have reasonable biological interpretations that afford insight into the integrated behavior of the genomic system. Consider the following:

1. A large fraction of the genes in Boolean networks converge to fixed states of activity, on or off, that contain the same genes on all cell-type attractors. The existence of this "stable core" predicts that most genes will be in the same state of activity on all cell types of an organism. Direct experimental testing of this prediction is possible using DNA chip technology today.
2. Nearby states in the state space of the system typically lie on trajectories that converge on each other in state space. This might be tested by cloning exogenous promoters upstream of a modest number of randomly chosen genes to transiently activate them, or by using inhibitory RNA to transiently inactivate a gene's RNA products, and following the trajectory of gene activities in unperturbed cells over time and perturbed cells where the gene's activity is transiently altered, using DNA chips to assess whether the states of activity become more similar.
3. The Boolean model predicts that if randomly chosen genes are transiently reversed in their activity, a downstream avalanche of gene activities will ensue. The size distribution of these avalanches is predicted to be a power law, with many small avalanches and few large ones. There is a rough maximum size avalanche that scales as about three times the square root of the number of genes, hence about 500 for human cells. This is testable, again by cloning upstream controllable promoters to transiently activate random genes, or inhibitory RNA to transiently inactivate random genes, and following the resulting avalanche of changes in gene activities over time using DNA chips.
4. The Boolean model assumes cell types are attractors. As such, cell-type attractors are stable to about 95 percent of the single gene perturbations—the system returns to the attractor from which it was perturbed. Similarly, it is possible to test whether cell types are stable in the same homeostatic way by perturbing the activity of many choices of single genes, one at a time.
5. The stable core leaves behind "twinkling islands" of genes that are functionally isolated from one another. These are the subcircuits that determine differentiation, since each island has its own attractors, and the attractors of the network as a whole are unique choices of attractor from each of the twinkling islands in a kind of combinatorial epigenetic code. Current techniques can test for such islands by starting avalanches from different single genes. Two genes in the same island should have overlapping downstream members of the avalanches they set off. Genes in different islands should not. The caveat here is that there may be genes downstream from more than one island, affected by avalanches started in each.

SOURCE: Stuart Kauffman, Santa Fe Institute, personal communication, September 20, 2002.

widely believed to be a pathology of the hereditary apparatus. However, it has been clear for some time that single-cause, single-effect etiologies cannot account for all or nearly all occurrences of cancer.[78]

[78] See, for example, T. Hunter, "Oncoprotein Networks," *Cell* 88(3):333, 1997; B. Vogelstein and K.W. Kinzler, "The Multistep Nature of Cancer," *Trends in Genetics* 9(4):138, 1993. (Cited in Szallasi and Liang, 1998.)

If the correspondence between attractor and cell is assumed, malignancy can be viewed as an attractor similar in most ways to that associated with a normal cell,[79] and the transition from normal to malignant is represented by a "phase transition" from one attractor to another. Such a transition might be induced by an external event (radiation, chemical exposure, lack of nutrients, and so on).

As one illustration, Szallasi and Liang argue that changes in large-scale gene expression patterns associated with conversion to malignancy depend on the nature of attractor transition in the underlying genetic network in three ways:

1. A specific oncogene can induce changes in the state of downstream genes (i.e., genes for which the oncogene is part of their regulatory network) and transition rules for those genes without driving the system from one attractor to another one. If this is true, inhibition of the oncogene will result a reversion of those downstream changes and a consequent normal phenotype. In some cases, just such phenomenology has been suggested,[80] although whether or not this mechanism is the basis of some forms of human cancer is unknown as yet.
2. A specific oncogene could force the system to leave one attractor and flow into another one. The new attractor might have a much shorter cycle time (implying rapid cell division and reproduction) and/or be more resistant to outside perturbations (implying difficulty in killing those cells). In this case, inhibition of the oncogene would not result in reversion to a normal cellular state.
3. A set of "partial" oncogenes may force the system into a new attractor. In this case, no individual partial oncogene would induce a phenotypical change by itself—however, the phenomenology associated with a new attractor would be similar.

These different scenarios have implications for both research and therapy. From a research perspective, the operation of the second and third mechanisms implies that the network's trajectory through state space is entirely different, a fact that would impede the effectiveness of traditional methodologies that focus on one or a few regulatory pathways or oncogenes. From a therapeutic standpoint, the operation of the latter two mechanisms implies that a focus on "knocking out the causal oncogene" is not likely to be very effective.

5.4.3.3 Genetic Regulation as Circuits

Genetic networks can also be modeled as electrical circuits.[81] In some ways, the electrical circuit analogy is almost irresistible, as can be seen from a glance at any of the known regulatory pathways: the tangle of links and nodes could easily pass for a circuit diagram of Intel's latest Pentium chip. For example, McAdams and Shapiro described the regulatory network that governs the course of a λ-phage infection in *E. coli* as a circuit, and included factors such as time delays, which are critical in biological networks (gene transcription and translation are not instantaneous, for example) and indeed, in electrical networks, as well.

More generally, nature's designs for the cellular circuitry seems to draw on any number of techniques that are very familiar from engineering: "The biochemical logic in genetic regulatory circuits provides real-time regulatory control [via positive and negative feedback loops], implements a branch-

[79] S.A. Kauffman, "Differentiation of Malignant to Benign Cells," *Journal of Theoretical Biology* 31:429, 1971. (Cited in Szallasi and Liang, 1998.)

[80] S. Baasner, H. von Melchner, T. Klenner, P. Hilgard, and T. Beckers, "Reversible Tumorigenesis in Mice by Conditional Expression of the HER2/c-erbB2 Receptor Tyrosine Kinase," *Oncogene* 13(5):901, 1996. (Cited in Szallasi and Liang, 1998.)

[81] H.H. McAdams and L. Shapiro, "Circuit Simulation of Genetic Networks," *Science* 269(5224):650-656, 1995.

TABLE 5.2 Points of Similarity Between Genetic Logic and Electronic Digital Logic in Computer Chips

Characteristic	Electronic Logic	Genetic Logic
Signals	Electron concentrations	Protein concentrations
distribution	Point-to-point (by wires or by electrically encoded addresses)	Distributed volumetrically by diffusion or compartment-to-compartment by active transport mechanisms
Organization	Hierarchical	Hierarchical
logic type	Digital, clocked, sequential logic	Analog, unclocked (can approximate asynchronous sequential logic; dependent on relative timing of signals)
Noise	Inherent noise due to discrete electron events and environmental effects	Inherent noise due to discrete chemical events and environmental effects
Signal-to-noise ratio	Signal-to-noise ratio high in most circuits	Signal-to-noise ratio low in most circuits
Switching speed	Fast ($>10^{-9}$ s^{-1})	Slow ($<10^{-2}$ s^{-1})

SOURCE: Excerpted with permission from H. McAdams and A. Arkin, "Simulation of Prokaryotic Genetic Circuits," *Annual Review of Biophysics and Biomolecular Structure* 27:199-224, 1998, available at http://caulo.stanford.edu/usr/hm/pdf/1998_McAdams_simulation_genetic_circuits.pdf. Originally published by *Annual Review of Biophysics and Biomolecular Structure.*

ing decision logic, and executes stored programs [in the DNA] that guide cellular differentiation extending over many cell generations."[82] Table 5.2 describes some of the similarities.

Of course, taking an engineering view of biological circuits does not make understanding them trivial. For example, consider that cellular regulatory circuits implement a complex adaptive control system. Understanding this system is greatly complicated by the fact that at the biochemical implementation level, the distinction between the controlling mechanisms and the controlled processes is not as clear as it is when such control is engineered into a human-designed artifact. In a biochemical environment, control reactions and controlled functions are composed of intermingled molecules interacting in ways that make identification of roles much more complex.

Nor does the analogy to electrical circuits always carry over perfectly. Because critical molecules are often present in the cell in extremely small quantities, to take the most notable example, certain critical reactions are subject to large statistical fluctuations, meaning that they proceed in fits and starts, much more erratically than their electrical counterparts.

5.4.3.4 Combinatorial Synthesis of Genetic Networks[83]

Guet et al. have demonstrated the feasibility of creating synthetic networks, composed of well-characterized genetic elements, that provide a framework for understanding how diverse phenotypi-

[82] H.H. McAdams and A. Arkin, "Simulation of Prokaryotic Genetic Circuits," *Annual Reviews of Biophysical and Biomolecular Structure* 27:199-224, 1998.

[83] Section 5.4.3.4 is based on C.C. Guet, M.B. Elowitz, W. Hsing, and S. Leibler, "Combinatorial Synthesis of Genetic Networks," *Science* 296(5572):1466-1470, 2002.

cal functionality can (but does not always) arise from changes in network topology rather than changes in the elements themselves. This functionality includes networks that exhibit the behavior associated with negative and positive feedback loops, oscillators, and toggle switches. By showing that functionality can change dramatically due to changes in topology, Guet et al. argue that once a simple set of genes and regulatory elements is in place, it is possible to jump discontinuously from one functional phenotype to another using the same "toolkit" of genes simply by modifying the regulatory connections. Such discontinuous changes are different from the more gradual effects driven by successive point mutations.

Such discontinuities reflect the nonlinear nature of genetic networks. Furthermore, the topology of connectivity of a network does not necessarily determine its behavior uniquely, and the behavior of even simple networks built out of a few well-characterized components cannot always be inferred from connectivity diagrams alone. Because genetic networks are nonlinear (and stochastic as well), the unknown details of interactions between components might be of crucial importance to understanding their functions. Combinatorially developed libraries of simple networks may thus be useful in uncovering the existence of additional regulatory mechanisms and exploring the limits of quantitative modeling of cellular systems.

The system of Guet et al. uses a small number of elements restricted to a single type of interaction (transcriptional regulation), but the range of biochemical interactions can be extended by including other modular genetic elements. For example, the approach can be extended to include linking input and output through cell-cell signaling molecules, such as those involved in quorum sensing. Also, this combinatorial strategy can be used to search for other dynamic behaviors such as switches, sensors, oscillators, and amplifiers, as well as for high-level structural properties such as robustness or noise resistance.

5.4.3.5 Identifying Systems Responses by Combining Experimental Data with Biological Network Information

Mawuenyega et al. have developed a method to identify specific subnetworks in large biological networks.[84] A biological network is constructed by identifying components (genes, proteins, transcription factors, chemicals) and interactions between components (protein-protein, protein-DNA, signal transduction, gene expression, catalysis) from genome context information as well as from external sources (databases, literature, and direct interaction with experimentalists). By superimposing experimental data such as expression values or identified proteins, it is possible to identify a best-scored subnetwork in the large biological network. This subnetwork is known as the *response network*, identifying a system's response with respect to the experimental scenario and data used.

Proteomic mass spectroscopy (MS) analysis was used to identify and characterize 1,044 *Mycobacterium tuberculosis* (TB) proteins and their corresponding cellular locations. From these 1,044 identified, 70 proteins were selected that are known to function in lipid biosynthesis (20) and fatty acid degradation (50). It is striking that the identified proteins involved in fatty acid degradation were distributed between the different cellular compartments in an almost exclusive fashion (e.g., in the subnetwork centered on *fadB2* and *fadB3*) (Figure 5.9).

In addition, Forst and colleagues performed a response network analysis of *mycobacterium tuberculosis* to isoniazid (INH) drug treatment.[85] The entirety of the FAS-II fatty acid synthase group (except

[84] K.G. Mawuenyega, C.V. Forst, K.M. Dobos, J.T. Belisle, J. Chen, M.E. Bradbury, A.R. Bradbury, and X. Chen, "Mycobacterium Tuberculosis Functional Network Analysis by Global Subcellular Protein Profiling," *Molecular Biology of the Cell* 16:396-404, 2005.

[85] L. Cabusora, E. Sutton, A. Fulmer, and C.V. Forst, "Differential Network Expression During Drug and Stress Response," *Bioinformatics* 21:2898-2905, 2005, available at http://bioinformatics.oupjournals.org/cgi/content/abstract/bti440v1.

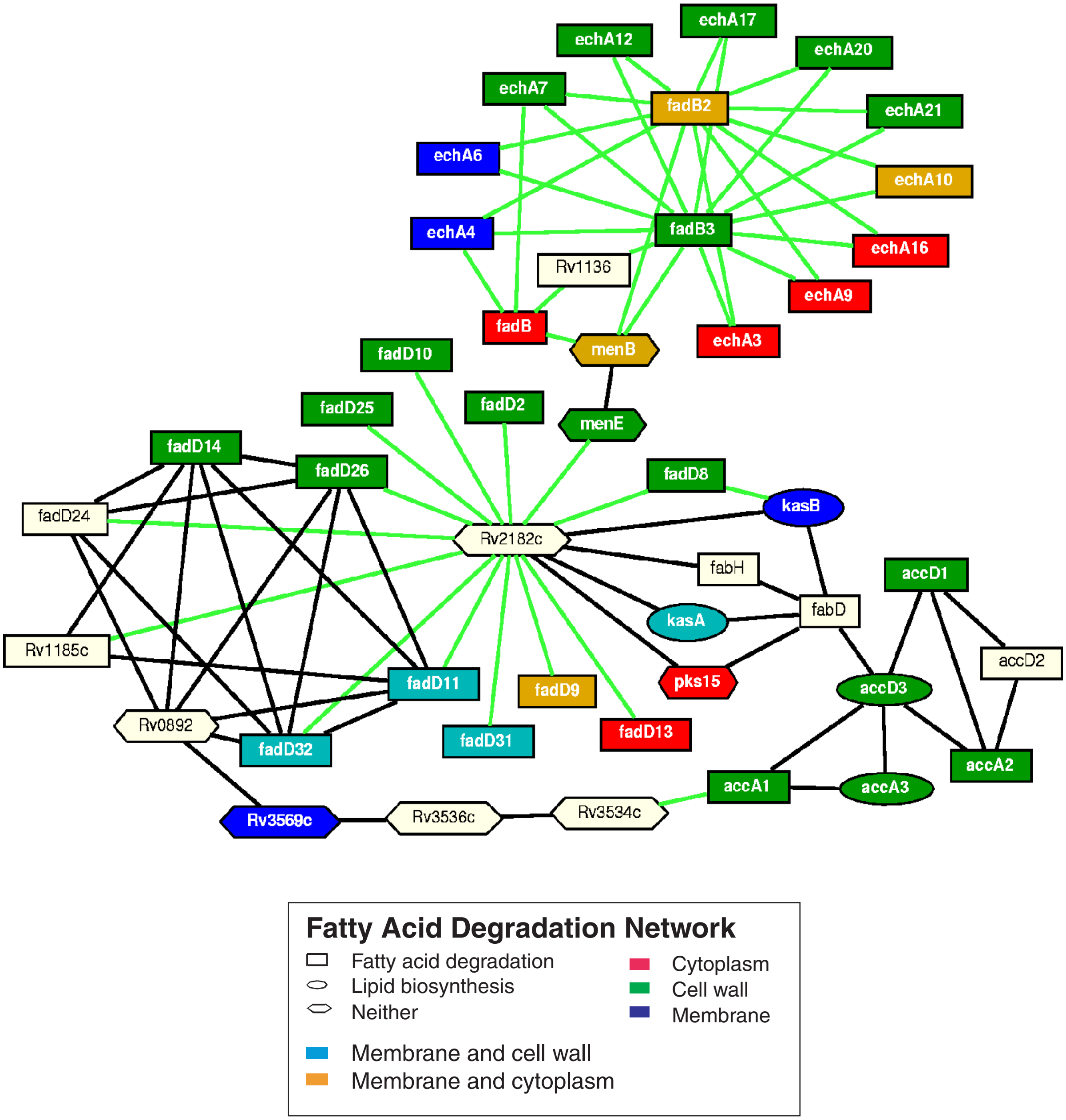


FIGURE 5.9 Fatty acid degradation network. SOURCE: Courtesy of Christian Forst, Los Alamos National Laboratories, December 8, 2004.

acpM, which was not included in the interaction data used to construct the original, whole network) showed up in the INH response network, all with significant up-regulation (Figure 5.10). Furthermore, the specific removal of these genes (*kasA*, *kasB*, *fabD*, *accD6*) from the initial set of genes did not affect their presence in the INH response subnetwork: the newly calculated network continued to contain each of them. Forst concluded that INH does directly interfere with the FAS-II fatty acid production pathway, in confirmation of earlier results.

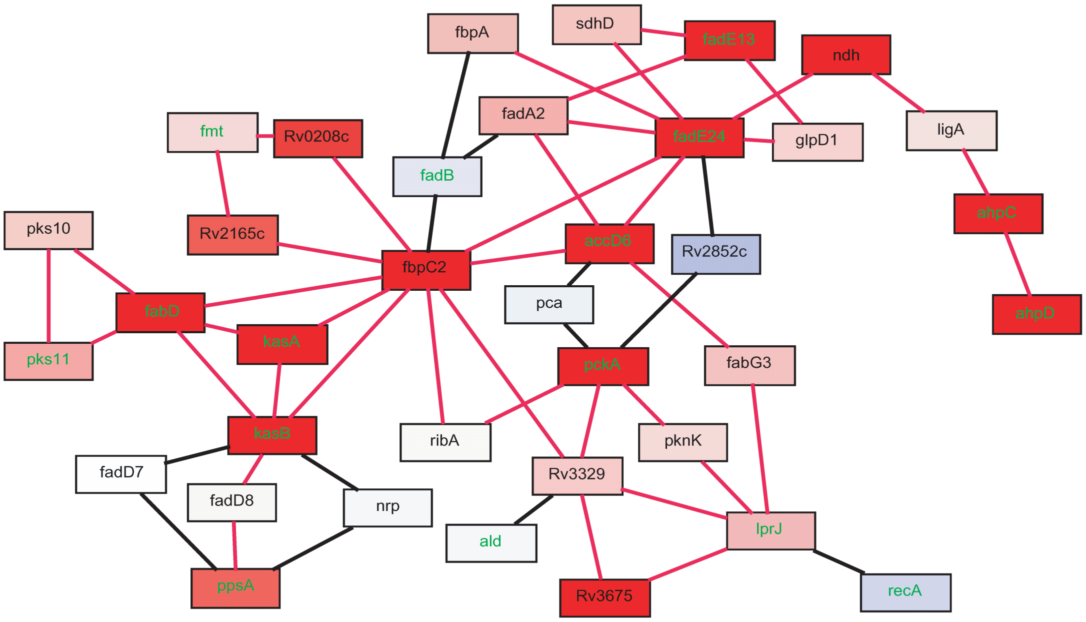


FIGURE 5.10 The isoniazid (INH) response network. Red nodes indicate up-regulated genes. Blue nodes indicate down-regulated genes. SOURCE: Courtesy of Christian Forst, Los Alamos National Laboratories, December 8, 2004.

5.4.4 Organ Physiology

Sydney Brenner has noted that "genes can only specify the properties of the proteins they code for, and any integrative properties of the system must be 'computed' by their interactions."[86] In this context, subcellular behavior and function represents a first level of "computed" interaction; cellular behavior and function, a second level. Organization of cells into organs provides a context for cellular behavior, and in the words of Denis Noble, "successful physiological analysis requires an understanding of the functional interactions between the key components of cells, organs, and systems, as well as how these interactions change in disease states. This information resides neither in the genome nor even in the individual proteins that genes code for. It lies at the level of protein interactions within the context of subcellular, cellular, tissue, organ, and system structures. There is therefore no alternative to copying nature and computing these interactions to determine the logic of healthy and diseased states. The rapid growth in biological databases; models of cells, tissues, and organs; and the development of powerful computing hardware and algorithms have made it possible to explore functionality in a quantitative manner all the way from the level of genes to the physiological function of whole organs and regulatory systems."[87]

5.4.4.1 Multiscale Physiological Modeling[88]

Physiological modeling is the modeling of biological units at a level of aggregation larger than that of an individual cell. Biological units can be successively decomposed into subunits (e.g., an organism may consist of subsystems for circulatory, pulmonary, digestive, and cognitive function; a digestive

[86] S. Brenner, "Biological Computation," *The Limits of Reductionism in Biology*. Wiley, Chichester, UK, 1998, pp. 106-116.

[87] D. Noble, "Modeling the Heart—from Genes to Cells to the Whole Organ," *Science* 295(5560):1678-1682, 2002.

[88] Much of the material in Section 5.4.4.1 is based on excerpts from A.D. McCulloch and G. Huber, "Integrative Biological Modelling in Silico," *Novartis Foundation Symposium* 247:4-19, 2002.

system may consist of esophagus, stomach, and intestines; and so on down to the level of organelles within cells and molecular functions within organelles), and every unit depends on the coordinated interaction of its subunits.

Given the complexity of physiological modeling, it makes sense to replicate this natural organization. Thus, models of tissue, organs, and even entire organisms are relevant subjects of physiological modeling. Functional behavior in each of these entities depends on activity at all spatial and temporal scales associated with structure from protein to cell to tissue to organ to whole organism (Box 5.11) and requires the integration of interacting physiological processes such as regulation, growth, signaling, metabolism, excitation, contraction, and transport processes. One term sometimes used for work that involves such integration is "physiome" (or by analogy to genomics,"physiomics").[89]

Integration of such models presents many intellectual challenges. Following McCulloch and Huber,[90] it is helpful to consider two different types of integration. *Structural integration* implies integration across physical scales of biological organization from protein to whole organism, while *functional integration* refers to the integrated representation of interacting physiological processes. Structurally integrative models (e.g., models of molecular dynamics and other strategies that predict protein function from structure) are driven by first principles and hence tend to be computation-intensive. Because they are based on first principles, they impose constraints on the space of possible organismic models. Functionally integrative models are strongly data-driven and therefore data-intensive, and are needed to bridge the multiple time and space scales of substructures within an organism without leaving the problem computationally intractable. Box 5.12 provides a number of examples of intersection between structurally and functionally integrated models.

Predictive simulations of subcomponents at various levels of the hierarchy of complexity are generally based on physicochemical first principles. Integrating such simulations, of which micromechanical tissue models and molecular dynamics models are examples, with each other across scales of biological organization is highly computationally intensive (and requires a computational infrastructure that enables distributed and heterogeneous computational resources to participate in the integration and facilitates the modular addition of new models and levels of organization).

5.4.4.2 Hematology (Leukemia)

Childhood acute lymphoblastic leukemia (ALL) is a lethal but highly treatable disease. However, successful treatment depends on the ability to deliver the correct intensity of therapy. Improper intensity can result in an excess of deaths caused by toxicity, decreased mental function over the long term, and undertreatment for high-risk cases.

The appropriate intensity is determined today through an extensive—and expensive—range of procedures including morphology, immunophenotyping, cytogenetics, and molecular diagnostics. However, Limsoon Wong has developed a relatively inexpensive single-platform microarray test that uses gene expression profiling to identify each of the known clinically important subgroups of childhood ALL (Figure 5.11) and hence the appropriate intensity of treatment.[91] This is confirmed using computer-assisted supervised learning algorithms, in which an overall diagnostic accuracy of 96 percent was achieved in a blinded test sample. To determine whether expression profiling at diagnosis

[89] J.B. Bassingthwaighte, "Toward Modeling the Human Physionome," pp. 331-339 in *Molecular and Subcellular Cardiology: Effects on Structure and Function*, S. Sideman and R. Beyar, eds., Plenum Press, New York, Volume 382 in Advanced Experiments in Medical Biology, 1995; http://www.physiome.org/.

[90] A.D. McCulloch and G. Huber, "Integrative Biological Modelling in Silico," pp. 4-25 in *'In Silico' Simulation of Biological Processes No. 247,* Novartis Foundation Symposium, G. Bock and J.A. Goode, eds., John Wiley & Sons Ltd., Chichester, UK, 2002.

[91] L. Wong, "Diagnosis of Childhood Acute Lymphoblastic Leukemia and Optimization of Risk-Benefit Ratio of Therapy," PowerPoint presentation presented at the Institute for Infocomm Research, 2003, Singapore, available at http://sdmc.lit.org.sg:8080/~limsoon/psZ/wls-aasbi03.ppt.

might further help identify those patients who are likely to relapse up to 4 years later, the expression profiles of four groups of leukemic samples with different outcomes were compared. Distinct gene expression profiles for each of these groups were identified.

5.4.4.3 Immunology

The immune system provides protection for human beings from pathogens. (For purposes of this discussion, the immune system of interest here refers to the *adaptive* immune system. The human body also has an innate immune system that provides a first response to pathogens that is essentially independent of the specific pathogen—in essence, its role is to give the adaptive immune system time to build a more specific response.) To do so, the immune system must first identify an entity within the body as a harmful pathogen that it should attack or eliminate and then mount a response that does so.

In principle, the identification of harmful pathogens might be based on a list of known pathogens. If an entity is found within the human body that is sufficiently similar to a known pathogen, it could be marked for later attack and destruction. However, a list-based approach to pathogen identification suffers from two major weaknesses. First, any such list would have to be large enough to include most of the possible pathogens that an organism might encounter in its lifetime; some estimates of the number of different foreign molecules that the human immune system is capable of recognizing are as high as 10^{16}.[92] Second, because pathogens evolve (and, thus, new pathogens are created), an a priori list could never be complete.

Accordingly, nature has developed an alternative mechanism for pathogen identification based on the notion of "self" versus "nonself." In this paradigm, entities or substances that are recognized as self are deemed harmless, while those that are nonself are regarded as potentially dangerous. Thus, the immune system has developed a variety of mechanisms to differentiate between these two categories. Note that this distinction is highly simplistic, as not all nonself entities are bad for the human body (e.g., transplanted organs that replace original organs damaged beyond repair). Nevertheless, the self-nonself distinction is not a bad point of departure for understanding the human immune system.

The immune system relies on a process that generates detectors for a subset of possible pathogens and constantly turns over those detectors for new detectors capable of identifying a different set of pathogens. When the immune system identifies a pathogen, it selects one of several immunological mechanisms (e.g., those associated with the different immunoglobulin [Ig] groups) to eliminate it. Furthermore, the immune system retains memory of the pathogen, in the form of detectors that are specifically configured for high affinity to that pathogen. Such memory enables the immune system to confer long-lasting resistance (immunity) to pathogens that may be encountered in the future and to mount a stronger response to such future encounters.

Many of the broad outlines of the immune system are believed to be understood, and computational modeling of the immune system has shed important light on its detailed workings, as described in Box 5.13. A medical application of simulation models in immunology has been to evaluate the effects of revaccinating someone yearly for influenza. Because of the phenomenon of immune memory, a vaccine that is too similar to a prior year's vaccine will be eliminated rapidly by the immune response (a negative interference effect). A simulation model by Smith et al. has examined this effect and suggests some circumstances under which individuals who are vaccinated annually will have greater or less protection than those with a first-time vaccination.[93] The Smith et al. results also suggested that in the production of flu vaccine, a choice among otherwise equivalent strains (i.e., strains thought to be

[92] J. Inman, "The Antibody Combining Region—Speculations on the Hypothesis of General Multispecificity," *Theoretical Immunology*, G. Bell, ed., Dekker, New York, 1978.

[93] D.J. Smith, S. Forrest, D.H. Ackley, and A.S. Perelson, "Variable Efficacy of Repeated Annual Influenza Vaccination," *Proceedings of the National Academy of Sciences* 96(24):14001-14006, 1999.

Box 5.11
Levels of Biological Organization

One helpful approach is to consider a set of different, but interrelated, levels of biological organization:

- *Organ system,* in which the entire organ can be represented by a lumped-parameter systems model that can be used to predict the gross behavior of the organ. In the case of the heart, one model can be based on the notion of arterial impedance, which can be used to generate the dynamic pressure boundary conditions acting on the cardiac chambers.
- *Whole organ continuum,* in which the physical behavior and dynamical responses of the organ can be calculated from finite element methods that solve the continuum equations for the mechanics of the organ. In the case of the heart, 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 fiber orientations have been used to represent the detailed structure of the whole organ with submillimeter resolution.[1]
- *Tissue,* in which constitutive laws for the continuum models are evaluated at each point in the whole organ continuum model and obtained by homogenizing the results of multicellular network models. That is, homogenization theory can be used to re-parameterize the results of a micromechanical analysis into a form suitable for continuum-scale stress analysis. In the case of tissue mechanics for the heart, the basic functional units of tissue are represented, such as laminar myocardial sheets as ensembles of cell and matrix micromechanics models and, in some cases, the microvascular blood vessels as well.[2] A variety of approaches for these models have been used, including stochastic models based on measured statistical distributions of myofiber orientations.[3] In cardiac electrophysiology, the tissue level is typically modeled as resistively coupled networks of discrete cellular models interconnected in three dimensions.[4]
- *Single cell,* in which different types of cells are represented. As a rule, single-cell models bridge to stochastic state-transition models of macromolecular function through subcellular compartment models of representative tissue structures (e.g., the sarcomere in the case of the heart). Heart cells of different types to be modeled are representative cells from different regions of the heart, such as epicardial cells, midventricular M-cells, and endocardial cells. For mechanical models, individual myofibrils and cytoskeletal structures are modeled by lattices and networks of rods, springs, and dashpots in one, two, or three dimensions.
- *Macromolecular complex,* in which representative populations of cross-bridges or ion channels are modeled. Such complexes are typically 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,* in which single cross-bridges and ion channels are represented. Cross-bridges move according to Brownian dynamics, and it is necessary to use weighted-ensemble dynamics to allow the simulation to clear the energy barriers. (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 motion (HCM) simulation of the channel complex.) The flexibility of the cross-bridges themselves can be derived from the HCM method, and the interactions with other molecules can be computed using continuum solvent approximations.
- *Atomic model,* in which molecules are represented in terms of the positions of their constituent atoms in crystallographic structures. (Such data can be found in public repositories such as the Protein Data Bank.) Such data feed molecular dynamics simulations in order to build the HCM model.

The approach described above—of integrating models across structural and functional lines—is generally adaptable to other tissues and organs, especially those with physical functions, such as lung and cartilage.

[1] F.J. Vetterand A.D. McCulloch, "Three-dimensional Analysis of Regional Cardiac Function: A Model of Rabbit Ventricular Anatomy," *Progress in Biophysics and Molecular Biology* 69(2-3):157-183, 1998.

[2] K. May-Newman and A.D. McCulloch, "Homogenization Modelling for the Mechanics of Perfused Myocardium," *Progress in Biophysics and Molecular Biology* 69(2-3):463-481, 1998.

[3] T.P. Usyk, J.H. Omens, and A.D. McCulloch, "Regional Septal Dysfunction in a Three-dimensional Computational Model of Focal Myofiber Disarray," *American Journal of Physiology* 281(2):H506-H514, 2001.

[4] L.J. Leon and F.A. Roberge, "Directional Characteristics of Action Potential Propagation in Cardiac Muscle: A Model Study," *Circulation Research* 69: 378-395, 1991.

SOURCE: Adapted from A.D. McCulloch and G. Huber, "Integrative Biological Modelling in Silico," pp. 4-25 in *'In Silico' Simulation of Biological Processes No. 247,* Novartis Foundation Symposium, G. Bock and J.A. Goode, eds., John Wiley & Sons Ltd., Chichester, UK, 2002.

Box 5.12
Examples of Intersection Between Structurally and Functionally Integrated Models

There are a number of examples of intersection between structurally and functionally integrated models, including the following:

- Linkage of biochemical networks and spatially coupled processes, such as calcium diffusion in structurally based models of cell biophysics;[1]
- Use of physicochemical constraints to optimize genomic systems models of cell metabolism;[2]
- Integration of genomic or cellular system models into multicellular network models of memory and learning,[3] developmental pattern formation,[4] or action potential propagation;[5]
- Integration of structure-based predictions of protein function into systems models of molecular networks;
- Development of kinetic models of cell signaling and coupling them to physiological targets such as energy metabolism, ionic currents or cell motility;[6]
- Use of empirical constraints to optimize protein folding predictions;[7] and
- Integration of systems models of cell dynamics into continuum models of tissue and organ physiology.[8]

[1] L.M. Loew, "The Virtual Cell Project," *Novartis Foundation Symposium* 247:151-161, 2002; L.M. Loew and J.C. Schaff, "The Virtual Cell: A Software Environment for Computational Cell Biology," *Trends in Biotechnology* 19(10):401-406, 2001.

[2] B.O. Palsson, "What Lies Beyond Bioinformatics?" *Nature Biotechnology* 15:3-4, 1997; C.H. Schilling, J.S. Edwards, D. Letscher, and B.O. Palsson, "Combining Pathway Analysis with Flux Balance Analysis for the Comprehensive Study of Metabolic Systems," *Biotechnology and Bioengineering* 71(4):286-306, 2000-2001.

[3] D. Durstewitz, J.K. Seamans, and T.J. Sejnowski, "Neurocomputational Models of Working Memory," *Nature Neuroscience* 3(Supplement):S1184-S1191, 2000; P.H. Tiesinga, J.M. Fellous, J.V. Jose, and T.J. Sejnowski, "Information Transfer in Entrained Cortical Neurons," *Network: Computation in Neural Systems* 13(1):41-66, 2002.

[4] E.H. Davison, J.P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.H. Yuh, T. Minokawa, et al., "A Genomic Regulatory Network for Development," *Science* 295(5560):1669-1678, 2002.

[5] R.M. Shaw and Y. Rudy, "Electrophysiologic Effects of Acute Myocardial Ischemia: A Mechanistic Investigation of Action Potential Conduction and Conduction Failure," *Circulation Research* 80(1):124-138, 1997.

[6] J.M. Levin, R.C. Penland, A.T. Stamps, and C.R. Cho, "Using in Silico Biology to Facilitate Drug Development.," in *Novartis Foundation Symposium* 247: 222-238, 2002.

[7] L. Salwinski and D. Eisenberg, "Motif-based Fold Assignment," *Protein Science* 10(12):2460-2469, 2001.

[8] R.L. Winslow, D.F. Scollan, A. Holmes, C.K. Yung, J. Zhang, M.S. Jafri, "Electrophysiological Modeling of Cardiac Ventricular Function: From Cell to Organ," *Annual Reviews of Biomedical Engineering* 2: 119-155, 2002; N.P. Smith, P.J. Mulquiney, M.P. Nash, C.P. Bradley, D.P. Nickerson, and P.J. Hunter, "Mathematical Modelling of the Heart: Cell to Organ," *Chaos, Solitons and Fractals* 13:1613-1621, 2002.

SOURCE: A.D. McCulloch and G. Huber, "Integrative Biological Modelling in Silico," pp. 4-19 in *'In Silico' Simulation of Biological Processes No. 247,* Novartis Foundation Symposium, G. Bock and J.A. Goode, eds., John Wiley & Sons Ltd., Chichester, UK, 2002. Reproduced with permission from John Wiley & Sons Ltd.

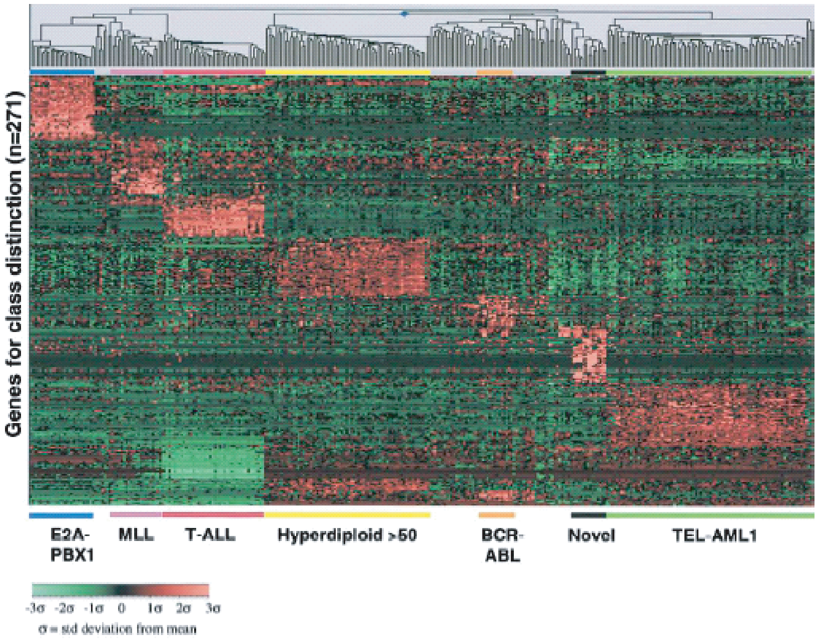


FIGURE 5.11 Microarray expression groupings indicating known clinically important subgroups of childhood acute lymphoblastic leukemia (ALL). Note in particular the second column from the right, labeled "novel." In this instance, the hierarchical clustering of gene expression reveals a novel subtype of childhood ALL. SOURCE: Courtesy of L. Wong, Institute for Infocomm Research, Singapore, 2003.

equally good guesses of the upcoming epidemic strain and equally appropriate for manufacture) should be resolved in favor of the strain that is most different from the one used in the previous year, because this choice would reduce the effects of negative interference and thus potentially increase vaccine efficacy in recipients of repeat vaccines.

5.4.4.4 The Heart

The heart is an organ of primary importance in vertebrates, and heart disease is one of the primary causes of death in the Western world. At the same time, the heart is an organ of high complexity. Although it is in essence an impulsive pump, it is a pump that must operate continuously and repair itself if necessary while in operation. Its output must be regulated according to various physiological conditions in the body, and its performance is affected by the characteristics of the arterial and vein networks to which it is connected.

The heart brings together many subsystems that interact mutually through fundamental physiological processes. As a general rule, physiological processes have both functional and structural dimensions. For example, cells are functionally specialized—blood cells and myocytes (heart cells) do different things. Furthermore, blood cells and heart cells are themselves part of a collective of other blood cells and heart cells; thus, the structure within which an individual cell is embedded is relevant.

An integrated computational model of the heart would bring together all of the relevant physiological processes (Box 5.14).[94] Were such a model available, it would be possible to investigate common

[94] A.D. McCulloch and G. Huber, "Integrative Biological Modelling in Silico," pp. 4-25 in *'In Silico' Simulation of Biological Processes No. 247,* Novartis Foundation Symposium, G. Bock and J.A. Goode, eds., John Wiley & Sons Ltd., Chichester, UK, 2002.

Box 5.13
Modeling in Immunology

In basic immunology, issues related to mutation also have been the focus of mathematical modeling and intense experimentation. . . . [For example,] during the course of an immune response, B lymphocytes within germinal centers can rapidly mutate the genes that code for antibody variable regions. The immune system thus provides an environment in which evolution occurs on a time scale of weeks. Among the large number of mutant B cells that are generated, selection chooses for survival those B cells that have increased binding affinity for the antigen that initiated the response. After 2 to 3 weeks, antibodies can have improved their equilibrium binding constant for antigen by one to two orders of magnitude, and may have sustained as many as 10 point mutations. How can the immune system generate and select variants with higher fitness this rapidly and this effectively? An optimal control model has suggested that mutation should be turned on and off episodically in order to allow new variants time to expand without being subjected to the generally deleterious effects of mutation. Time-varying mutation could be implemented by having cells recycle through one region of the germinal center, mutating while there, and proliferating in a different region of the germinal center. This suggestion has generated new experimental investigations of events that occur within germinal centers. Opportunities exist for a range of models that address basic questions about in vivo cell population dynamics and evolution, as well as more detailed questions involving the immunological mechanisms underlying affinity maturation.

Control of the immune response is another area ripe for modeling. What determines the intensity of a response? How is the response shut off when the antigen is eliminated? Feedback mechanisms may exist to control the response intensity, response length, and type of response (cellular or antibody). Some models of a basic feedback mechanism involving two types of helper T cells, T_H1 and T_H2, have been developed; others are needed. Regulatory mechanisms involve interactions among many cell populations that communicate by direct cell-cell contact and through the secretion of cytokines. Diagrams representing the elements of regulatory schemes commonly have scores of elements. Because of the complexities involved, theorists have an opportunity to lead experimentation by providing suggestions as to what needs to be measured and how such measurements can be used to provide an insightful view of possible control mechanisms.

A fundamental feature of the immune system is its diversity. Successful recognition of antigens appears to require a repertoire of at least 10^5 different lymphocyte clones. The diversity of the immune system has challenged experimentalists, and many recent advances have come from developing experimental models with limited immune diversity. However, models based on ecological concepts may provide insights into the control of clonal diversity, and modern computational methods now make it practical to consider models with tens of thousands of clones. Thus, it is possible to develop models that start to approach the size of small immune systems. Simulations have suggested that from simple rules of cell response, emergent phenomena arise that may have immunological significance. The challenge in using computation is to develop models that address important questions, are realistic enough to capture the relevant immunology, and yet are simple enough to be revealing.

SOURCE: Reprinted by permission from S.A. Levin, B. Grenfell, A. Hastings, and A.S. Perelson, "Mathematical and Computational Challenges in Population Biology and Ecosystems Science," *Science* 275(5298):334-343. Copyright 1997 AAAS. (References omitted.)

heart diseases and to probe cardiac structure and function in different places in the heart—a point of some significance in light of the fact that heart failure is usually regional and nonhomogeneous. The graphic in Box 5.14 emphasizes functional integration in the heart, and the majority of functional interactions take place at the scale of the single cell. However, an organism's behavior depends on interactions that span many orders of magnitude of space and time (from molecular structures and events to whole-organ anatomy and physiology). Thus, high-fidelity modeling of an organism or organ system within an organism demands the integration of information across similar scales.

An example of a functional model of a single cell is the work of Winslow et al. in modeling the cardiac ventricular cell, and specifically the relationship between various current flows in the cell and

Box 5.14
Computational Modeling of the Heart

. . . [Integrative cardiac modelling has sought] to integrate data and theories on the anatomy and structure, hemodynamics and metabolism, mechanics and electrophysiology, regulation and control of the normal and diseased heart. The challenge of integrating models of many aspects of such an organ system, including its structure and anatomy, biochemistry, control systems, hemodynamics, mechanics and electrophysiology, has been the theme of several workshops over the past decade or so.

Some of the major components of an integrative cardiac model that have been developed include [models of] ventricular anatomy and fiber structure, coronary network topology and hemodynamics, oxygen transport and substrate delivery, myocyte metabolism, ionic currents, impulse propagation, excitation-contraction coupling, neural control of heart rate and blood pressure, cross-bridge cycling, tissue mechanics, cardiac fluid dynamics and valve mechanics, and ventricular growth and remodelling. . . .

. . . . [T]hese models can be extended and integrated with others [by considering the role in] several major functional modules . . . as shown in the figure below. . . . 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.

. . . While [Figure 5.14.1] 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. . . . [A functionally integrated] 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.

. . . 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. . . . 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. . . .

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; biophysical models of the dynamics of cross-bridge interactions at the level of the cardiac contractile filaments; whole-cell biophysical models of the regulation of muscle contraction; microstructural constitutive models of the mechanics of multicellular tissue units; continuum models of myocardial tissue mechanics and electrical impulse propagation; and anatomically detailed whole organ models.

They have also investigated methods to bridge some of the boundaries between the different levels of organization. We [McCulloch and Huber] and others have developed finite-element models of the whole heart, incorporating microstructural constitutive laws and the cellular biophysics of thin filament activation. 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. At the other end of the hierarchy, Huber 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.

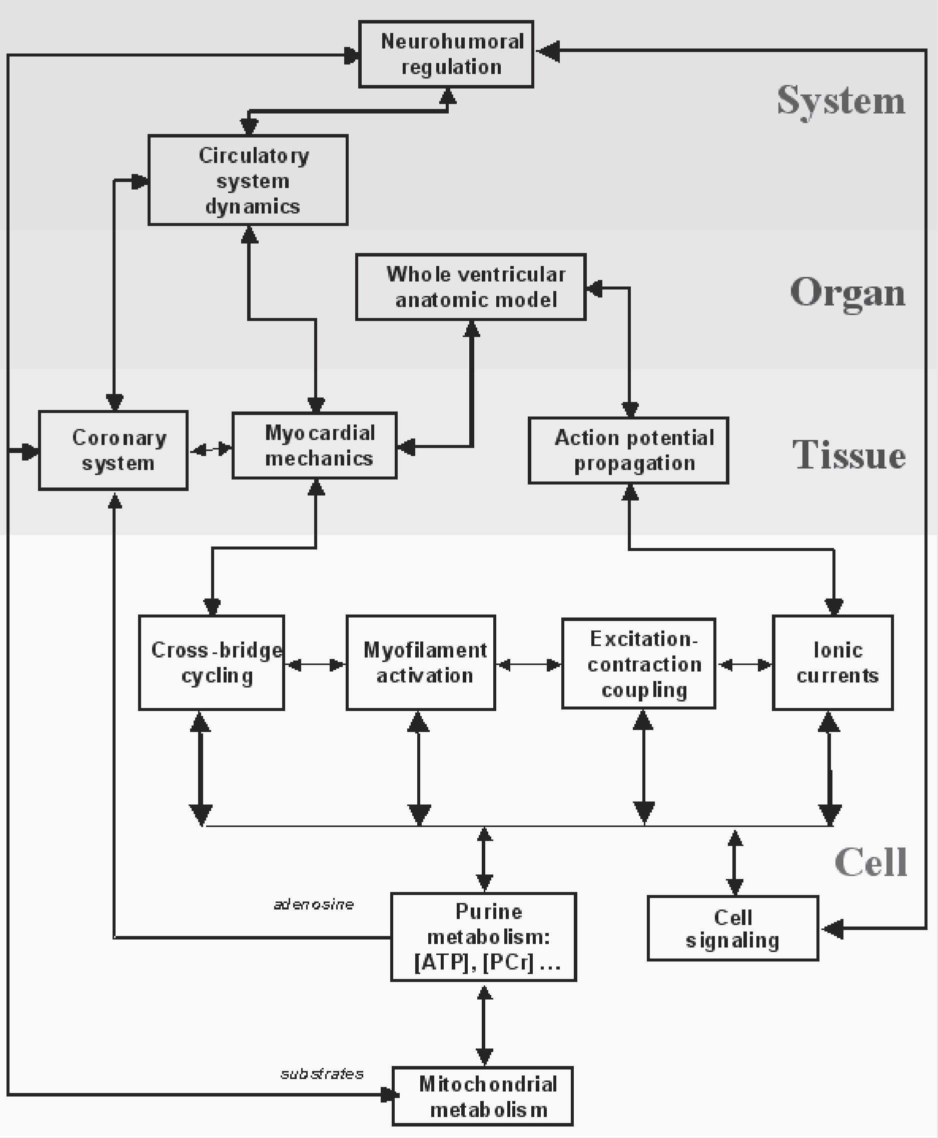


FIGURE 5.14.1 Some major functional subsystems of an integrated heart model and their hierarchical relationships from cell to tissue to organ and cardiovascular system.

SOURCE: A.D. McCulloch and G. Huber, "Integrative Biological Modelling in Silico," pp. 4-19 in *'In Silico' Simulation of Biological Processes No. 247,* Novartis Foundation Symposium, G. Bock and J.A. Goode, eds., John Wiley & Sons Ltd., Chichester, UK, 2002. Text and figure reproduced with permission from John Wiley & Sons Ltd. (References omitted.)

Box 5.15
Illustrations of Functional Models of Cellular Behavior

Example 1: Results from Single-cell Modeling

Winslow et al. have developed and applied a model of the normal and failing canine ventricular myocyte to analysis of the functional significance of changes in gene expression during tachycardia pacing-induced heart failure. Using the data on mRNA and protein expression levels cited above, these investigators defined a minimal model of end-stage heart failure as (1) 33 percent reduction of I_{K1}; (2) 66 percent reduction of I_{to1}; (3) 68 percent reduction of the SR [sacroplasmic reticulum] Ca^{2+}-ATPase; and (4) 75 percent upregulation of the Na^{+}-Ca^{2+} exchanger. They incorporated these changes sequentially into the computational model and used the model to predict the functional consequences of each alteration of gene expression in this disease. Results show that the minimal HF [heart failure] model can reproduce the increased APD [action potential duration] observed in failing myocytes relative to normal myocytes. The minimal model can also account for the reduced amplitude and slowed relaxation of the Ca^{2+} transients observed in failing versus normal myocytes. Most importantly, model simulations reveal that reduced expression of the outward potassium currents I_{to1} and I_{K1} have relatively little impact on APD, whereas altered expression of the Ca^{2+} handling proteins has a profound impact on APD.

These results suggested a strong interplay between APD and the properties of Ca^{2+} handling in canine myocytes. The nature of this interplay was examined in the model. The model indicated that reductions in expression level of the SR Ca^{2+}-ATPase and increased expression of the Na^{+}-Ca^{2+} exchanger both contribute to a reduction of JSR Ca^{2+} load. This reduction in the junctional SR (JSR) Ca^{2+} load in turn produces a smaller Ca^{2+} release from SR, reduced subspace Ca^{2+} levels, and therefore reduced Ca^{2+}-mediated inactivation of the Ca^{2+} current. The enhanced Ca^{2+} current then contributes to prolongation of APD. This is an important insight, because identifies the heart failure-induced reduction in JSR Ca^{2+} load as a critical factor in APD prolongation and in the accompanying increased risk of arrhythmias related to repolarization abnormalities.

Analyses of the type described above are likely to become increasingly important in determining the functional role of altered gene and protein expression in various disease states as more comprehensive large-scale data on genome and protein expression in disease become available.

its contractile behavior.[95] In particular, Winslow has used this model to show that the reduced contractility (i.e., reduction in the strength with which a ventricular muscle contracts, which is associated with heart failure) is caused largely by changes in the calcium ion currents in those cells, rather than changes in potassium ion currents as was widely speculated before this work (Example 1 in Box 5.15). Such an insight suggests that the development of drugs to cope with heart failure would thus be better focused on those that can regulate calcium flow. Examples 2 and 3 in Box 5.15 illustrate some of the scientific insights that can be gained with a computational model integrated across functional and structural lines.

Integrating these various perspectives on the heart (and other organs as well) is the mission of the Physiome Project, which seeks to construct models that incorporate the detailed anatomy and tissue structure of an organ in a way that allows the inclusion of cell-based models and spatial structure and distribution of proteins. The Physiome project has developed a computational framework for integrating the electrical, mechanical, and biochemical functions of the heart:[96]

[95] R.L. Winslow, D.F. Scollan, A. Holmes, C.K. Yung, J. Zhang, and M.S. Jafri, "Electrophysiological Modeling of Cardiac Ventricular Function: From Cell to Organ," *Annual Reviews of Biomedical Engineering* 2:119-155, 2002.

[96] P.J. Hunter, "The IUPS Physiome Project: A Framework for Computational Physiology," *Progress in Biophysics and Molecular Biology* 85(2-3):551-569, 2004.

Results from Integrated Modeling (Examples 2 and 3)

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 analyze the mechanism by which this genetic defect can lead to sudden cardiac death, for example, by exploring the effects of altered repolarization on the inducibility and stability of reentrant activation patterns in the whole heart. A recent study made excellent progress in 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.

. . . [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. In this case, the precise physical mechanisms at each level remain speculative, although much detail has been elucidated recently, so an integrative model will be useful for testing various hypotheses regarding the mechanisms. The modeling approach could be based on the same integrative paradigm commonly used by experimental biologists, in which the integrated effect of a specific molecular defect or structure can be analysed using techniques such as in vivo gene targeting.

SOURCE: R.L. Winslow, D.F. Scollan, A. Holmes, C.K. Yung, J. Zhang, and M.S. Jafri, "Electrophysiological Modeling of Cardiac Ventricular Function: From Cell to Organ," *Annual Review of Biomedical Engineering* 2:119-156, 2000. Adapted by permission from *Annual Review of Biomedical Engineering*. (References and citations omitted.)

- The underlying anatomical descriptions are based on finite element techniques, and orthotropic constitutive laws based on the measured fiber-sheet structure of myocardial tissue drive the dynamics of the large deformation soft-tissue mechanics involved.
- Patterns of electrical current flow in the heart are computed using reaction-diffusion equations on a grid of deforming material points which access systems of ordinary differential equations representing the cellular processes underlying the cardiac action potential; these result in representations of the activation wavefronts that spread around the heart and initiate contraction.
- Coronary blood flow is computed based on the Navier-Stokes equations in a system of branching blood vessels embedded in the deforming myocardium and the delivery of oxygen and metabolites is coupled to the energy-dependent cellular processes.

These models of different cardiac phenomena have been been implemented with "horizontal" integration of mechanics, electrical activation and metabolism, together with "vertical" integration from cell to tissue to organ. Thus, these models can be said to deconstruct an organ into a set of (submodels for) constituent functions, with explicit feedback and connection between them represented in the overall model of the whole organ.

5.4.5 Neuroscience

In recent years, neuroscience has expanded its horizons beyond the microstructure of the brain—neurons, synapses, neurotransmitters, and the like—to focus on the brain's large-scale cognitive architecture. Drawing on dramatic advances in mapping techniques, such as functional magnetic resonance imaging (MRI) and magnetoencephalography, neuroscientists hope to give a computational account of precisely what each specialized region of the brain is doing and how it interacts with all the other active regions to produce high-level thought and behavior.

5.4.5.1 The Broad Landscape of Computational Neuroscience

Neuroscience seeks to probe the details of the brain and the mechanisms by which the nervous systems develops, is organized, processes information, and establishes mental abilities. Research in neuroscience spans many levels of organization, from atomic and molecular events on the order of one-tenth to one nanometer, up to the entire nervous system on the order of a meter or more. In addition, there are on the order of 10^{11} neurons and thousands to tens of thousands of synapses per neuron.

Information processing in the brain occurs through the interactions and spread of chemical and electrical signals both within and among neurons. Acting within the extensive but intricate architecture of the neurons and their interconnections, the mechanisms are nonlinear and span a wide range of spatial and temporal scales.[97] Understanding how the nervous system and brain work thus requires an interdisciplinary approach to the challenging multiscale integration of experimental data, computational data, and theory.

It is helpful to describe the nervous system's functional processes and their mechanisms at several different levels of detail, depending on the goal of a given effort. Table 5.3 and Figure 5.12 describe the numerous spatial and temporal scales relevant to neuroscience research, and provide some indication of the complexity of such research.

To illustrate, a low level of analysis might involve consideration of individual neurons. In this analysis, functional properties of neurons such as electronic structure, nerve cell connections (synapses), and voltage-gated ion channels are important. At a higher level, it is recognized that individual neurons connect in networks—an analysis at this level examines how individual neurons interact to form functioning circuits. The mathematics of dynamic systems and visual neuroscience are notably relevant at this level. At a still higher level, individual networks—each with its own specific architecture and information-processing capabilities—interact to form neural nets and carry out cognition, speech

TABLE 5.3 Scales of Neuroscience Research

Spatial Scale	Component
1 meter	Central nervous system
10 centimeters	Systems
1 centimeter	Maps
1 millimeter	Networks
100 microns	Neurons
1 micron	Synapses
10 angstroms	Molecules

[97] N.T. Carnevale and S. Rosenthal, "Kinetics of Diffusion in a Spherical Cell: I. No Solute Buffering," *Journal of Neuroscience Methods* 41(3):205-216, 1992.

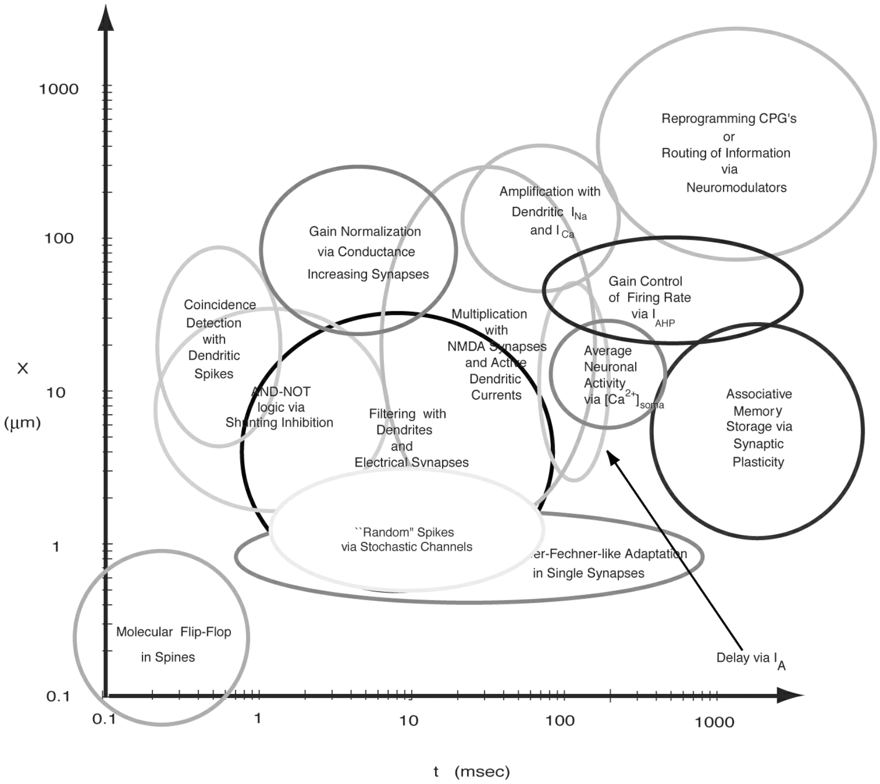

1000
100
10
1
0.1
X
(μm)
0.1
1
10
100
1000
t (msec)
Reprogramming CPG's or Routing of Information via Neuromodulators
Amplification with Dendritic I_{Na} and I_{Ca}
Gain Normalization via Conductance Increasing Synapses
Coincidence Detection with Dendritic Spikes
Multiplication with NMDA Synapses and Active Dendritic Currents
Gain Control of Firing Rate via I_{AHP}
Average Neuronal Activity via $[Ca^{2+}]_{soma}$
AND-NOT logic via Shunting Inhibition
Filtering with Dendrites and Electrical Synapses
Associative Memory Storage via Synaptic Plasticity
``Random" Spikes via Stochastic Channels
er-Fechner-like Adaptation in Single Synapses
Molecular Flip-Flop in Spines
Delay via I_A

FIGURE 5.12 Temporal and spatial scales of neuroscience research. SOURCE: Courtesy of Christof Koch, Caltech.

perception, and imaging. At this level, computational analysis of nervous system networks and (connectionist) modeling of psychological processes is the primary focus.

Computational neuroscience provides the basis for testing models of the nervous system's functional processes and their mechanisms, and computational modeling at several levels of detail is important, depending on the purposes of a given effort. Box 5.16 describes simulators that operate at different levels of detail for different purposes.

5.4.5.2 Large-scale Neural Modeling[98]

To better understand a system as complex as the human brain, it is necessary to develop techniques and tools for supporting large-scale, similarly complex simulations. Recent advances in understanding how single neurons represent the world,[99] how large populations of neurons cooperate to build more complex representations,[100] and how neurobiological systems compute functions over their representations make large-scale neural modeling a highly anticipated next step.

[98] Section 5.4.5.2 is based largely on material supplied by Chris Eliasmith, University of Waterloo, September 7, 2004.

[99] F. Rieke, D. Warland, R. de Ruyter van Steveninick, and W. Bialek, *Spikes: Exploring the Neural Code*, MIT Press, Cambridge, MA, 1997; D. Warland, M. Landolfa, J. Miller, and W. Bialek, "Reading Between the Spikes in the Cercal Filiform Hair Receptors of the Cricket," *Analysis and Modeling of Neural Systems*, F. Eeckman, ed., Kluwer Academic Publishers, Boston, MA, 1992.

[100] L. Abbott and T. Sejnowski, *Neural Codes and Distributed Representations: Foundations of Neural Computation*, MIT Press, Cambridge, MA, 1999; R.S. Zemel, P. Dayan, and A. Pouget, "Probabilistic Interpretation of Population Codes," *Neural Computation* 10, 1998.

Box 5.16
Simulators for Computational Neuroscience

The nervous system is extraordinarily complex. A single cubic centimeter in the brain's cerebral cortex contains on the order of 5 billion synapses, and these differ in size and shape. The transmission of chemical signals is very complex, with many molecules involved, and is an area of intense study. With the introduction of more powerful computer hardware and advances in algorithms, quantitative modeling and realistic simulation in three-dimensions of the interplay of biological ultrastructure and neuron physiology have become possible and have provided insight into the variability in signaling and plasticity of the system.

To deal with the complexity, multiscale range of space and time, and nonlinearity of neural phenomena, a number of specialized computational tools have been developed.

MCell (a Monte Carlo simulator of cellular microphysiology) simulates individual connections or synapses between neurons and groups of synapses. MCell simulations provide insights into the behavior and variability of real systems comprising finite numbers of molecules interacting in spatially complex environments. MCell incorporates high-resolution physical structure into models of ligand diffusion and signaling and thus can take into account the large complexity and diversity of neural tissue at the subcellular level. Monte Carlo algorithms are used to simulate ligand diffusion using three-dimensional random walk movements for individual molecules. Effector sites and surface positions are mapped spatially, and the encounters during ligand diffusion are detected. Bulk solution rate constants are converted into Monte Carlo probabilities so that the diffusing ligands can undergo stochastic chemical interactions with individual binding sites such as receptor proteins, enzymes, and transporters.

GENESIS (the General Neural Simulation System) is a tool for building structurally realistic simulations of biological neural systems that quantitatively embed what is known about the anatomical structure and phys-

Recent theoretical work has suggested that it is possible to generally characterize the dynamics, representation, and computational properties of any neural population (Figure 5.13).[101] Applications of these methods have been used successfully to generate models of working memory, rodent navigational tasks (path integration; see Figure 5.14), eye position control, representation of self-motion, lamprey and fish motor control, and deductive reasoning (Figure 5.15).

Box 5.17 illustrates the use of computational modeling to understand how dopamine functions in the prefrontal cortex. The box also illustrates the often-present tension between those who believe that simple models (in this case, advocates of a connectionist model) can provide useful insight and those who believe that simple models cannot capture the implications of the complex dynamics of individual neurons and their synapses and that the addition of considerable biophysical and physiological detail is needed for real understanding. Many of these models require large numbers of individual, spiking neurons to be simulated concurrently, which results in significant computational demands. In addition, calculating the necessary connection weights requires the inversion of extremely large matrices. Thus, high-performance computing resources are essential for expanding these simulations to include more neural tissue, and hence more complex neural function.

[101] C. Eliasmith and C.H. Anderson, *Neural Engineering: Computation, Representation and Dynamics in Neurobiological Systems*, MIT Press, Cambridge, MA, 2003.

iological characteristics of the neural system of interest. GENESIS reflects the modeling perspective that spatial organization and structure are important for understanding neural function. GENESIS is organized around neurons constructed out of components such as compartments (short sections of cellular membrane) and variable conductance ion channels that receive inputs, perform calculations on them, and then generate outputs. Neurons in turn can be linked to form neural circuits. GENESIS originally was used largely for realistic simulations of cortical networks and of the cerebellar Purkinje cell and, more recently, to interconnect cell and network properties to biochemical signaling pathways.

NEURON is similar to GENESIS in many ways, but contains optimizations that enable it to run very fast on networks in which cable properties play a crucial role, that involve system sizes ranging from parts of single cells to small numbers of cells, and that involve complex branched connections. Furthermore, the performance of NEURON degrades very slowly with increasing complexity of morphology and membrane mechanisms, and it has been applied to very large network models (10^4 cells with six compartments each and a total of 10^6 synapses in the network. Using a high-level language known as NMODL, NEURON has also been extended to investigate new kinds of membrane channels. The morphology and membrane properties of neurons are defined with an object-oriented interpreter, allowing for voltage control, manipulation of current stimuli, and other biological parameters.

SOURCES: For more information, see http://www.mcell.cnl.salk.edu; J.R. Stiles and T.M. Bartol, Jr., "Monte Carlo Methods for Simulating Realistic Synaptic Microphysiology Using MCell," pp. 87-127 in *Computational Neuroscience: Realistic Modeling for Experimentalists,* E. de Shutter, ed., Boca Raton, FL, CRC Press, 2000; J.R. Stiles, T.M. Bartol, Jr., E.E. Salpeter, M.M. Salpeter, T.J. Sejnowski, "Synaptic Variability: New Insights from Reconstructions and Monte Carlo Simulations with MCell," pp. 681-731 in *Synapses,* W.M. Cowan, T.C. Sudhof, C.F. Sudhof, eds., Johns Hopkins University Press, Baltimore, 2001; J.M. Bower, D. Beeman, and M. Hucka, "The GENESIS Simulation System," *The Handbook of Brain Theory and Neural Networks,* Second Edition, M.A. Arbib, ed., MIT Press, Cambridge, MA, 2003, pp. 475-478, available at http://www.genesis-sim.org/GENESIS/hbtn2e-bower-etal/hbtn2e-bower-etal.html; M.L. Hines and N.T. Carnevale, "The NEURON Simulation Environment," *Neural Computation* 9(6):1179-1209, 1997, available at www.neuron.yale.edu/neuron/papers/nc97/nsimenv.pdf.

5.4.5.3 Muscular Control

Muscles are controlled by action potentials—brief, rapid depolarizations of membranes in nerves and muscles. The timing of action potentials transmitted from motor neurons coordinates the contraction of the muscles they innervate. Rhythmic activity of the nervous system often takes the form of complex bursting oscillations in which intervals of action potential firing and quiescent intervals of membrane activity alternate. The relative timing of action potentials generated by different neurons is a key ingredient in the function of the nervous system.

Changes in the electrical potential of membranes are mediated by ion channels that selectively permit the flow of ions such as sodium, calcium, and potassium across the membrane. Individual channels are protein complexes containing membrane-spanning pores that open and close randomly at rates that depend on many factors. Cellular and network models of membrane potential represent these systems as electrical circuits in which voltage gated channels function as "nonlinear" resistors whose conductance depends on membrane potential. Information is transmitted from one neuron to another through synapses where action potentials trigger the release of neurotransmitters that bind to channels of adjacent cells, stimulating changes in the ionic currents of these cells. (The action potential is the basic neuronal signaling "packet" of ionic flow through a cell membrane.)

The most basic model of this mechanism is the Hodgkin-Huxley model, which refers to a set of differential equations that describe the action potential.[102] Specifically, the Hodgkin-Huxley equations

[102] A.L. Hodgkin and A.F. Huxley, "A Quantitative Description of Membrane Current and Its Application to Conduction and Excitation in Nerve," *Journal of Physiology* 117(4):500-544, 1952.

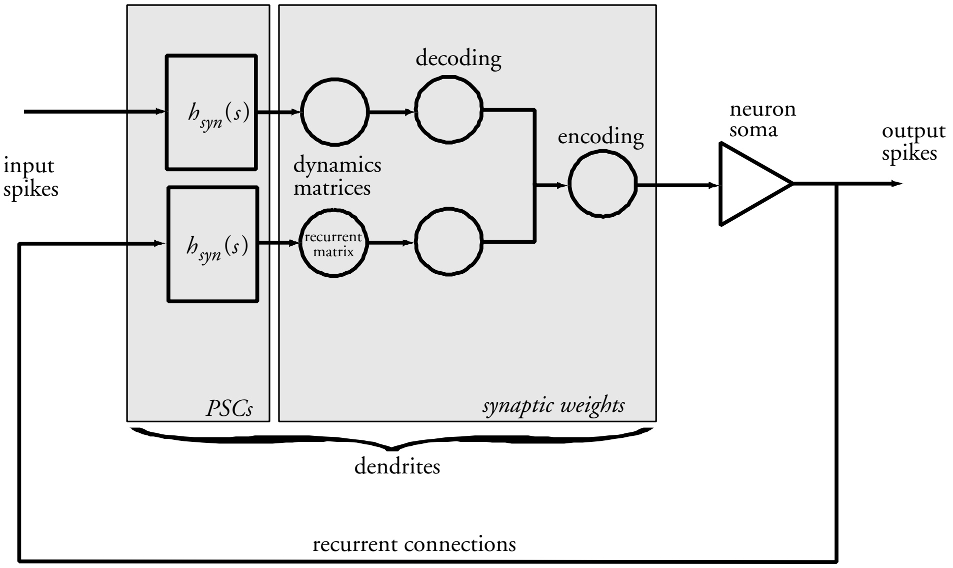


FIGURE 5.13 A generic neural subsystem. The diagram depicts the mathematical analysis of a neural subsystem and its mapping onto the biological system—a population of neurons. Labels outside the gray boxes indicate the relevant biological structures and processes. Neural action potentials (spikes) coming from a previous neural population generate weighted post-synaptic currents (PSCs) in the dendrites of the neurons to which they are connected. The subsequent voltage changes travel to the neural somata, where action potentials are generated, resulting in output spikes. Because the input and output are neural spikes, this kind of subsystem can be linked to others like it, permitting the construction of larger, more complex neural circuits (see Figure 5.15 for an example). Note that labels inside the gray boxes are generated based on understanding of the purpose of the neural system being modeled and on current understanding of neural representation (encoding), computation (decoding), and dynamics (dynamics matrices and *hsyn*). Building simulations using these methods leads to a better understanding of how neural systems perform the complex functions they do. SOURCE: Courtesy of Chris Eliasmith, University of Waterloo.

express quantitatively the feedback entailed in the relationship between changing ionic flows and changing membrane potential. Originally based on data collected from experiments on the giant axon of the squid, the physical model used is that of a membrane separating two infinite regions, each of which is homogeneous on its side of the membrane.

In the nervous system, different kinds of ions pass through the membrane, and the flow of ions through these channels is voltage dependent. In the model, a circuit is used to represent the ion flows and potential differences that drive ion flow. The semipermeable cell membrane separating the interior of the cell from the extracellular liquid is modeled as a capacitor, and each ion channel is modeled as a separately variable resistor. In series with each variable resistor is a battery representing the Nernst potential arising from the difference in ion concentration on each side of the membrane. All of these components are connected in parallel and are driven by a time-varying current source to ground. If a time-varying input current is injected into the cell, it may add further charge on the capacitor, or the added charge may leak through the channels in the cell membrane. Because of active ion transport through the cell membrane, the ion concentration inside the cell is different from that in the extracellular liquid. The potential generated by the difference in ion concentration is represented by a battery.

Elementary circuit theory allows the construction of a set of differential equations relating the different ion currents to the potential difference across the membrane. Using this set of differential equations, certain essential features of neural behavior can be modeled. For example, assuming appro-

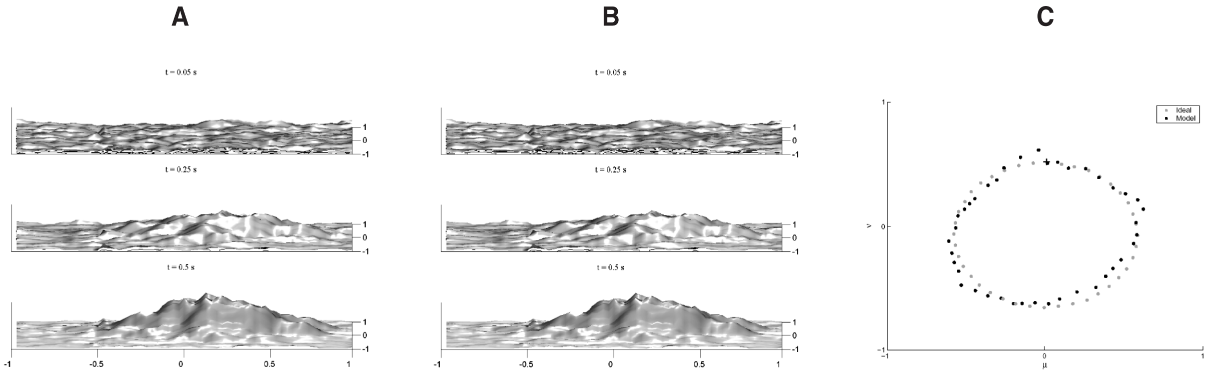


FIGURE 5.14 Rodent navigation. These figures depict the behavior of a neurally realistic simulation of the path integrator in a rat. The simulation was generated by using a single (recurrent) generic neural subsystem. (A) When the simulation is given random noise, it spontaneously generates a stable, localized bump of neural activity over the neural sheet, which represents the rat's current location. This demonstrates that a stable attractor (a widely accepted model of how the rat's path integrator is organized) has been implemented. (B) This model also implements control (i.e., updating of the current location based on the rat's motion) of the path integrator in a neurally plausible way. Here, straight-line motion in a rightward direction is shown. (C) The model correctly integrates the circular path of the rat, demonstrating that it can path integrate in any direction that the rat might move. This simulation has very little error compared to the simulations of past models. SOURCE: Chris Eliasmith, University of Waterloo, personal communication, September 7, 2004, and A. Samsonovich and B.L. McNaughton, "Path Integration and Cognitive Mapping in a Continuous Attractor Model," *Journal of Neuroscience* 17(15):5900-5920, 1997.

priate parameter values, a constant input current larger than a certain critical value and turned on at a given instant of time results in the potential difference across the membrane taking the form of a regular spike train—which is reminiscent of how a real neuron fires. More realistic current inputs (e.g., stochastic ones) result in a much more realistic-looking output.

Despite lack of information about much of the cellular and molecular basis of neuronal excitation at the time, Hodgkin and Huxley were able to provide a relatively accurate quantitative description of how an action potential was generated by voltage-dependent ionic conductivities. The Hodgkin-Huxley model provided the basis for research for more than five decades, spinning off a new field of neurophysiology: in large part, this field rests on the foundation created by their model. Recent research on membrane ion channels can be related directly to the seminal ideas and (more importantly) precise mechanism that their model described.

The "plain vanilla" Hodgkin-Huxley model is still interesting today. For example, a recent study demonstrated previously unobserved dynamics in the Hodgkin-Huxley model, namely, the existence of chaotic solutions in the model with its original parameters.[103] The significance of chaos in this context is that the excitability of a neural membrane with respect to firing is likely to be more complex than can be explained by a simple sub- or super-threshold potential.

Simulation and mathematical analysis of models have become essential tools in investigations of the complicated processes underlying rhythm generation in the nervous system. There are many types of channels and synapses. The number of channels and synapses and their locations distinguish different types of neurons from one another. Simulation of networks consisting of model neurons with

[103] J. Guckenheimer and R.A. Oliva, "Chaos in the Hodgkin-Huxley Model," *SIAM Journal on Applied Dynamical Systems* 1(1):105-114, 2002.

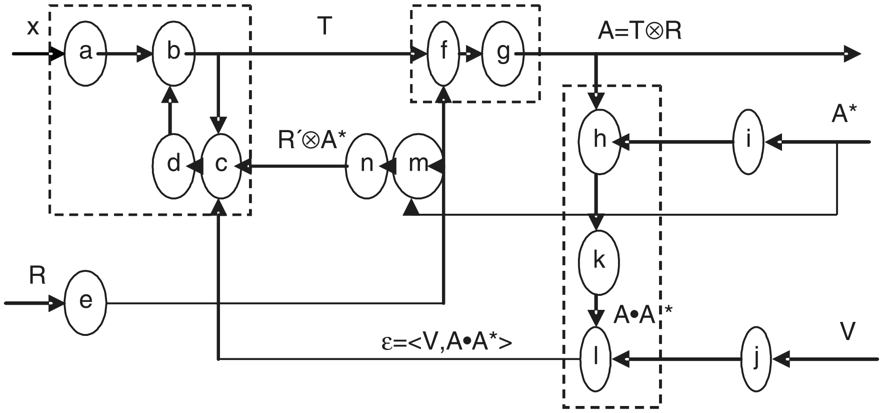


FIGURE 5.15 System for learning and performing deductive reasoning. The graphic describes the proposed system used during solution of the Wason card selection task; see P.C. Wason and P.N. Johnson-Laird, *Psychology of Reasoning: Structure and Content,* Harvard University Press, Cambridge, MA, 1972. This task requires determining when a logical rule is valid or invalid, and so is a form of deductive reasoning. Humans perform notoriously badly on many versions of this task, but well on other versions. This kind of context/content sensitivity is captured by this model; see C. Eliasmith, "Learning Context Sensitive Logical Inference in a Neurobiolobical Simulation," pp. 17-19 in *Compositional Connectionism in Cognitive Science: Papers from the AAAI Fall Symposium,* S.D. Levy and R. Gayler, Program Co-chairs, October 21-24, 2004, The AAAI Press, Arlington, VA, Technical Report FS-04-03, 2004. The depicted large-scale circuit consists of 14 neural subsystems, distributed across frontal and ventral areas of the brain. This is a good example of the degree of complexity that can be built into a neurally realistic simulation using these new techniques. Populations *a-d* learn and apply the appropriate context for interpretation of the rule (*R*) encoded by population *e*. Populations *f* and *g* apply the relevant transformation (*T*) to the rule, giving the current answer (*A*). Populations *h*, *k*, and *l* determine the degree of correctness or incorrectness of the suggested answer (either given the correct answer, or given a reward or punishment signal), resulting in an error signal e. Populations *m* and *n* provide a guess at the best possible transformation. This guess and the error signal are integrated into the learning algorithm. SOURCE: Courtesy of Chris Eliasmith, University of Waterloo.

specified conductances and synapses enables researchers to test their intuitions regarding how these networks function. Simulations also lead to predictions of the effects of neuromodulators and disorders that affect the electrical excitability of the systems. Nonetheless, simulation alone is not sufficient to determine the information we would like to extract from these models. The models have large numbers of parameters, many of which are difficult or impossible to measure, and the goal is to determine how the system behavior depends on the values of all of these parameters.

Dynamical systems theory provides a conceptual framework for characterizing rhythms. This theory explains why there are only a small number of dynamical mechanisms that initiate or terminate bursts of action potentials, and it provides the foundations for algorithms that compute parameter space maps delineating regions with different dynamical behaviors. The presence of multiple time scales is an important ingredient of this analysis because the rates at which different families of channels respond to changes in membrane potential or ligand concentration vary over several orders of magnitude.

Figure 5.16 illustrates this type of analysis using a model for bursting in the pre-Bötzinger complex, a neural network in the brain stem that controls respiration. The first panel shows voltage recordings from intracellular recordings of a medullar slice from neonatal rats. Butera and colleagues measured conductances in this preparation and constructed a model for this system.[104] Simulations of the burst-

[104] R.J. Butera, Jr., J. Rinzel, and J.C. Smith, "Models of Respiratory Rhythm Generation in the Pre-Bötzinger Complex. I. Bursting Pacemaker Neurons," *Journal of Neurophysiology* 82(1):382-397, 1999.

Box 5.17
Computational Perspectives on Dopamine Function in the Prefrontal Cortex

Connectionist Models of Dopamine Neuromodulation

A long-held hypothesis suggests that catecholamine neurotransmitters, including dopamine (DA), modulate target neuron responses, by increasing their signal-to-noise (SNR) ratio (i.e. by increasing the differentiation between background or baseline firing rates and those that are evoked by afferent stimulation). For example, studies in the striatum showed that DA potentiated the response of target neurons to the effect of both excitatory and inhibitory signals. However, the precise biophysical mechanisms underlying these effects were not well understood. Moreover, the view that DA acts as a modulator in the pre-frontal cortex (PFC) has been controversial, because, for many years, DA application or stimulation of DA neurons reliably inhibited spontaneous PFC activity. Thus, many investigators argued that DA served as an inhibitory transmitter in PFC.

The first explicit computational models of the neuromodulatory function of catecholamines, and DA in particular, were developed within the connectionist framework, and focused on their effects on information processing. Although such models do not typically incorporate biophysical detail, by virtue of their simplicity they have the advantage of simulating system level function and performance in a wide variety of cognitive tasks. Within this framework, DA effects were simulated as a change in the slope (or gain) of the sigmoidally shaped input-output activation function of processing units. Thus, in the presence of DA, both the excitatory and inhibitory influences of afferent inputs are potentiated. Computational analyses showed that this modulatory function would not improve the SNR characteristics of single neurons, but could do so at the network level. Models implementing these ideas proved useful for accounting for a wide range of phenomena, including the pharmacological effects of DA on performance in tasks thought to rely on PFC and the effects of disturbances of DA in schizophrenia.

Biophysically Detailed Models

In recent work, computational studies have focused on more biophysically detailed accounts of DA action within PFC. Models by Durstewitz et al. and Brunel and Wang, all include data on the different biophysical effects of DA on specific cellular processes. These models have been used to simulate the dynamics of activity in networks that closely parallel the patterns observed in vivo within PFC. . . .

These models synthesize the rapidly growing, but often confusing literature on the neurophysiology of DA within PFC. For example, the biophysical effects of DA are shown to produce a suppressive influence on spontaneous activity, explaining its apparent inhibitory actions, while at the same time causing an enhanced excitability in response to afferent drive. Furthermore, the selective enhancement of inputs from recurrent versus external afferents provides a mechanism for stabilizing sustained activity patterns within PFC that are resistant to interference from external inputs. These computational analyses support the characterization of DA as a modulatory neurotransmitter, rather than a classical excitatory or inhibitory one, and explain its role in support sustained activity within PFC.

Strikingly, these models are remarkably consistent with the original hypothesis that DA increases SNR within the PFC, and the expression of this idea in earlier connectionist models. The underlying assumption in both types of models is that short-term storage of information in PFC occurs through recirculating activity within local recurrent networks, which can be described as fixed-point attractor systems. DA activity helps to stabilize attractor states, both by making high activity states more stable (active maintenance), and low activity states (spontaneous background activity) less likely to spuriously transition to high activity states in the absence of strong afferent input. This is accomplished by the concurrent potentiation of excitatory and inhibitory transmission, implemented as changes in ion channel properties in biophysically detailed models and “summarized” as a change in the gain of the sigmoidal activation function in connectionist models.

These mechanisms can be used to simulate the effects of DA on performance in cognitive tasks that rely on PFC function. For example, in a task emphasizing the role of PFC in working memory, increased DA activation in the Durstewitz et al. model enhanced the stability of PFC working memory representations by making them less susceptible to interference from the intervening distractors. Within connectionist models, similar effects have been demonstrated by changing the gain of the activation function, and simulating human performance in tasks known to rely on PFC, tasks similar to those simulated by Durstewitz et al. and Brunel and Wang.

SOURCE: Reprinted by permission from J.D. Cohen, T.S. Braver, and J.W. Brown, “Computational Perspectives on Dopamine Function in Prefrontal Cortex,” *Current Opinion in Neurobiology* 12(2):223-229. Copyright 2002 Elsevier. (References omitted.)

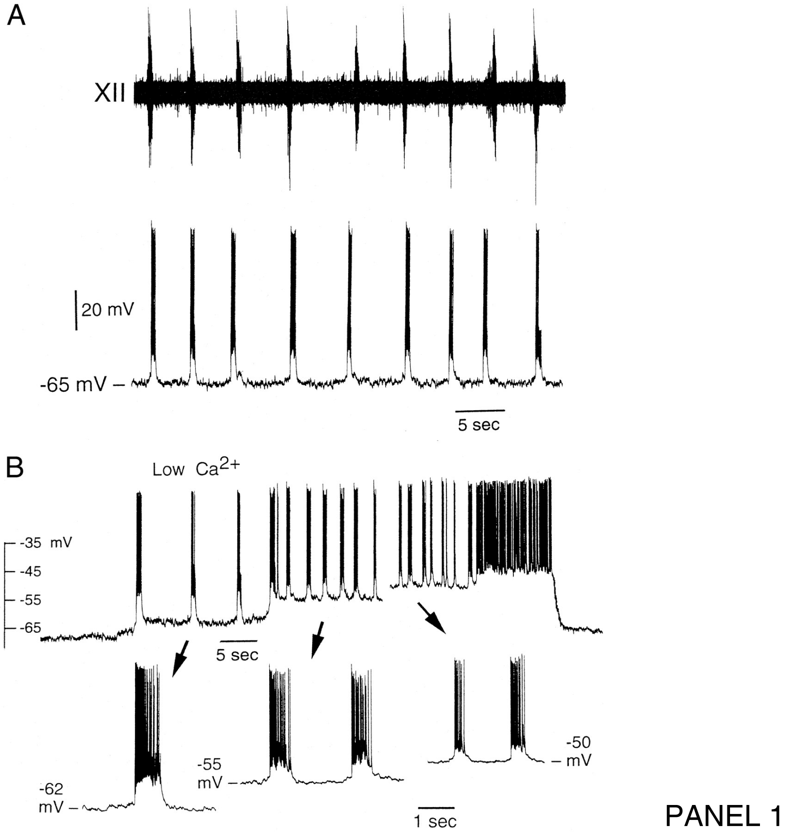


FIGURE 5.16 Bursting in the pre-Botzinger complex.

Panel 1: Example of voltage-dependent properties of pre-Bötzinger complex (pre-BötC) inspiratory bursting neurons. Traces show whole-cell patch-clamp recordings from a single candidate pacemaker neuron in the pre-BötC of a 400-μm-thick neonatal rat transverse medullary slice with rhythmically active respiratory network. Recordings in A and B were obtained respectively before and after block of synaptic transmission by low Ca^{2+} conditions identical to those described in Johnson et al. (1994) (i.e., 0.2 mM Ca^{2+}, 4 mM Mg^{2+}, 9 mM K^{+} in slice bathing solution). Patch pipette solution and procedure for whole-cell recording were as described previously (Smith et al. 1991, 1992). Before block of synaptic transmission, the neuron bursts in synchrony with the inspiratory phase of network activity as monitored by the inspiratory discharge recorded on the hypoglossal (XII) nerve (Smith et al. 1991). After block of synaptic activity (30 minutes under low-Ca^{2+} conditions), the cell exhibits intrinsic voltage-dependent oscillatory behavior. As the cell is depolarized by constant applied current, it undergoes a transition from silence (baseline potential below 65 mV, left) to oscillatory bursting to beating (baseline potential above 45 mV, right). In the bursting regime, the burst period and duration decreases (see expanded time-base traces in B) as the baseline membrane potential is depolarized. SOURCE: Reprinted by permission from R.J. Butera, Jr., J. Rinzel, and J.C. Smith, "Models of Respiratory Rhythm Generation in the Pre-Bötzinger Complex. I. Bursting Pacemaker Neurons," *Journal of Neurophysiology* 82(1):382-397, 1999. Copyright 1999 American Physiological Society.

continued

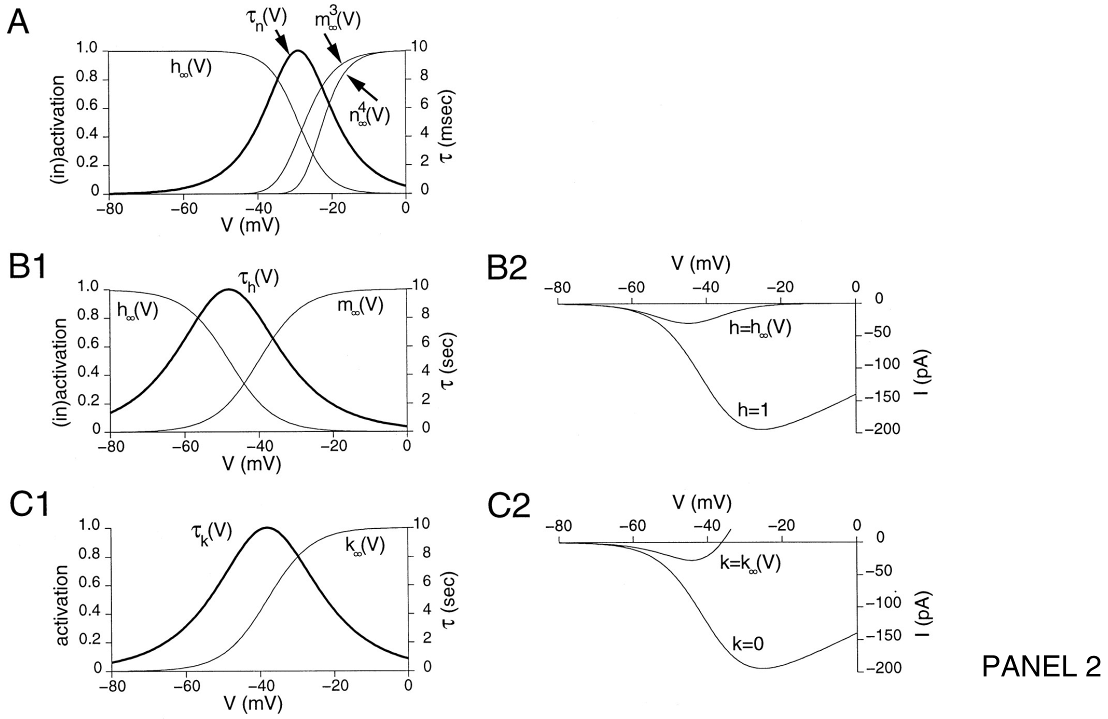


FIGURE 5.16 Continued
Panel 2: Gating and *I-V* characteristics of components of *models 1* and *2*. (A) spike-generating kinetics: $m_\infty{}^3(V)$ and $h_\infty(V)$ of I_{Na} and $n_\infty(V)$ and $\tau_{\mathrm{n}}(V)$ of I_{K}; note that $h = 1 - n$. (B1) gating characteristics of $\mathrm{I}_{\mathrm{NaP}}$: $m_\infty(V)$, $h_\infty(V)$, and $\tau_{\mathrm{h}}(V)$ (bold); *left: y*-axis scale for steady-state gating functions; *right: y*-axis scale for $\tau_{\mathrm{h}}(V)$. (B2) I-V plots of I_{NaP} for $h = h_\infty(V)$ and $h = 1$. First case results in a small window current at subthreshold potentials; second case corresponds to $I_{\mathrm{NaP\text{-}h}}$ with complete removal of inactivation. (C1) gating characteristics of I_{KS}: $k_\infty(V)$ and $\tau_{\mathrm{k}}(V)$ (bold); *left: y*-axis scale for activation function; *right: y*-axis scale for $\tau_{\mathrm{k}}(V)$. (C2) I-V plots of $I_{\mathrm{NaP}} + I_{\mathrm{KS}}$ for $k = k_\infty(V)$ and $k = 0$. First case results in a small current at subthreshold potentials; second case corresponds to I_{NaP} with complete removal of the opposing I_{KS}.
SOURCE: Reprinted by permission from R.J. Butera, Jr., J. Rinzel, and J.C. Smith, "Models of Respiratory Rhythm Generation in the Pre-Bötzinger Complex. I. Bursting Pacemaker Neurons," *Journal of Neurophysiology* 82(1):382-397, 1999. Copyright 1999 American Physiological Society.

ing rhythms displayed by this model are shown in the second panel. The third panel shows a map of the simulated trajectory that illustrates the relationship of the bursting to slow and fast variables in the system.

5.4.5.4 Synaptic Transmission

The intercellular signaling process of synaptic transmission is a much-studied problem. Much has been learned about synaptic structure and function through the classical techniques of neuropharmacology, electron microscopy (EM) neuroanatomy, and electrophysiology, and correlation of the observations made through these various techniques has led to the development of computational models of synaptic microphysiology. However, the scope of previous modeling attempts has been limited by available computing power, modeling framework, and lack of high-resolution three-dimensional ultrastructural data in an appropriate machine representation.

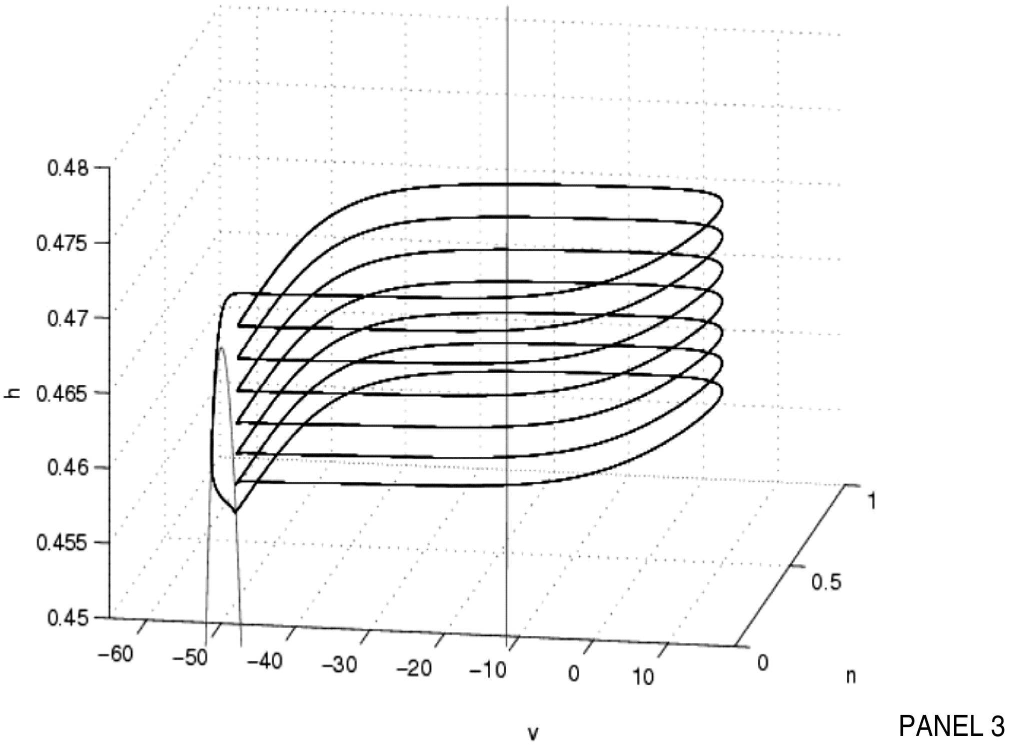


FIGURE 5.16 Continued
Panel 3: Projection of trajectory onto fixed points of fast subsystem. The axes are v: membrane potential; h: inactivation of the HH sodium channel (there is also a persistent sodium channel in the model); and n: activation of the HH "delayed rectifier" potassium channel. The voltage traces show the changes of voltage as a function of time. The values of *h* and *n* also change with time. Think of *v*, *n*, *h* as the three coordinates of a point moving through space. This plot depicts the path taken by this point in a bursting oscillation of the model. The curves are states at which the motion through this space is particularly slow, becoming zero in the limit so that the slower currents in the model are not allowed to change at all. SOURCE: Derived from Figure 4, Panel A3, in R.J. Butera Jr., J. Rinzel, and J.C. Smith, "Models of Respiratory Rhythm Generation in the Pre-Bötzinger Complex. I. Bursting Pacemaker Neurons," *Journal of Neurophysiology* 82(1):382-397, 1999. Copyright 1999 American Physiological Society. Used by permission.

What has been missing is an appropriate set of tools for acquiring, building, simulating, and analyzing biophysically realistic models of subcellular microdomains. Coggan et al. have developed and used a suite of such computational tools to build a realistic computational model of nicotinic synaptic transmission based on serial electron tomograms of a chick ciliary ganglion somatic spine mat.[105]

The chick ciliary ganglion somatic spine mat is a complex system with more than one type of neurotransmitter receptor, possible alternative locations for transmitter release, and a tortuous synaptic geometry that includes a spine mat and calyx-type nerve terminal. Highly accurate models of the synaptic ultrastructure are obtained through large-scale, high-resolution electron tomography; com-

[105] J.S. Coggan, T.M. Bartol, E. Esquenazi, J.R. Stiles, S. Lamont, M.E. Martone, D.K. Berg, M.H. Ellisman, and T.J. Sejnowski, "Evidence for Ectopic Neurotransmission at a Neuronal Synapse," *Science* 309(5733):446-451, 2005.

puter-aided methods for extracting accurate surfaces and defining in-silico representations of their molecular properties; and physiological underpinnings from a variety of studies conducted by the involved laboratories and from the literature.

These data are then used as the framework for advanced simulations using MCell running on high-performance supercomputers as well as distributed or grid-based computational resources. This project pushes development of tools for acquisition of improved large-scale tomographic reconstructions of cellular interfaces down to supramolecular scales. It also drives improvements in the software tools both for the distribution of molecular components within the surface models extracted from the tomographic reconstructions and for the deposition and retrieval of relevant information for the MCell simulator (Box 5.18) in the tomography and Cell-Centered Database (CCDB) environment.

Realistic modeling of synaptic microphysiology (as illustrated in Figure 5.17) requires the following:

1. Acquisition of high-resolution, three-dimensional synaptic ultrastructure—this is accomplished with serial EM tomography.
2. Segmentation of pre- and postsynaptic membrane from the tomographic volume—this is accomplished using the tracing tool in Xvoxtrace.
3. Three-dimensional reconstruction of the membrane surface topology to form a triangle mesh—this is accomplished using the marching cubes isosurface extraction tool in Xvoxtrace.
4. Subdivision of the membrane surface meshes into physiologically relevant regions (e.g., spine versus nonspine membrane and PSD [phosphorylation site domain] versus non-PSD regions)—this is accomplished using the mesh tagging tool in DReAMM.
5. Placement of effector molecules (e.g., receptors, enzymes, reuptake transporters) onto membrane surfaces with the desired distribution and density—this is accomplished using the MCell model description language (MDL). Effector distribution and density may be determined by labeling and imaging studies.
6. Specification of the diffusion constant, quantity, and location of neurotransmitter release—this is accomplished using MCell MDL.
7. Specification of the reaction mechanisms and kinetic rate constants governing the mass action kinetics interaction of neurotransmitter and effector molecules—this is accomplished using MCell MDL.
8. Specification of what quantitative measures should be made during the simulation—this is accomplished using MCell MDL.
9. Simulation of the defined system—this is accomplished using the MCell compute kernel.
10. Analysis of the results at various points in the parameter space defined by the system—this is accomplished using analysis tools of the investigator's discretion.

Analysis of miniature excitatory postsynaptic currents (mEPSCs) recorded in electrophysiological experiments shows that mEPSCs in the CG somatic spine mat occur in a broad spectrum of amplitudes, rise times, and fall times. The differential kinetics and complementary distributions of α3 and α7 nAChRs are expected to lead to mEPSCs whose characteristics are highly dependent on the location of neurotransmitter release within the spine mat. Realistic simulation makes it possible to explore and quantify the degree to which this hypothesis is true and to make quantitative comparisons of the simulation and electrophysiological results. Figure 5.18 summarizes the results of simulations designed to explore the limits of mEPSC behavior by virtue of the choice of neurotransmitter release locations. The results not only confirm the qualitative expectations at each site but also predict their quantitative behavior, allowing fine discriminations to be made.

The process briefly outlined above represents a significant advance in the ability to create realistic computational models of subcellular microdomains from actual cellular ultrastructure. The preliminary results presented are just the beginning of exciting computational experiments that can now be performed on the CG model in an effort to illuminate and inform further bench experiments. Among all of the things learned, perhaps the most important is which of the physical characteristics of the CG are the

Box 5.18
The MCell Simulator

MCell is a general Monte Carlo simulator of cellular microphysiology. MCell simulations provide insights into the behavior and variability of real systems comprising finite numbers of molecules interacting in spatially complex environments. MCell incorporates high-resolution physical structure into models of ligand diffusion and signaling, and thus can take into account the large complexity and diversity of neural tissue at the subcellular level.

MCell is based on the use of rigorously validated Monte Carlo algorithms to track the evolution of biochemical events in time and three-dimensional space for individual ligand and effector molecules. That is, the Monte Carlo approach is based on the use of random numbers and probabilities to effect the simulation of individual cases of the system's behavior.

In the MCell models used in neural signaling employing a Brownian dynamics random walk algorithm, individual ligand molecules move according to a three-dimensional Brownian dynamics random walk and encounter membrane boundaries and effector molecules as they diffuse. Bulk solution rate constants are converted into Monte Carlo probabilities so that the diffusing ligands can undergo stochastic chemical interactions with individual binding sites such as receptor proteins, enzymes, and transporters. These interactions are governed by user-specified reaction mechanisms.

The diffusion algorithms are grid-free, and the reaction algorithms are at the level of interactions between individual molecules and thus do not involve solving systems of differential equations. Membrane boundaries are represented as triangle meshes and may be of arbitrary complexity.

The Monte Carlo approach has certain important advantages over the finite element (FE) approach often used to include spatial information in kinetic modeling. The FE approach divides three-dimensional space into a regular grid of contiguous subcompartments, or voxels. It assumes well-mixed conditions within each voxel and uses differential equations to compute fluxes between, and reactions within, each voxel. Mass action equations are based on continuum processes and predict average concentrations. In large, simple volumes with great numbers of a few types of molecules (e.g., reactions in a test tube), fluctuations are relatively small, and knowledge of average concentrations accounts most of the interesting phenomena. However, synaptic signaling is inherently discrete and stochastic because the number of molecules involved is small; hence, the FE method will fail to describe accurately the biochemistry of synaptic signaling because these methods provide only averaged data. Furthermore, complex cellular structures—such as the structures that characterize the synapse—require that the voxel grid be very fine and irregular in shape, making an FE approach both computationally expensive and difficult to implement.

MCell is very general because it includes a high-level model description language (MDL), which allows the user to build subcellular structures and signaling pathways of virtually any configuration. MCell's algorithms scale smoothly from typical workstations to shared-memory multiprocessor machines to massively parallel supercomputers.

SOURCE: For more information, see http://www.mcell.cnl.salk.edu; J.R. Stiles and T.M. Bartol, Jr., "Monte Carlo Methods for Simulating Realistic Synaptic Microphysiology Using MCell," pp. 87-127 in *Computational Neuroscience: Realistic Modeling for Experimentalists,* E. de Schutter, ed., CRC Press, Boca Raton, FL, 2000; J.R. Stiles, T.M. Bartol, Jr., E.E. Salpeter, M.M. Salpeter, and T.J. Sejnowski, "Synaptic Variability: New Insights from Reconstructions and Monte Carlo Simulations with MCell," pp. 681-731 in *Synapses*, W. Cowan, T.C. Sudhof, and C.F. Stevens, eds., Johns Hopkins University Press, Baltimore, MD, 2001. Discussion of the pros and cons of FE versus MC is from K.M. Franks and T.J. Sejnowski, "Complexity of Calcium Signaling in Synaptic Spines," *BioEssays* 24(12):1130-1144, 2002.

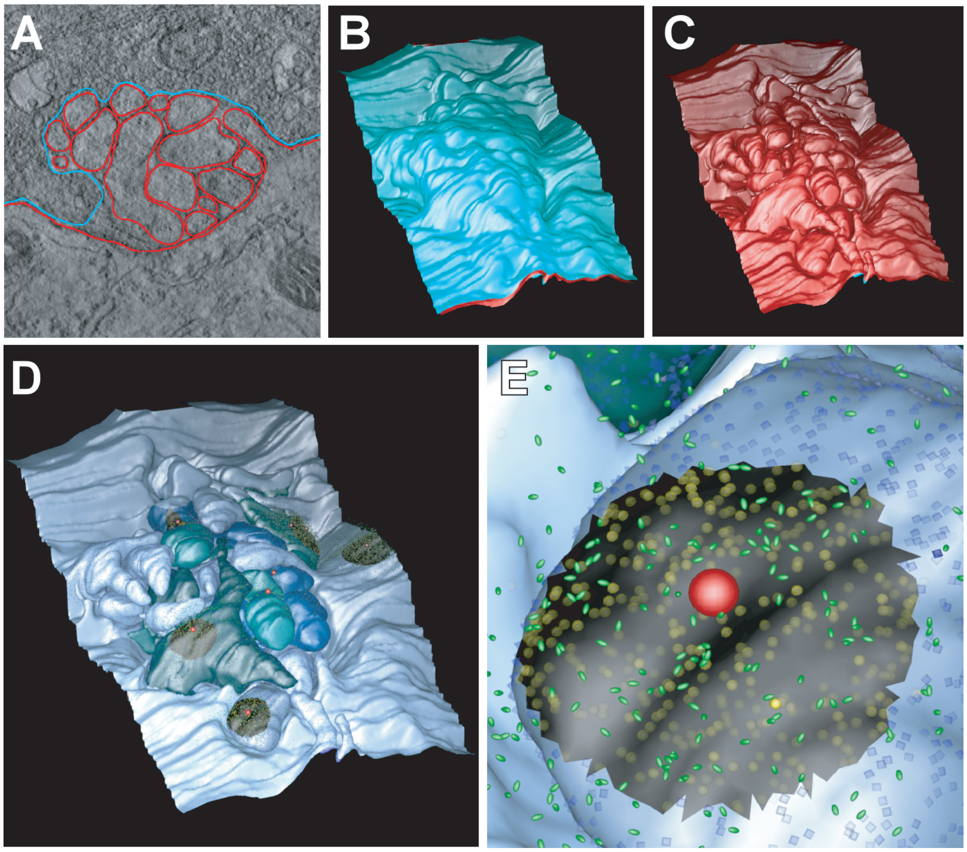


FIGURE 5.17 Constructing the geometry of a chick ciliary ganglion (CG) somatic spine mat model. A serial EM tomogram of a CG spine mat was obtained at ~4 nm per voxel resolution. The serial tomogram encompassed a volume of ~27 mm^3 (~3 µm × 3 µm × 3 µm).

(A) A typical slice through the tomographic volume together with hand-traced contours of the pre- and postsynaptic membranes. Tracing and segmentation of presynaptic (cyan) and postsynaptic (red) membrane contours generated using Xvoxtrace.

(B) Three-dimensional reconstruction of pre- and postsynaptic membrane surfaces as triangle meshes—view looking down onto intracellular face of presynaptic membrane (visualized using DReAMM). The presynaptic mesh is composed of 100,000 triangles and the postsynaptic mesh is composed of 300,000 triangles.

(C) Postsynaptic membrane surface—view of extracellular face of membrane (presynaptic membrane invisible).

(D) Completed model including postsynaptic membrane subdivided into distinct spines, PSD areas (black regions with yellow borders), receptor molecules (tiny blue particles on membrane surface), and several neurotransmitter release sites (red spheres). The membrane was subsequently populated with the desired distributions and densities of nicotinic acetylcholine receptor (nAChR) types and acetylcholine esterase (AChE) enzyme. Also visible in (D) are several acetylcholine (ACh) vesicular release sites whose locations are most clearly illustrated in Figure 5.18A.

(E) Magnified view of the state of a simulation of synaptic transmission model as simulated by MCell. State of system 300 ms after release of 5,000 molecules of acetylcholine (small green ellipsoids) is shown. α7 nAChR types are shown in blue, and α3* nAChR types are shown in yellow (inactive receptors are semitransparent, and open receptors are opaque).

SOURCE: Courtesy of Tom Bartol, Salk Institute, San Diego, California.

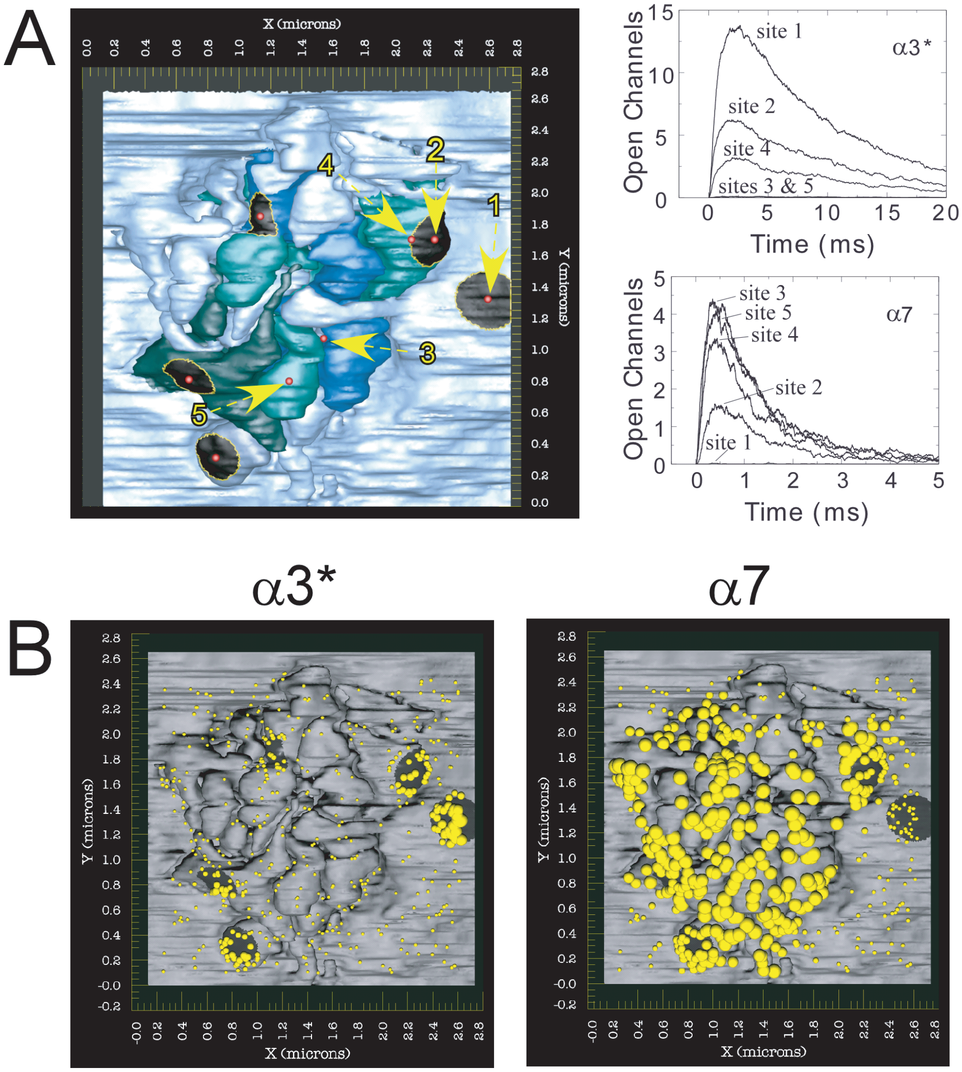


FIGURE 5.18 Summary of synaptic transmission simulations.
(A) Location of selected transmitter release sites and their associated simulated mEPSC traces, each decomposed into their α3 and α7 nAChR components. Each trace is the average of 100 simulations using MCell. Site 1 is located at a PSD on nonspine membrane. This site is expected to have a large α3 response and a very small α7 response. At the other extreme of behavior, sites 3 and 5 are placed over non-PSD spine membrane. Rich in α7 receptors and poor in α3 receptors, these sites are expected to have large α7 responses and minimal α3 responses. The other sites are placed at locations expected to give rise to mEPSCs of mixed nAChR origin.
(B) mEPSC amplitudes (decomposed into their α3 and α7 nAChR components) at each of 550 distinct vesicular release sites. The mEPSC amplitudes are indicated by the diameter of the yellow spherical glyph and demonstrate a strong dependence on location and underlying geometry.
SOURCE: Courtesy of Tom Bartol, Salk Institute, San Diego, CA.

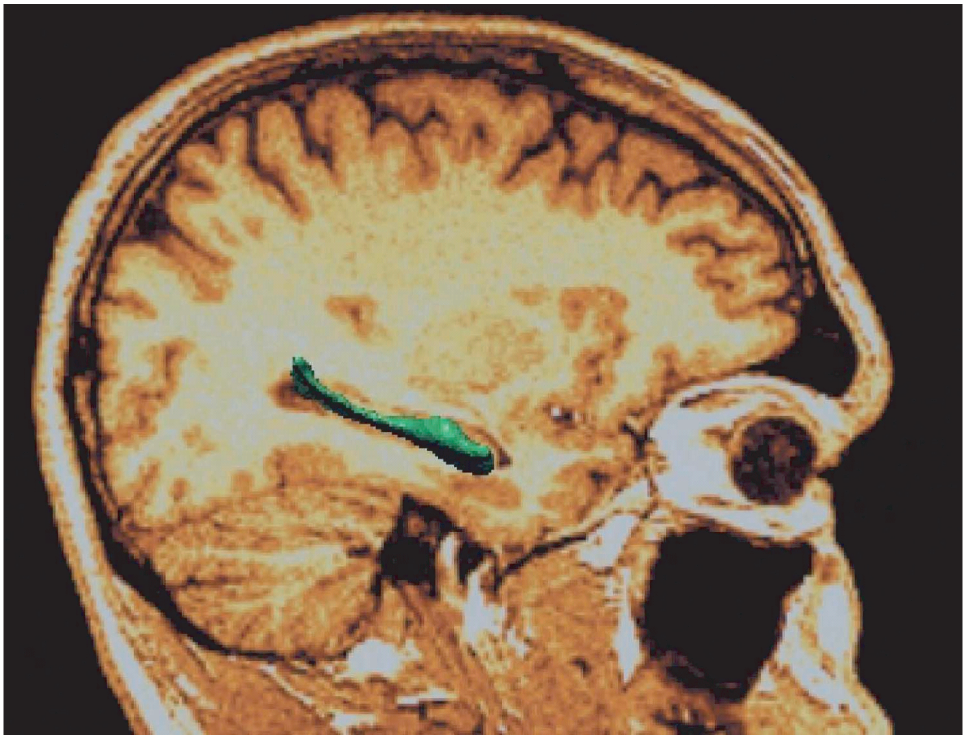


FIGURE 5.19 The hippocampus in situ. SOURCE: Courtesy of Michael Miller, Johns Hopkins University.

least constrained, are least understood, and have the greatest impact on synaptic function. Specifically, the results clearly demonstrate that synaptic geometry, receptor distribution, and vesicle release location each have a profound quantitative impact on the efficacy of the postsynaptic response. This means that attention to accuracy in the model-building process must be a prime concern.

5.4.5.5 Neuropsychiatry[106]

The field of computational neuropsychiatry has been exploding with applications of large-deformation brain mapping technology that provide mechanisms for discovering neuropsychiatric disorders of many types. The hippocampus is a region of the brain (depicted in green in Figure 5.19) that has been implicated in schizophrenia and other neurodegenerative diseases such as Alzheimer's. Using large-deformation brain mapping tools in computational anatomy, researchers can define, visualize, and measure the volume and shape of the hippocampus. These methods allow for precise assessment of changes in hippocampal formation.

Researchers at the Center for Imaging Science (CIS) used mapping tools to compare the left and right hippocampi (Figure 5.20) in 15 pairs of schizophrenic and control subjects. In the schizophrenic

[106] Section 5.4.5.5 is based on L. Wang, S.C. Joshi, M.I. Miller, and J.G. Csernansky, "Statistical Analysis of Hippocampal Asymmetry in Schizophrenia," *Neuroimage* 14(3):531-545, 2001; J.G. Csernansky, L. Wang, S. Joshi, J.P. Miller, M. Gado, D. Kido, D. McKeel, et al., "Early DAT Is Distinguished from Aging by High-dimensional Mapping of the Hippocampus," *Neurology* 55(11):1636-1643, 2000.

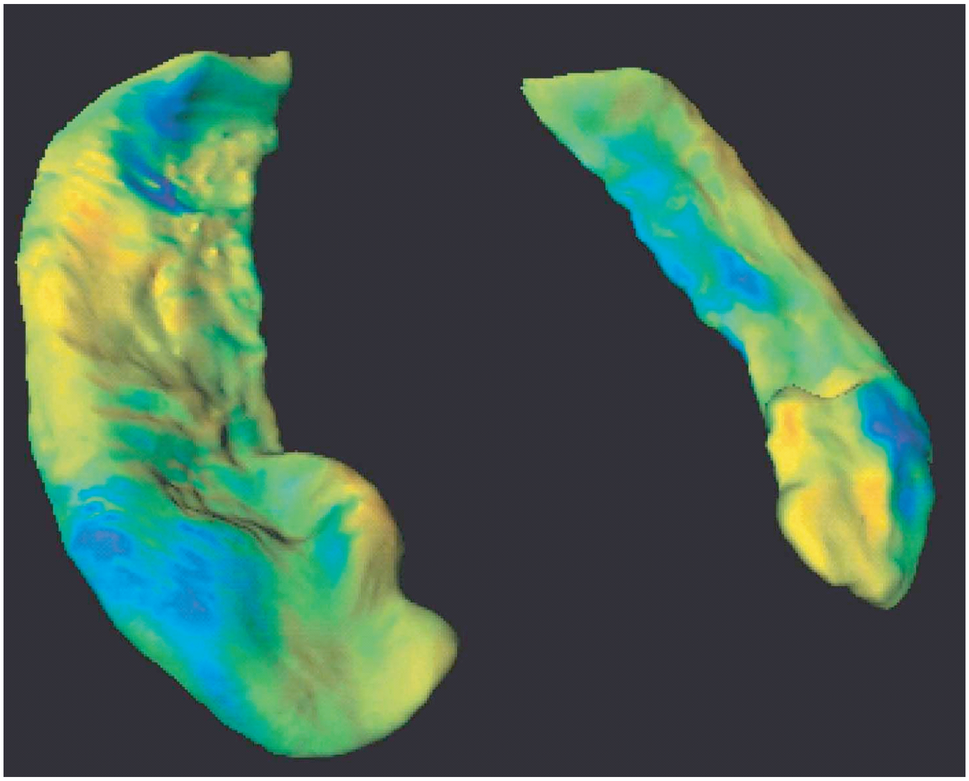


FIGURE 5.20 Left and right hippocampuses. SOURCE: Courtesy of Michael Miller, Johns Hopkins University.

subjects, deformations were localized to hippocampal subregions that send projections to the prefrontal cortex. The deformations strongly distinguish schizophrenic subjects from control subjects. The pictures indicate inward deformations by cooler colors, outward deformations by warmer colors, and little deformation by a neutral green color. These results support the current hypothesis that schizophrenia involves a disturbance of hippocampal-prefrontal connections.

In a separate study, CIS researchers also compared asymmetry between the left and right hippocampi. The left and the right side of normal brains develop at different rates. Structures on both sides of the brain are similar, but not identical. This is normal brain asymmetry. If a different asymmetry pattern exists in schizophrenic subjects, it may indicate a disturbance of the left-right balance during early stages of brain development. Researchers found that the left hippocampus was narrower along the outside edge than the right hippocampus. This asymmetry was similar in schizophrenic and normal subjects (Figure 5.21, left image). However, further comparison revealed a significant difference in asymmetry patterns of the hippocampal area called the subiculum (Figure 5.21, right image). People with schizophrenia tend to have a more pronounced depression and a downward bend in the surface of that structure.

As part of Washington University's Healthy Aging and Senile Dementia (HASD) program, CIS researchers have also applied brain mapping tools to assess the structure of the hippocampus in older human subjects (depicted in Figure 5.22). They compared measurements of hippocampal volume and shape in 18 subjects with early dementia of the Alzheimer type (DAT) with 18 healthy elderly and 15 younger control subjects. Hippocampal volume loss and shape deformities observed in subjects with DAT distinguished them from both elderly and younger control subjects. The pattern of hippocampal

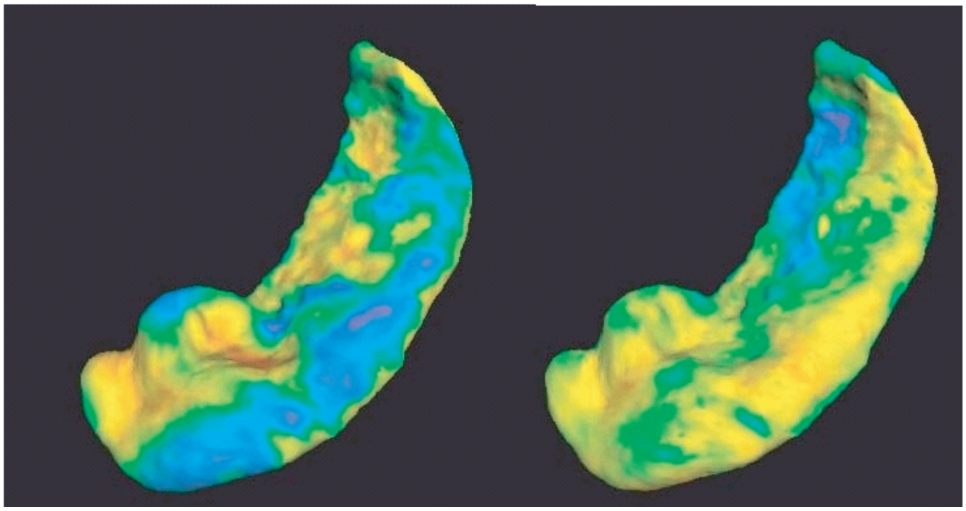


FIGURE 5.21 Asymmetry in schizophrenia. SOURCE: Michael Miller, Johns Hopkins University.

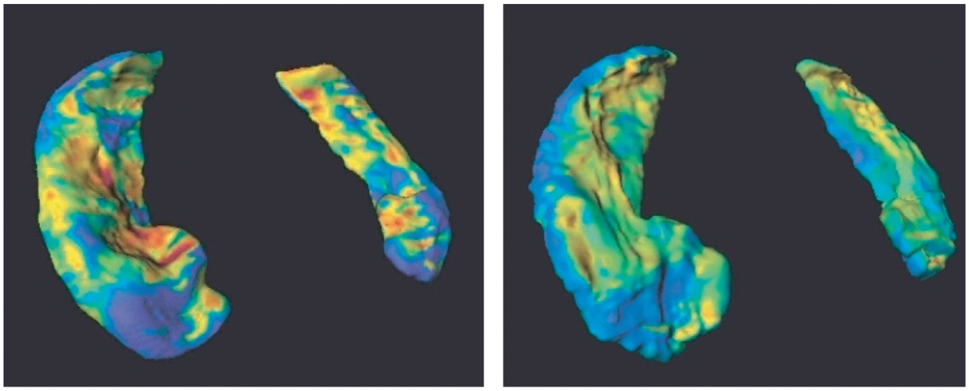


FIGURE 5.22 Hippocampal structure in normal aging (left) versus in Alzheimer's disease patients (right). SOURCE: Courtesy of Michael Miller, Johns Hopkins University.

deformities in subjects with DAT was largely symmetric and suggested damage to the CA1 hippocampal subfield.

Hippocampal shape changes were also observed in healthy elderly subjects, which distinguished them from healthy younger subjects. These shape changes occurred in a pattern distinct from the pattern seen in DAT and were not associated with substantial volume loss. These assessments indicate that hippocampal volume and shape derived from computational anatomy large deformation brain mapping tools may be useful in distinguishing early DAT from healthy aging.

5.4.6 Virology

Mathematical and computational methods are increasingly important to virology. For example, a primary and surprising phenomenological aspect of HIV infection is that progression to AIDS usually

Box 5.19
Modeling the In Vivo Dynamics of HIV-1 Infection

Mathematical models of HIV infection and treatment have provided quantitative insights into the major biological processes that underlie HIV pathogenesis and helped establish the treatment of patients with combination therapy. This in turn has changed HIV from a fatal disease to a treatable one. The models successfully describe the changes in viral load in patients under therapy and have yielded estimates of how rapidly HIV is produced and cleared in vivo, how long HIV-infected cells survive while producing HIV, and how fast HIV mutates and evolves drug resistance. They have also provided clues into the process of T-cell depletion that characterizes AIDS. The models have also provided means to rapidly screen antiviral drug candidates for potency in vivo, thus hastening the introduction of new antiretroviral therapies.

On average, HIV takes about 10 years to advance from initial infection to immune dysfunction (or AIDS). During this period the amount of virus measured in a person's blood hardly changes. Because of this slow progression and the unchanging level of virus it was initially thought that this infection was slow and it was unclear whether treating this disease early, when symptoms were not apparent, was worthwhile.

Recognizing that constant levels of virus meant only that the rates of viral production and clearance were in balance, but not necessarily slow, Perelson and David Ho from Rockefeller University used experimental drug therapy to "perturb" the viral steady state. Mathematically modeling the response to this perturbation using a system of ordinary differential equations that kept track of the concentrations of infected cells and HIV, and fitting the experimental data to the model, revealed a plethora of new features of HIV infection.

Figure 5.19.1 shows that after therapy is initiated at time 0, levels of HIV RNA (a surrogate for virus) fall ten- to a hundredfold in the first week or two of therapy. This suggested that HIV has a half-life ($t_{1/2}$) of 1-2 days, and thus maintaining the pre-therapy constant level of virus requires enormous virus production—in fact, the amount of virus in the body must double every 1-2 days.

Detailed analysis showed that this viral decay was governed by two processes, clearance of free virus particles ($t_{1/2} < 6$ hours) and loss of productively infected cells ($t_{1/2} < 1.6$ days). From this rapid clearance of virus one could compute that at steady state, ~10^{10} virions are produced daily and given the mutation rate of HIV, that each single and most double mutations of the HIV genome are produced daily. Thus, effective drug therapy

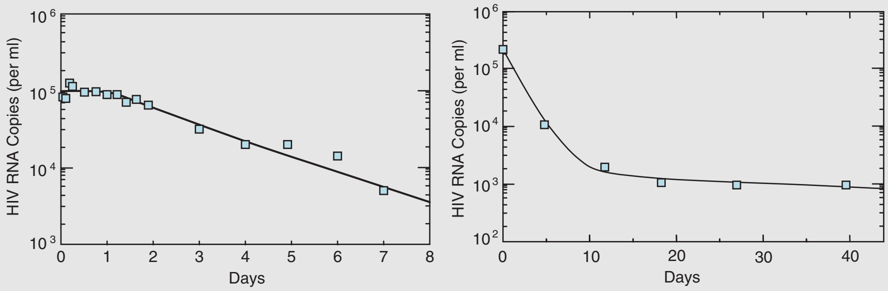


FIGURE 5.19.1 Model predictions (lines) of the biphasic decay of HIV viral load compared with typical patient data (symbols). SOURCE: Courtesy of A.S. Perelson, Los Alamos National Laboratory.

continued

Box 5.19 Continued

would require drug combinations that can sustain at least three mutations before resistance arises, and this engendered the idea of triple combination therapy. Other analyses showed that the slope of viral decay was proportional to the drug combinations' antiviral efficacy, providing a means of comparing therapies.

Following the rapid 1-2 week "first phase" loss, the rate of HIV RNA decline slows. Models of this "second phase" of decline, when fitted to the kinetic data, suggested that a small fraction of infected cells might live a period of weeks while infected ($t_{1/2}$ ~ 14 days).

Following upon the success of these joint modeling and experimental efforts, many similar studies were undertaken and revealed a fourth, much longer time-scale for the decay of latently infected cells of 6-44 months. Latently infected cells, which harbor the HIV genome but do not produce virus, can hide from the immune system and reignite infection when the cells are stimulated into proliferation. Clearing latently infected cells is one of the last remaining obstacles to eradicating HIV from the body.

takes a very long time, and individuals who have not progressed to full-blown AIDS are asymptomatic. As Box 5.19 suggests, computational models have been able to shed considerable light on this phenomenology, and these insights have altered the view of AIDS from a static picture in which the virus is essentially dormant and does not do very much for a long time to a much more dynamic picture of a rough balance between the virus and the immune system, both working very hard, for that period of time. These findings have had tangible impact, because they have affected drug treatment regimes considerably.

More specifically, the average rate of HIV production in the human body is on the order of 10^{10} copies per day as noted in Box 5.19. Empirical data indicate that errors in HIV replication occur at a rate on the order of 10^{-4} to 10^{-5} per base per generation, and since the HIV genome is 10,000 base pairs long, the likelihood that a replicated genome will contain at least one error is 10 percent to nearly unity (and the vast majority of these errors are errors in a single base). Because there are only four possible bases in DNA (and hence each base can change into only one of three other bases), there are only 30,000 possible single-base mutations of a given genome. An error rate of 10^{-4} to 10^{-5} per base per generation distributed among 10^{10} copies each with 10^4 bases means that each generation produces 10^9 to 10^{10} mutations, which are distributed over the set of 30,000 possible mutations. Put differently, every new day brings to life on the order of 10^5 instances of every possible single-base variant of HIV.

Thus, a drug known to bind to a particular sequence of amino acids at a certain location in a protein today will face 10^5 to 10^6 new variants tomorrow against which its effectiveness will be questionable. This fact suggests that drug treatment regimes must target multiple binding sites, and hence combination drug therapy is likely to be more effective because drug-resistant variants must then be the result of multiple errors in the replication process (which occur much less frequently). This in fact reflects recent experience with combination drug regimes.[107]

5.4.7 Epidemiology

Epidemiology is the study of the dynamics of disease in a population of individuals. Of particular interest is the epidemiology of infectious diseases, which arise from contact between an environmental

[107] For further discussion, see A.G. Rodrigo, "HIV Evolutionary Genetics," *Proceedings of the National Academy of Sciences* 96(19):10559-10561, 1999; B.A. Cipra, "Will Viruses Succumb to Mathematical Models?" *SIAM News* 32(2), 1999, available at http://www.siam.org/siamnews/03-99/viruses.pdf.

Box 5.20
Spatial Heterogeneity in Epidemiology: An Example

One of the best illustrations of [the significance of spatial heterogeneity] is provided by the highly dynamic spatiotemporal epidemic pattern of measles. An important set of analyses of simple, homogeneous models predicted the possibility of chaotic dynamics; however, the resulting large-amplitude [predicted] epidemics generate unrealistically low persistence of infection in small communities. Adding successive layers of social and geographical space—and moving from deterministic to stochastic models—improves spatial realism and may reduce the propensity for chaos.

The major computational challenge in these highly nonlinear stochastic systems is to represent hierarchical spatial complexity and especially its impact on vaccination strategies. Depending on the problem, all scales—from the individual level to big cities—may be important, both in terms of social space [family and school infection dynamics] and in terms of geographic spread and coherency.

. . . [A] central question is: How spatially aggregated and parsimonious a model can provide useful results in a given context? This is particularly important in comparisons between directly transmitted human infections—where long-range movements may bring infection dynamics comparatively close to mean field behavior (in which every individual is assumed to have equal contact with every other individual, thus experiencing the mean or average field)—and the equivalent infections in natural populations, where more restricted movements and host population dynamics add extra complexities.

It is risky to model at a given level of detail without having data at the relevant spatial grain. Notifiable infectious diseases are unusually well [documented], with large and often as yet uncomputerized spatiotemporal data sets. These data provide a huge potential testbed for developing methods for characterizing spatiotemporal dynamics in nonlinear, nonstationary stochastic systems. An encouraging development is that the current, generally nonparametric, approaches to characterizing chaos and other nonlinear behaviors are increasingly incorporating lessons from mechanistic epidemiological models.

SOURCE: Reprinted by permission from S.A. Levin, B. Grenfell, A. Hastings, and A.S. Perelson, "Mathematical and Computational Challenges in Population Biology and Ecosystems Science," *Science* 275(5298):334-343, 1997. Copyright 1997 AAAS. (References omitted.)

agent and an individual (e.g., an insect that bites an individual) or between individuals (e.g., an individual who sneezes in a room filled with people) that leads to the transmission of disease. The dynamics of infectious diseases depend on many things, such as the likelihood of transmission between carrier agent and infected individual given that contact has been made, the geographical distribution of carrier agents and individuals, and the susceptibility of individuals to the disease.

A central problem in epidemiology is how the dynamics of disease play out across geographical space.[108] Problems of spatial heterogeneity play out at many different levels of aggregation: individuals, families, work groups and firms, neighborhood, and cities. Box 5.20 provides an example taken from the study of measles.

At the same time, spatial heterogeneity is not the only inhomogeneity of interest. For example, the epidemiology of sexually transmitted diseases (STDs) cannot be separated from a consideration of their dynamics in different social groups. For example, patterns of STDs in prostitutes and intravenous drug

[108] K. Dietz, "The Estimation of the Basic Reproduction Number for Infectious Diseases," *Statistical Methods in Medical Research* 2(1):23-41, 1993; A.D. Cliff and P. Haggett, *Atlas of Disease Distributions: Analytic Approaches to Epidemiologic Data*, Blackwell LTD, Oxford, UK, 1988; D. Mollison and S.A. Levin, "Spatial Dynamics of Parasitism," pp. 384-398 in *Ecology of Infectious Diseases in Natural Populations*, B.T. Greenfell and A.P. Dobson, eds., Cambridge University, Cambridge, UK, 1995.

Box 5.21
Social Heterogeneity in Epidemiology: An Example

The main focus for modeling social space (the space of social interactions) and disease is, of course, on AIDS and other sexually transmitted infections. Simple models illustrated clearly that heterogeneities in contact rates can substantially alter the predicted course of epidemics. This area has seen an explosion of research, both in data analysis of contact structures and in graph-theoretic and other approaches to modeling. Models and data analysis are most productive when combined, especially in allowing the observations to limit the universe of possible networks.

The major computational challenge is how to deal with the complexity of networks, where concurrency of partnerships often means that closure to a few moments of the distribution is difficult. This problem is especially acute given the sensitivity of obtaining data for STD networks, in that the nature of the network is generally only partially and imperfectly known. The use of mathematical models for human immunodeficiency virus (HIV) transmission will be especially important in assessing the impact of potential vaccines. Another major computational challenge—which developed with the AIDS epidemic and is currently being applied to another pathogen, the bovine spongiform encephalopathy agent—is to estimate the parameters of transmission models from disease incidence and other demographic data.

One hope for the future for both of these areas is network information embedded in viral genomes. A body of recent work indicates exciting possibilities for estimating epidemiological parameters from the birth and death processes of pathogen evolutionary trees. More generally, new mathematical and computational techniques will be needed to understand the epidemiological implications of the rapidly accumulating data on pathogen sequences, especially in the context of parasite genetic diversity and the host immunological response to it.

SOURCE: Reprinted by permission from S.A. Levin, B. Grenfell, A. Hastings, and A.S. Perelson, "Mathematical and Computational Challenges in Population Biology and Ecosystems Science," *Science* 275(5298):334-343,1997. Copyright 1997 AAAS. (References omitted.)

users exhibit different dynamical patterns than those in the general population because of factors such as rates of sexual contact with others (both inside and outside the individual's own social group) and different sexual practices of individuals in each group (e.g., use of condoms). Box 5.21 elaborates on this notion in greater detail.

5.4.8 Evolution and Ecology

5.4.8.1 Commonalities Between Evolution and Ecology

No two fields in biology encompass such a broad range of levels of biological organization as ecology and evolutionary biology. Although the two fields ask different questions, they both contend with factors of space and time, and share common theories about relationships between individuals, populations, and communities. The two intertwined fields view these relationships in different ways. Evolutionary biologists want to understand and quantify the effect of environment (e.g., natural selection) on individuals and populations; ecologists want to understanding the role of individuals and populations in shaping their environment (ecological inheritance, niche construction).

The two fields encompass a diverse assemblage of topics with applications in resource management, epidemiology, and global change. In these fields, data have been relatively difficult to collect in ways that relate directly to mathematical or computational models, although this has been changing over the past 10 years. Thus, both fields have relied heavily on theory to advance their insights. In fact, ecology and evolution have been the substrate for the development of important mathematical concepts. The quantitative study of biological inheritance and evolution provided the context for statistics, probability theory, stochasticity, and dynamical systems theory.

Among the fundamental questions in the study of evolution are those that seek to know the relative strengths of natural selection, genetic drift, dispersal processes, and genetic recombination in shaping the genome of a population—essentially the forces that provide genetic variability in a species. Both ecologists and evolutionary biologists want to know how these forces lead to morphological changes, speciation, and ultimately, survival over time. The fields seek theory, models, and data that can account for genetic changes over time in large heterogeneous populations in which genetic information is exchanged routinely in an environment that also exerts its influence and changes over time.

In addition to interest in genetic variability and fitness within a single species, the two fields are interested in relationships between multiple species. In ecology, this manifests itself in questions of how the individual forces of variability within and between species affect their relative ability to compete for resources and space that leads to their survival or extinction—in other words, forces that determines the biodiversity of an ecosystem (i.e., a set of biological organisms interacting among themselves and their environment). Ecologists want to understand what determines the minimum viable population size for a given population, the role of keystone species in determining the diversity of the ecosystem, and the role of diversity in preservation of the ecosystem.

For evolutionary biologists, questions regarding relationships between species focus on trying to understand the flow of genetic information over long periods of time as a measure of the relatedness of different species and the effects of selection on the genetic contribution to phenotypes. Among the great mysteries for evolutionary biologists is whether and how evolution relates to organismal development, an interaction for which no descriptive language currently exists.

How will ecologists and evolutionary biologists answer these questions? These fields have had few tools to monitor interactions in real time. But new opportunities have emerged in areas from genomics to satellite imaging and in new capabilities for the computer simulation of complex models.

5.4.8.2 Examples from Evolution

A plethora of genomic data is beginning to help untangle the relationship between traits, genes, developmental processes, and environments. The data will serve as the substrate from which new statistical conclusions can be drawn, for example, new methods for identifying inherited gene sequences such as those related to disease. To answer question about the process of genome rearrangement, the possibility of comparing gene sequences from multiple organisms provides the basis for testing tools that discern repeatable patterns and elucidate linkages.

As more detailed DNA and protein sequence information is compiled for more genes in more organisms, computational algorithms for estimating parameters of evolution have become extremely complex. New techniques will be needed to handle the likelihood functions and produce satisfactory statistics in a reasonable amount of time. Studies of the role of environmental and genetic plasticity in trait development will involve large-scale simulations of networks of linked genes and their interacting products. Such simulations may well suggest new approaches to such old problems as the nature-nurture dichotomy for human behaviors.

New techniques and the availability of more powerful computers have also led to the development of highly detailed models in which a wide variety of components and mechanisms can be incorporated. Among these are individual unit models that attempt to follow every individual in a population over time, thereby providing insight into dynamical behavior (Box 5.22).

Levin argues that such models are "imitation[s] of reality that represent at best individual realization of complex processes in which stochasticity, contingency, and nonlinearity underlie a diversity of possible outcomes."[109] From the collective behaviors of individual units arise the observable dynamics

[109] S.A. Levin, B. Grenfell, A. Hastings, and A.S. Perelson, "Mathematical and Computational Challenges in Population Biology and Ecosystems Science," *Science* 275(5298):334-343, 1997.

Box 5.22
The Dynamics of Evolution

Avida is a simulation software system developed at the Digital Life Laboratory at the California Institute of Technology.[1] In it, digital organisms have genomes comprised of a sequence of instructions that operate on a virtual machine. These instructions include the ability to perform simple mathematical operations, copy values from memory location to memory location, provide input and output, and check conditions. Through a sequence of instructions, these organisms can copy their genome, thereby reproducing asexually. Since the software can simulate many hundreds of thousands of generations of evolution for thousands of organisms, their digital evolution not only can be observed in reasonable lengths of time, but also can be precisely inspected (since there are no inconvenient gaps in the fossil record). Moreover, alternate scenarios can be explored by going back into evolutionary history and reversing the effects of mutations, for example. At a minimum, this can be seen as experiment by analogy, revealing potential avenues for investigation or hypotheses to test in actual biological evolution. A stronger argument holds that evolution is an abstract mathematical process and will operate under similar dynamics whether embodied in DNA in the physical world or in digital simulations of it.

Avida has been used to explore how complex features can arise through mutation, competition, and selective pressure.[2] In a series of experiments, organisms were provided with a limited supply of energy units necessary for the execution of their genome of instructions. However, organisms that performed any of a set of complex logical operations were rewarded with an increased allowance and thus increased opportunities to reproduce. More complicated logical operations provided proportionally greater rewards.

The experiment was seeded with an ancestral form that could perform none of those operations, containing only the instructions to reproduce. Mutation arose through imperfect copying of the genome during reproduction. EQU, the most complex logical operation checked for [representing the logical statement (A and B) or (~A and ~B)], arose in 23 out of 50 populations studied where the simpler operations also provided rewards. The sequence of instructions that evolved to perform the operation varied widely in length and implementation. However, in other simulations where only EQU was rewarded, no lineages ever evolved it. This evidence agrees with the standard theory of biological evolution—stated as early as Darwin—that complex structures arise through the combination and modification of useful intermediate forms.

[1] C. Adami, *Introduction to Artificial Life*, Springer-Verlag, New York, 1998.
[2] R.E. Lenski, C. Ofria, R.T. Pennock, and C. Adami, "The Evolutionary Origin of Complex Features," *Nature* 423:139-144, 2003.

of the system. "The challenge, then, is to develop mechanistic models that begin from what is understood about the interactions of the individual units, and to use computation and analysis to explain emergent behavior in terms of the statistical mechanics of ensembles of such units." Such models must extrapolate from the effects of change on individual plants and animals to changes in the distribution of individuals over longer time scales and broader space scales and hence in community-level patterns and the fluxes of nutrients.

5.4.8.2.1 Reconstruction of the Saccharomyces *Phylogenetic Tree* Although the basic structure and mechanisms underlying evolution and genetics are known in principle, there are many complexities that force researchers into computational approaches in order to gain insight. Box 5.23 addresses complexities such as multiple loci, spatial factors, and the role of frequency dependence in evolution, and discusses a computational perspective on the evolution of altruism, a behavioral characteristic that is counterintuitive in the context of individual organisms doing all that they can to gain advantage in the face of selection pressures.

Box 5.23
Genetic Complexities in Evolutionary Processes

The dynamics of alleles at single loci are well understood, but the dynamics of alleles at two loci are still not completely understood, even in the deterministic case. As a rule, two-locus models require the use of a variety of computational approaches, from straightforward simulation to more complex analyses based on optimization or the use of computer algebra systems. Three-locus models can be understood only through numerical approaches, except for some very special cases.

Compare these analytical capabilities to the fact that the number of loci exhibiting genetic variation in populations of higher organisms is well into the thousands. Thus, the number of possible genotypes can be much larger than the population. In such a situation, the detailed population simulation (i.e., a detailed consideration of events at each locus) leads to problems of substantial computational difficulty.

An alternative is to represent the population as phenotypes—that is, in terms of traits that can be directly observed and described. For example, certain traits of individuals are quantitative in the sense that they represent the sum of multiple small effects. Efforts have been undertaken to integrate statistical models of the dynamics of quantitative traits with more mechanistic genetic approaches, though even under simplifying assumptions concerning the relation between genotype and phenotype, further approximations are required to obtain a closed system of equations.

Frequency dependence in evolution refers to the phenomenon in which the fitness of an individual depends both on its own traits and on the traits of other individuals in the population—that is, selection is dependent on the frequency with which certain traits appear in the population, not just on pressures from the environment.

This point arises most strongly in understanding how cooperation (altruism) can evolve through individual selection. The simplest model is the game of prisoner's dilemma, in which the game-theoretic solution for a single encounter between parties is unconditional noncooperation. However, in the iterated prisoner's dilemma, the game theoretic solution is a strategy known as "tit-for-tat," which begins with cooperation and then uses the strategy employed by the other player in the previous interaction. (In other words, the iterated prisoner's dilemma stipulates repeated interactions over time between players.)

Although the iterated prisoner's dilemma yields some insight into how cooperative behavior might emerge under some circumstances, it is a highly and perhaps oversimplified model. Most importantly, it does not account for possible spatial localizations of individuals—a point that is important in light of the fact that individuals who are spatially separated have low probabilities of interacting. Because the evolution of traits dependent on population frequency requires knowledge of which individuals are interacting, more realistic models introduce some explicit spatial distribution of individuals—and, for these, simulations are required to dynamical understanding. These more realistic models suggest that spatial localization affects the evolution of both cooperative and antagonistic behaviors.

SOURCE: Adapted from S.A. Levin, B. Grenfell, A. Hastings, and A.S. Perelson, "Mathematical and Computational Challenges in Population Biology and Ecosystems Science," *Science* 275(5298):334-343, 1997. (References in the original.)

Along these lines, a particularly interesting work on the reconstruction of phylogenies was reported in 2003 by Rokas et al.[110] One of the primary goals of evolutionary research has been understanding the historical relationships between living organisms—reconstruction of the phylogenetic tree of life. A primary difficulty in phylogenetic reconstruction is that different single-gene datasets often result in different and incongruent phylogenies. Such incongruences occur in analyses at all taxonomic levels, from phylogenies of closely related species to relationships between major classes or phyla and higher taxonomic groups.

Many factors, both analytical and biological, may cause incongruence. To overcome the effect of some of these factors, analysis of concatenated datasets has been used. However, phylogenetic analyses of different sets of concatenated genes do not always converge on the same tree, and some studies have yielded results at odds with widely accepted phylogenies.

Rokas et al. exploited genome sequence data for seven *Saccharomyces* species and for the outgroup fungus *Candida albicans* to construct a phylogenetic tree. Their results suggested that datasets consisting of a single gene or a small number of concatenated genes had a significant probability of supporting conflicting topologies, but that use of the entire dataset of concatenated genes resulted in a single, fully resolved phylogeny with the maximum likelihood. In addition, all alternative topologies resulting from single-gene analyses were rejected with high probability. In other words, even though the individual genes examined supported alternative trees, the concatenated data exclusively supported a single tree. They concluded that "the maximum support for a single topology regardless of method of analysis is strongly suggestive of the power of large data sets in overcoming the incongruence present in single-gene analyses."

5.4.8.2.2 Modeling of Myxomatosis Evolution in Australia Evolution also provides a superb and easy-to-understand example of time scales in biological phenomena. Around 1860, a nonindigenous rabbit was introduced into Australia as part of British colonization of that continent. Since this rabbit had no indigenous foe, it proliferated wildly in a short amount of time (about 20 years). Early in the 1950s Australian authorities introduced a particular strain of virus that was deadly to the rabbit.

The data indicated that in the short term (say, on a time scale of a few months), the most virulent strains of the virus were dominant (i.e., the virus had a lethality of 99.8 percent). This is not surprising, in the sense that one might expect virulence to be a measure of viral fitness. However, in the longer term (on a scale of decades), similar measurements indicate that these more virulent strains were no longer dominant, and the dominant niche was occupied by less virulent strains (lethality of 90 percent or less). The evolutionary explanation for this latter phenomenon is that an excessively virulent virus would run the risk of killing off its hosts at too rapid a rate, thereby jeopardizing its own survival. The underlying mechanism responsible for this counterintuitive phenomenon is that transmission of the virus depended on mosquitoes feeding from live rabbits. Rabbits that were infected with the more virulent variant died quickly, and thus, fewer were available as sources of that variant.

The above system was modeled in closed form based on a set of coupled differential equations; this model was successful in reproducing the essential qualitative features described above.[111] In 1990, this model was extended by Dwyer et al. to incorporate more biologically plausible features.[112] For example, the evolution of rabbit and virus reacting to each other was modeled explicitly. A multiplicity of

[110] A. Rokas, B.L. Williams, N. King, and S.B. Carroll, "Genome-scale Approaches to Resolving Incongruence in Molecular Phylogenies," *Nature* 425(6960):798-804, 2003.

[111] S. Levin and D. Pimentel, "Selection of Intermediate Rates of Increase in Parasite-Host Systems," *The American Naturalist* 117(3), 1981.

[112] G. Dwyer, S.A. Levin, and L.A. Buttel, "A Simulation Model of the Population Dynamics and Evolution of Myxomatosis," *Ecological Monographs* 60(4):423-447, 1990.

virus vectors was modeled, each with different transmission efficiencies, rather than assuming a single vector. The inclusion of such features, cou pled with exploitation of a wealth of data available on this system, allowed Dwyer et al. to investigate questions that could not be addressed in the earlier model. These questions included whether the system will continue to evolve antagonistically and whether the virus will be able to control the rabbit population in the future.

More broadly, this example illustrates the important lesson that both time scales are equally significant from an evolutionary perspective, and one is not more "fundamental" than the other when it comes to understanding the dynamical behavior of the system. Furthermore, it demonstrates that pressures for natural selection can operate at many different levels of complexity.

5.4.8.2.3 The Evolution of Proteins By making use of simple physical models of proteins, it is possible to model evolution under different evolutionary, structural, and functional scenarios. For example, cubic lattice models of proteins can be used to model enzyme evolution involving binding to two hydrophobic substrates. Gene duplication coupled to subfunctionalization can be used to predict enzyme gene duplicate retention patterns and compare with genomic data.[113] This type of physical modeling can be expanded to other evolutionary models, including those that incorporate positive selective pressures or that vary population genetic parameters. At a structural level, they can be used to address issues of protein surface-area-to-volume ratios or the evolvability of different folds. Ultimately, such models can be extended to real protein shapes and can be correlated to the evolution of different folds in real genomes.[114]

The role of structure in evolution during potentially adaptive periods can also be analyzed. A subset of positive selection will be dictated by structural parameters and intramolecular coevolution. Common interactions, like RKDE ionic interactions can be detected in this manner. Similarly, less common interactions, like cation-p interactions, can also be detected and the interconversion between different modes of interactions can be assessed statistically.

One important tool underlying these efforts is the Adaptive Evolution Database (TAED), a phylogenetically organized database that gathers information related to coding sequence evolution.[115] This database is designed to both provide high-quality gene families with multiple sequence alignments and phylogenetic trees for chordates and embryophytes and to enable answers to the question, "What makes each species unique at the molecular genomic level?"

Starting with GenBank, genes have been grouped into families, and multiple sequence alignments and phylogenetic trees have been calculated. In addition to multiple sequence alignments and phylogenetic trees for all families of chordate and embryophyte sequences, TAED includes the ratio of nonsynonymous to synonymous nucleotide substitution rates (K_a/K_s) for each branch of every phylogenetic tree. This ratio, when significantly greater than 1, is an indicator of positive selection and potentially a change of function of the encoded protein in closely related species, and has been useful in the construction of phylogenetic trees with probabilistic reconstructed ancestral sequences calculated using both parsimony and maximum likelihood approaches. With a mapping of gene tree to species tree, the branches whose ratio is significantly greater than 1 are collated together in a phylogenetic context.

[113] F.N. Braun and D.A. Liberles, "Retention of Enzyme Gene Duplicates by Subfunctionalization;" *International Journal of Biological Macromolecules* 33(1-3):19-22, 2003.

[114] H. Hegyi, J. Lin, D. Greenbaum, and M. Gerstein, "Structural Genomics Analysis: Characteristics of Atypical, Common, and Horizontally Transferred Folds," *Proteins* 47(2):126-141, 2002.

[115] D.A. Liberles, "Evaluation of Methods for Determination of a Reconstructed History of Gene Sequence Evolution." *Molecular Biology and Evolution* 18(11):2040-2047, 2001; D.A. Liberles, D.R. Schreiber, S. Govindarajan, S.G. Chamberlin, and S.A. Benner, "The Adaptive Evolution Database (TAED)," *Genome Biology* 2(8):research0028.1-0028.6, 2001; C. Roth, M.J. Betts, P. Steffansson, G. Sælensminde, and D.A. Liberles, "The Adaptive Evolution Database (TAED): A Phylogeny-based Tool for Comparative Genomics," *Nucleic Acids Research* 33(Database issue):D495-D497, 2005.

The TAED framework is expandable to incorporate other genomic-scale information in a phylogenetic context. This is important because coding sequence evolution (e.g., as reflected in the K_a/K_s ratio) is only one part of the molecular evolution of genomes driving phenotypic divergence. Changes in gene content[116] and phylogenetic reconstructions of changes in gene expression and alternative splicing data[117] can indicate where other significant lineage-specific changes have occurred. Altogether, phylogenetic indexing of genomic data presents a powerful approach to understanding the evolution of function in genomes.

5.4.8.2.4 The Emergence of Complex Genomes How did life get started on Earth? Today, life is based on DNA genomes and protein enzymes. However, biological evidence exists to suggest that in a previous era, life was based on RNA, in the sense that genetic information was contained in RNA sequences and phenotypes were expressed as catalytic properties of RNA.[118]

An interesting and profound issue is therefore to understand the transition from the RNA to the DNA world, one element of which is the fact that DNA genomes are complex structures. In 1971, Eigen found an explicit relationship between the size of a stable genome and the error rate inherent in its replication, specifically that the size of the genome was inversely proportional to the per-nucleotide replication error rate.[119] Thus, for a genome of length L to be reasonably stable over successive generations, the maximum tolerable error rate in replication could be no more than $1/L$ per nucleotide. However, more precise replication mechanisms tend to be more complex. Given that the replication mechanism must itself be represented in the genome, the puzzle is that a precise replication mechanism is needed to maintain a complex genome, but a complex genome is required to encode such a mechanism.

The only possible answer to this puzzle is that complex genomes evolved from simpler ones. Szabó et al. investigated this possibility through computer simulations.[120] They constructed a population of digital genomes subject to evolutionary forces and found that under a certain set of circumstances, both genome size and replication fidelity increased with the run time of the simulation. However, such behavior was dependent on the existence of a sufficient amount of spatial isolation of the evolving population. In the absence of separation (i.e., in the limit of very rapid diffusion of genomes across the two-dimensional surface to which they were confined), genome complexity and replication fidelity were both limited. However, if diffusion is slow (i.e., the characteristic time constant of diffusion is less than the time scale of replication), both complexity and fidelity increase.

In addition, Johnston et al. have synthesized in the laboratory a catalytic RNA molecule that contains about 200 nucleotides and synthesizes RNA molecules of up to 14 nucleotides, with an error rate of about 3 percent per residue.[121] This laboratory demonstration, coupled with the computational finding described above, suggest that a small RNA genome that operates as an RNA replicase with

[116] E.V. Koonin, N.D. Fedorova, J.D. Jackson, A.R. Jacobs, D.M. Krylov, K.S. Makarova, R. Mazumder, et al., "A Comprehensive Evolutionary Classification of Proteins Encoded in Complete Eukaryotic Genomes," *Genome Biology* 5(2):R7, 2004. (Cited in Roth et al., "The Adaptive Evolution Database," 2005.)

[117] R. Rossnes, "Phylogenetic Reconstruction of Ancestral Character States for Gene Expression and mRNA Splicing Data," M.Sc. thesis, Universtiy of Bergen, Norway, 2004. (Cited in Roth et al., 2005.)

[118] See, for example, G.F. Joyce, "The Antiquity of RNA-based Evolution," *Nature* 418(6894):214-221, 2002.

[119] M. Eigen, "Selforganization of Matter and the Evolution of Biological Macromolecules," *Naturwissenschaften* 58(10):465-523, 1971.

[120] P. Szabó, I Scheuring, T. Czaran, and E. Szathmary, "In Silico Simulations Reveal That Replicators with Limited Dispersal Evolve Towards Higher Efficiency and Fidelity," *Nature* 420(6913):340-343, 2002. A very helpful commentary on this article can be found in G.F. Joyce, "Molecular Evolution: Booting Up Life," *Nature* 420(6894):278–279, 2002. The discussion in Section 5.4.8.2.4 is based largely on this article.

[121] W.K. Johnston, P.J. Unrau, M.S. Lawrence, M.E. Glasner, and D.P. Bartel, "RNA-catalyzed RNA Polymerization: Accurate and General RNA-Templated Primer Extension," *Science* 292(5520):1319-1325, 2001.

modest efficiency and fidelity could evolve a succession of ever-larger genomes and ever-higher replication efficiencies.

5.4.8.3 Examples from Ecology[122]

Simulation-based study of an ecosystem considers the dynamic behavior of systems of individual organisms as they respond to each other and to environmental stimuli and pressures (e.g., climate) and examines the behavior of the ecosystem in aggregate terms. However, no individual run of such a simulation can be expected to predict the detailed behavior of each individual organism within an ecosystem. Rather, the appropriate test of a simulation's fidelity is the extent to which it can, through a process of judicious averaging of many runs, predict features that are associated with aggregation at many levels of spatial and/or temporal detail. These more qualitative features provide the basis for descriptions of ecosystem dynamics that are robust across a variety of dynamical scenarios that are different at a detailed level and also provide high-level descriptions that can be more readily interpreted by researchers.

Because of the general applicability of the approach described above, simulations of dynamical behavior can be developed for aggregations of any organisms as long as they can be informed by adequate understandings of individual-level behavior and the implications of such behavior for interactions with other individuals and with the environment.

Note also the key role played by ecosystem heterogeneity. Spatial heterogeneity is one obvious way in which nonuniform distributions play a role. But in biodiversity, functional heterogeneity is also important. In particular, essential ecosystem functions such as the maintenance of fluxes of certain nutrients and pollutants, the mediation of climate and weather, and the stabilization of coastlines may depend not on the behavior of all species within the ecosystem but rather on a limited subset of these species. If biodiversity is to be maintained, the most fragile and functionally critical subsets species must be identified and understood.

The mathematical and computational challenges range from techniques for representing and accessing datasets, to algorithms for simulation of large-scale spatially stochastic, multivariate systems, to the development and analysis of simplified description. Novel data acquisition tools (e.g., a satellite-based geographic information system that records changes for insertion in the simulations) would be welcome in a field that is relatively data poor.

5.4.8.3.1 Impact of Spatial Distribution in Ecosystems An important dimension of ecological environments is how organisms interact with each other. One often-made computationally simple assumption is that an organism is equally likely to interact with every other organism in the environment. Although this is a pragmatic assumption, actual ecosystems are physical and organisms interact only with a very small number of other organisms—namely, the ones that are nearby in a spatial sense. Moreover, localized selection—in which a fitness evaluation is undertaken only under nearest neighbors—is also operative.

Introducing these notions increases the speciation rate tremendously, and the speculation is that in a nonlocalized environment, the pressures on the population tend toward population uniformity—everything looks similar, because each entity faces selection pressure from every other entity. When localization occurs, different species emerge in different spatial areas. Further, the individuals that are evolving will start to look quite different from each other, even though they have (comparably) high

[122] Section 5.4.8.3 is based largely on material taken from S.A. Levin, B. Grenfell, A. Hastings, and A.S. Perelson, "Mathematical and Computational Challenges in Population Biology and Ecosystems Science," *Science* 275(5298):334-343, 1997.

fitness ratings. (This phenomenon is known as convergent evolution, in which a given environment might evolve several different species that are in some sense equally well adapted to that environment.)

As an example of spatial localization, Kerr et al. developed a computational model to examine the behavior of a community consisting of three strains of *E. coli*,[123] based on a modification of the lattice-based simulation of Durrett and Levin.[124] One of the strains carried a gene that created an antibiotic called colicin. (The colicin-producing strain, C, was immune to the colicin it produced.) A second strain was sensitive to colicin (S), while a third strain was resistant to colicin (R). Furthermore, the factors that make the S strain sensitive also facilitate its consumption of certain nutrients, and the R strain is less able to consume these nutrients. However, because the R strain does not have to produce colicin, it avoids a metabolic cost incurred by the C strain. The result is that C bacteria kill S bacteria, S bacteria thrive where R bacteria do not, and R bacteria thrive where C bacteria do not. The community thus satisfies a "rock-paper-scissors" relationship.

The intent of the simulation was to explore the spatial scale of ecological processes in a community of these three strains. It was found found (and confirmed experimentally) that when dispersal and interaction were local, patches of different strains formed, and these patches chased one another over the lattice—type C patches encroached on S patches, S patches displaced R patches and R patche invaded C patches. Within this mosaic of patches, the local gains made by any one type were soon enjoyed by another type; hence the diversity of the system was maintained. However, dispersal and interaction were no longer exclusively local (i.e., in the "well-mixed" case in which all three strains are allowed to interact freely with each other): continual redistribution of C rapidly drove S extinct, and R then came to dominate the entire community

5.4.8.3.2 Forest Dynamics[125] To simulate the growth of northeastern forests, a stochastic and mechanistic model known as SORTIE has been developed to follow the fates of individual trees and their offspring. Based on species-specific information on growth rates, fecundity, mortality, and seed dispersal distances, as well as detailed, spatially explicit information about local light regimes, SORTIE follows tens of thousands of trees to generate dynamic maps of distributions of nine dominant or subdominant species of tree that look like real forests and match data observed in real forests at appropriate levels of spatial resolution. SORTIE predicts realistic forest responses to disturbances (e.g., small circles within the forest boundaries within which all trees are destroyed), clear-cuts (i.e., large disturbances), and increased tree mortality.

SORTIE consists of two units that account for local light availability and species life history for each of nine tree species. Local light availability refers to the availability of light at each individual tree. This is a function of all of the neighboring trees that shade the tree in question. Information on the spatial relations among these neighboring tree crowns is combined with the movement of the sun throughout the growing season to determine the total, seasonally averaged light expressed as a percentage of full sun. In other words, the growth of any given tree depends on the growth of all neighboring trees.

The species life history (available for each of nine tree species) provides the relationship between radial growth rates as a function of its local light environment and is based on empirically estimated life-history information. Radial growth predicts height growth, canopy width, and canopy depth in accordance with estimated allometric relations. Fecundity is estimated as an increasing power function of tree size, and seeds are dispersed stochastically according to a relation whereby the probability of

[123] B. Kerr, M.A. Riley, M.W. Feldman, and B.J. Bohannan, "Local Dispersal Promotes Biodiversity in a Real-life Game of Rock-Paper-Scissors," *Nature* 418(6894):171-174, 2002.

[124] R. Durrett and S. Levin, "Allelopathy in Spatially Distributed Populations," *Journal of Theoretical Biology* 185(2):165-171, 1997.

[125] Section 5.4.8.3.2 is based largely on D.H. Deutschman, S.A. Levin, C. Devine, and L.A. Buttel, "Scaling from Trees to Forests: Analysis of a Complex Simulation Model," *Science Online* supplement to *Science* 277(5332), 1997, available at http://www.sciencemag.org/content/vol277/issue5332. *Science Online* article available at http://www.sciencemag.org/feature/data/deutschman/home.htm.

dispersal declines with distance. Mortality risk is also stochastic and has two elements: random mortality and mortality associated with suppressed growth.

Because SORTIE is intended to aggregate statistical properties of forests, an ensemble of simulation runs is necessary, in which different degrees of smoothing and aggregation are used to determine how much information is lost by averaging and to find out where error is compressed and where it is enlarged in the course of this process. SORTIE is a computation-intensive simulation even for individual simulations, because multiple runs are needed to generate the necessary ensembles for statistical analysis. In addition, simulations carried out for heterogeneous environments require an interface between large dynamic simulations and geographic information systems, providing real-time feedbacks between the two.

5.5 TECHNICAL CHALLENGES RELATED TO MODELING

A number of obstacles and difficulties must be overcome if modeling is to be made useful to life scientists more broadly than is the case today. The development of a sophisticated computational model requires both a conceptual foundation and implementation. Challenges related to conceptual foundations can be regarded as mathematical and analytical; challenges related to implementation can be regarded as computational or, more precisely, as related to computer science (Box 5.24).

Today's mathematical tools for modeling are limited. Nonlinear dynamics and bifurcation theory provide some of the most well-developed applied mathematical techniques and offer great successes in illuminating simple nonlinear systems of differential equations. But they are inadequate in many situations, as illustrated by the fact that understanding global stability in systems larger than four equations is prohibitively hard, if not unrealistic. Visualization of high-dimensional dynamics is still problematic in computational as well as analytical frameworks; the question remains as to how to represent such complex dynamics in the best, most easily understood ways. Moreover, many high-dimensional systems have effectively low-dimensional dynamics. A challenge is to extract the dynamical behavior from the equations without first knowing what the low-dimensional subspace is. Box 5.25 describes one new and promising approach to dealing with high-dimensional multiscale problems.

Other mathematical methods and new theory will be needed to find solutions that apply not only to biological problems, but to other scientific and engineering applications as well. These include methods for global optimization and for reverse engineering of structure (of any "black box," be it a network of genes, a signal transduction pathway, or a neuronal system) based on data elicited in response to stimuli and perturbations.

Identification of model structure and parameters in nonlinear systems is also nontrivial. This is especially true in biological systems due to incomplete knowledge and essentially limitless types of interactions. Decomposition of complex systems into simpler subsystems ("modules") is an important challenge to our ability to analyze and understand such systems (a point discussed in Chapter 6). Development of frameworks to incorporate moving boundaries and changing geometries or shapes is essential to describing biological systems. This is traditionally a difficult area. Ideally, it would be desirable to be able to synthesize and analyze models that have nonlinear deterministic as well as stochastic elements, and continuous as well as discrete, algebraic constraints, with other more traditional nonlinear dynamics. (See Section 5.3.2 for greater detail.) All of these can be viewed as challenges in nonlinear dynamics aspects of modeling.

Further developing both computational (numerical simulation) methods and analytical methods (bifurcation, perturbation methods, asymptotic analysis) for large nonlinear systems will invariably mean great progress in the ability to build more elaborate and detailed models. However, with these large models come large challenges. One is how to find methodical ways of organizing parameter space exploration for systems that have numerous parameters. Another is the development of ways to codify and track assumptions that have gone into the construction of a model. Understanding these assumptions (or simplifications) is essential to understanding the limitations of a model and when its predictions are no longer biologically relevant.

Box 5.24
Modeling Challenges for Computer Science

Integration Methods

- Methods for integrating dissimilar mathematical models into complex and integrated overall models
- Tools for semantic interoperability

Models

- High-performance, scalable algorithms for network analyses and cell modeling
- Methods to propagate measures of confidence from diverse data sources to complex models

Validation

- Robust model and simulation-validation techniques (e.g., sensitivity analyses of systems with huge numbers of parameters, integration of model scales)
- Methods for assessing the accuracy of genome-annotation systems

SOURCE: U.S. Department of Energy, *Report on the Computer Science Workshop for the Genomes to Life Program*, Gaithersburg, MD, March 6-7, 2002, available at http://DOEGenomesToLife.org/compbio/.

Box 5.25
Equation-free Multiscale Computation: Enabling Microscopic Simulators to Perform System-level Tasks

Yannis Kevrikides of Princeton University and his colleagues have developed a framework for computer-aided multiscale analysis. This framework enables models at a "fine" (microscopic, stochastic) level of description to perform modeling tasks at a "coarse" (macroscopic, systems) level. These macroscopic modeling tasks, yielding information over long time and large space scales, are accomplished through appropriately initialized calls to the microscopic simulator for only short times and small spatial domains: "patches" in macroscopic space-time.

In general, traditional modeling approaches require the derivation of macroscopic equations that govern the time evolution of a system. With these equations in hand (usually partial differential equations (PDEs)), a variety of analytical and numerical techniques for their solution is available. The framework of Kevrikides and colleagues, known as the equation-free (EF) approach can, when successful, bypass the derivation of the macroscopic evolution equations when these equations conceptually exist but are not available in closed form.

The advantage of this approach is that the long-term behavior of the system bypasses the computationally intensive calculations needed to solve the PDEs that describe the system. That is, the EF approach enables an alternative description of the physics underlying the system at the microscopic scale (i.e., its behavior on relatively short time and space scales) provide information about the behavior of the system over relatively large time and space scales directly without expensive computations. In effect, the EF approach constitutes a systems identification-based, "closure on demand" computational toolkit, bridging microscopic-stochastic simulation with traditional continuum scientific computation and numerical analysis.

SOURCE: The EF approach was first introduced by Yannis Kevrikides and colleagues in K. Theodoropoulos et al., "Coarse Stability and Bifurcation Analysis Using Timesteppers: A Reaction Diffusion Example," *Proceedings of the National Academy of Sciences* 97:9840, 2000, available at http://www.pnas.org/cgi/reprint/97/18/9840.pdf. The text of this box is based on excerpts from an abstract describing a presentation by Kevrikides on April 16, 2003, to the Singapore-MIT Alliance program on High Performance Computation for Engineered Systems (HPCES); abstract available at http://web.mit.edu/sma/events/seminar/kevrekidis.htm.

In the second category, issues related to implementing the model arise. Often such issues involve the actual code used to implement the model. Computational models are, in essence, large computer programs; issues of software development come to the fore. As the desire for and utility of computational modeling increase, the needs for software are growing rather than diminishing as hardware becomes more capable. On the other hand, progress in software development and engineering over the last several decades has not been nearly as dramatic as progress in hardware capability, and there appears to be no magic bullets on the horizon that will revolutionize the software development process.

This is not to say that good software engineering does not or should not play a role in the development of computational models. Indeed, the Biomedical Information Science and Technology Initiative (BISTI) Planning Workshop of January 15-16, 2003, explicitly recommended that NIH require grant applications, proposing research in bioinformatics or computational biology to adopt as appropriate, accepted practices of software engineering.[126] Section 4.5 describes some of the elements of good software engineering in the context of tool development, and the same considerations apply to model development.

A second important challenge as large simulation models become more prevalent is a standard specification language to unambiguously specify the model, its parameters, annotations, and even the means by which it is to be scored against data. The challenge will be to provide a language flexible enough to capture all interesting biological processes and incorporate models at different levels of abstraction and in different mathematical paradigms, including stochastic differential, partial differential, algebraic, and discrete equations. It may prove necessary to develop a set of nested languages—for example, a language that specifies the biological process at a very high level and a linked language that specifies the mathematical representation of each process. There are some current attempts at these languages based on the XML framework. SBML and CellML are attempts in this direction.

Finally, many biological modeling applications involve a problem space that is not well understood and may even be intended to explore queries that are not well formulated. Thus, there is a high premium on reducing the labor and time involved to produce an application that does something useful. In this context, technologies for "rapid prototyping" of biological models have considerable interest.[127]

[126] See http://www.bisti.nih.gov/2003meeting/report.cfm.

[127] Note, however, that in the rapid prototyping process often used to create commercial applications, there is a dialogue between developer and user that reveals what the user would find valuable: once the developer knows what the user really wants, the software development effort is straightforward. By contrast, in biological applications, it is nature that determines the appropriate structuring and formulation of a problem, and a problem cannot be structured in a certain way simply because it is convenient to do so. Thus, technologies for the rapid prototyping of biological models must afford the ability to rearrange model components and connections between components with ease.

6

A Computational and Engineering View of Biology

Because 21st century biology is very concerned with function, it is helpful to have abstractions available that characterize the functionality of interest. By doing so, insights derived from study of those abstractions in other contexts become available for biological use. In addition, because biological systems are the products of eons of evolutionary history and decision making, viewing them through the lens of engineering yields insights that are not otherwise available from an analysis that might be based on first principles.

6.1 BIOLOGICAL INFORMATION PROCESSING[1]

As noted in Chapter 2, biological systems are extraordinarily complex—and partly as a consequence, poorly understood. Yet it is clear that biological systems demonstrate and exemplify functionality at different levels.

Artifacts such as computer hardware and software also exhibit functionality and multiple levels. To facilitate the understanding and construction of such artifacts, computer science has developed information abstractions that seek to capture and encapsulate certain kinds of functional behavior in manipulating and managing information; such abstractions are a primary focus of study of the computer scientist (Box 6.1).

One key connection to 21st century biology is that many biological problems now require the simultaneous consideration of phenomena at different scales. For example, biologists can think of genetics at the level of individual nucleotides, at the level of chromosomes, at the level of genomes, and at the level of populations. From nucleotide to population is a span of many orders of magnitude, and it is difficult to conceptualize such a range without moving seamlessly between different levels of abstraction.

Section 6.1 describes several such abstractions and their specific biological applications already in use, but the description is not intended to be exhaustive, and there are likely many more such abstractions capable of providing biological insight, including new or as yet undiscovered techniques or concepts. As such, this area represents opportunities for both biologists and computer scientists.

[1] Much of the discussion in Section 6.1 about cells as information-processing devices is adapted from R. Aviv and E. Shapiro, "Cellular Abstractions: Cells as Computation," *Nature* 419:343, 2002.

Box 6.1
On the Abstractions of the Computer Scientist and Engineer

Abstraction is a generic technique that allows the scientist or engineer to focus only on certain features of a system while hiding others. Scientists in all disciplines typically use abstractions as a way to simplify calculations for purposes of analysis, but computer scientists also use abstractions for purposes of design: to build working computer systems. Because building systems is the central focus of much work in computer science, the use of abstractions to cope with complexity over a wide range of scale, size, and levels of detail is central to a computer scientist's way of thinking.

The focus of the computer scientist in creating an abstraction is to hide the complexity of operation "underneath the abstraction" while offering a simple and useful set of services "on top of it." Using such abstractions is the principal technique for organizing and constructing very sophisticated computer systems, and they enable computer scientists to deal with large differences of scale. For example, one particularly useful abstraction uses hardware, system software, and application software as successive layers on which useful computer systems can be built. This illustrates one very important use of abstraction in computer systems: each layer provides the capability to specify that a certain task be carried out without specifying *how* it should be carried out. In general, computing artifacts embody many different abstractions that capture many different levels of detail.

A good abstraction is one that captures the important features of an artifact and allows the user to ignore the irrelevant ones. (The features decided to be important collectively constitute the interface of the artifact to the outside world.) By hiding details, an abstraction can make working with an artifact easier and less subject to error. But hiding details is not cost-free—in a particular programming problem, access to a hidden detail might in fact be quite helpful to the person who will use that abstraction. Thus, deciding how to construct an abstraction (i.e., deciding what is important or irrelevant) is one of the most challenging intellectual issues in computer science. A second challenging issue is how to manage all of the details that are hidden. The fact that they are hidden beneath the interface does not mean that they are irrelevant, only that the computer scientist must design and implement approaches to handle these details "automatically" (i.e., without external specification).

SOURCE: Adapted from Computer Science and Telecommunications Board, National Research Council, *Computing the Future: A Broader Agenda for Computer Science and Engineering,* National Academy Press, Washington, D.C., 1991.

Consider that biological processes, such as catalysis, protein synthesis, and other metabolic systems, are consumers, processors, or creators of information. As Loewenstein puts it, in biological systems, "in addition to flows of matter and energy, there is also flow of information. Biological systems are information-processing systems and this must be an essential part of any theory we may construct."[2] Sydney Brenner goes farther, arguing that ". . . this information flow, not energy per se, is the prime mover of life—that molecular information flowing in circles brings forth the organization we call 'organism' and maintains it against the ever-present disorganizing pressures in the physics universe. So viewed, the information circle becomes the unit of life."[3]

The current state of intellectual affairs with respect to biological information and complexity may have some historical analogy with the concept of energy at the beginning of the 19th century. Although the concept was intuitively obvious, it was not formally defined or measured at that time. Carnot's analysis of the performance of steam engines formalized the meaning of energy, creating the basis for

[2] W. Loewenstein, *The Touchstone of Life: Molecular Information, Cell Communication, and the Foundations of Life*, Oxford University Press, New York, 1998, p. xiv.

[3] S. Brenner, "Theoretical Biology in the Third Millennium," *Philosophical Transactions of the Royal Society B* 354(1392):1963-1965, 1999.

the science of thermodynamics. Only after energy had been identified and studied in the artificial realm of steam engines was it recognized as a prime aspect of natural systems as well.

Similarly, the existing state of the theory of biological information (or, indeed, information of any sort) is based on the work of Claude Shannon, who studied the processing of information in human technological channels of communication, and the field of computational complexity, which was created to analyze the performance characteristics of algorithms running on human-built computers. However, just as thermodynamics successfully widened its scope to the natural world from steam engines, information and computation theory may become a powerful lens for describing, measuring, and understanding processes in the natural world.

Biological information is likely to have a close relationship to information in the Shannon sense of the term, if only because biological entities depend on information to coordinate their internal activity. Cells coordinate their internal activity because they have harnessed intracellular Shannon information channels. Multicellular organisms coordinate their internal activity because they have harnessed intercellular Shannon information channels. These channels are the conduits through which genes transfer their information content to proteins, proteins serve as signaling agents, and nervous systems work. Also, Shannon's insight about the nature of information transmission allows us to understand how signals can reliably be sent through a noisy unpredictable environment (whether cell telephone signals, Internet packets, or hormone signaling proteins) and received accurately at the other end.

On the other hand, Shannon information applies in the strict sense only when it is possible to identify a sender and receiver connected by a channel. There are some places in which this applies, such as the projection of the retina to the brain. Yet in the context of information feedback and loops rather than channels, it is not clear that Shannon information continues to have a well-defined meaning.

There have been a number of attempts to generalize Shannon information to problems at the cellular and subcellular levels, of which the conceptualization by Manfred Eigen of hypercycles, quasi-species, and sequence space is one of the most notable.[4] But whether these concepts are the right ones is not as important as the recognition that new concepts are needed.

A more specific connection between biology and computation can be seen in the biological use of information to enhance the survival and reproductive functions of an organism. That is, biological organisms use information about the environment to stimulate or drive responses that boost the likelihood of survival and successful reproduction. This process is effectively a computation that transforms the inputs (which describe environmental conditions) into the appropriate outputs (the organism's behavior).[5] For example, Hartwell et al. note that signals from the environment entrain circadian biological clocks to produce responses to predicted fluctuations in light intensity and temperature.[6]

Embedded within cells are complex signaling mechanisms that transfer information from one part of a cell to another and intercellular mechanisms that transfer information from one part of a multicellular organism to another. Indeed, signal transduction pathways—and the proteins associated with them—appear to serve the functions of information processing and transfer,[7] rather than those of more "traditional" biology (e.g., chemical transformation of metabolic intermediates or the building of cellular structures).

[4] M. Eigen, "The Origin of Biological Information," presented at the Seventh International Conference on Intelligent Systems for Molecular Biology, August 6-10, 1999; Heidelberg, Germany, available at http://bioinf.mpi-sb.mpg.de/conferences/ismb99/WWW/abstracts/abs-eigen.html.

[5] Indeed, it has been asserted that the history of life can be described as the evolution of systems that manipulate one set of symbols representing inputs into another set of symbols that represent outputs. J.J. Hopfield, "Physics, Computation, and Why Biology Looks So Different," *Journal of Theoretical Biology* 171:53-60, 1994.

[6] L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl):C47-C52, 1999.

[7] D. Bray, "Protein Molecules as Computational Elements in Living Cells," *Nature* 376(6538):307-312, 1995. The examples in the next paragraph are also Bray's.

For example, a simple enzyme protein could be viewed as a computational element that takes an input—the concentration of its "substrate," the molecule with which it interacts—and produces an output: a concentration of the catalyzed reaction product. An enzyme that becomes active only when it binds to two separate regulator molecules will function something like a Boolean AND gate, and so on. Circuits formed from these elements can be as simple as a switch or an oscillator, or as complex as to drive a bacterium's chemotaxis response. Indeed, the cell even possesses a kind of short-term, "random-access" memory, in the sense that events in its environment have profoundly shaped the concentration and activity of many thousands of molecules in the cell. In short, these protein-based circuits constitute a kind of nervous system for the cell, providing it with much of what it needs to control its behavior. Box 6.2 provides some additional perspective on this subject.

Additional insights can be gained from the notion that both computational processes and biological pathways can be viewed as processes that affect the state of a system according to well-defined (though possibly probabilistic) rules. Thus, it is possible to describe regulatory, metabolic, and signaling pathways, as well as multicellular processes such as immune responses, as systems of interacting computations operating in parallel. In particular, languages such as Petrinets, Statecharts (discussed in Section 4.3.1), and the Pi-calculus, originally developed for the specification and study of systems of interacting computations, can be used to represent such systems.[8] Such representations enable researchers to simulate their behavior, and to support qualitative and quantitative reasoning on the properties of these systems.

To cite two prominent researchers in this area:

> Processes, the basic interacting computational entities of these languages, have an internal state and interaction capabilities. Process behavior is governed by reaction rules specifying the response to an input message based on its content and the state of the process. The response can include state change, a change in interaction capabilities, and/or sending messages. Complex entities are described hierarchically—for example, if a and b are abstractions of two molecular domains of a single molecule, then (a parallel b) is an abstraction of the corresponding two-domain molecule. Similarly, if a and b are abstractions of the two possible behaviors of a molecule in one of two conformational states, depending on the ligand it binds, then (a choice b) is an abstraction of the molecule, with the choice between a and b determined by its interaction with a ligand process.[9]

Abstractions of the cell as a computing or information-processing device allow one to distinguish between two conceptual levels: a "low-level" view that focuses on implementation (i.e., how the system is built—where the wires go or the detailed molecular processes involved) and a "high-level" view that focuses on functionality (what the system does—analogous to a logic gate or a computational device).[10] For example, one might distinguish between the pathways involved in regulating the circadian rhythm of an organism and its functional behavior as an oscillator.

The difference between these levels of abstraction enables biologically significant comparisons to be made. For example, it would be instructive if two different organisms implemented the same function in different ways. In other words, functional equivalence between related implementations in different organisms could be regarded as a measure of the behavioral similarity of entire systems. (In the literature of evolutionary biology, the implementation of the same function in different ways is called "analogous" implementation.) Perhaps more importantly, a functional perspective is an enabler for the integration of knowledge about the function, activity, and interaction of cellular molecular systems.

[8] R. Aviv and E. Shapiro, "Cellular Abstractions: Cells as Computation," *Nature* 419:343, 2002.

[9] R. Aviv and E. Shapiro, "Cellular Abstractions," 2002.

[10] In many circumstances, different parts of a biological system may play different roles at different times or even different roles at different time scales at the same time. This is especially true in splicing variants, where the expression of a gene may produce proteins with quite different functions according to the behavior of the splicing mechanism. Indeed, in some cases, different splicings have opposite functions. Nevertheless, in understanding a given role at a given time and time scale, the high-level abstraction focused on functionality is meaningful and scientifically significant.

Box 6.2
Role of Computation in Complex Regulatory Networks

Computation . . . [is] a crucial ingredient when dealing with the description of biocomplexity and its evolution, because it turns out to be much more relevant than the underlying physics. Its dynamics is governed mainly by the transmission, storage and manipulation of information, a process which is highly nonlinear. This nonlinearity is well illustrated by the nature of signaling in cells: local events involving a few molecules can produce a propagating cascade of signals through the whole system to yield a global response. . . . If we try to make predictions about the outcomes of these signaling events in general, we are faced with the inherent unpredictability of computational systems. It is at this level where computation becomes central and where idealized models of regulatory networks seem appropriate enough to capture the essential features at the global scale.

Cells are probably the most complete example of this traffic of signals at all levels. . . . The cellular network can be divided into three major self-regulated sub-webs:

- The *genome*, in which genes can affect each other's level of expression;
- The *proteome*, defined by the set of proteins and their interactions by physical contact; and
- The metabolic network (or the *metabolome*), integrated by all metabolites and the pathways that link each other.

All these subnetworks are very much intertwined since, for instance, genes can only affect other genes through special proteins, and some metabolic pathways, regulated by proteins themselves, may be the very ones to catalyze the formation of nucleotides, in turn affecting the process of translation. . . . It is not difficult to appreciate the enormous complexity that these networks can achieve in multicellular organisms, where large genomes have structural genes associated with at least one regulatory element and each regulatory element integrates the activity of at least two other genes. . . .

Luckily, all this extraordinary complexity can be abstracted, at least at some levels, to simplified models which can help in the study of the inner-workings of cellular networks. Overall, irrespective of the particular details, biological systems show a common pattern: some low-level units produce complex, high-level dynamics coordinating their activity through local interactions. Thus, despite the many forms of interaction found at the cellular level, all come down to a single fact: the state of the elements in the system is a function of the state of the other elements it interacts with. What models of network functioning try, therefore, is to understand the basic properties of general systems composed of units whose interactions are governed by nonlinear functions. These models, being simplifications, do not allow one to make predictions at the level of the precise state of particular units. Their average overall behavior, however, can shed light into the way real cells behave as a system. . . .

. . . [M]any entities in cellular networks can be identified as the basic units of regulation, mainly distinguished by their unique roles with respect to interaction with other units. These basic units are genes, each of the proteins that the genes can produce, each of the forms of a protein, protein complexes, and all related metabolites. These units have associated values that either represent concentrations or levels of activation. Their values depend on the values of the units that affect them due to the mechanisms discussed, plus some parameters that govern each special form of interaction. . . . Computer modeling of [the] network [the segment polarity network of *Drosophila melanogaster*] has provided insight into various questions. A very important result is the fact that this network seems to be a conserved module. Evidence for this has been obtained by simulations demonstrating its robustness against the change of parameters. . . .

SOURCE: Reprinted from P. Fernandez and R.V. Sole, "The Role of Computation in Complex Regulatory Networks," Santa Fe Institute Working Paper, 2003, available at http://www.santafe.edu/sfi/publications/Working-Papers/03-10-055.pdf; to appear in a chapter in *Power Laws, Scale-Free Networks and Genome Biology,* Landes Bioscience. Reprinted with permission.

This perspective on cells as computational devices should not be taken as an argument that cells process information the way a digital computer does. The organizations are radically different. To name just a few differences, in a cell there is no clean separation between the data store and the central processing unit: the cell's memory is the same protein reaction network that does its processing. Real proteins rarely respond or act in a completely binary fashion—the levels of concentration matter. Apart from DNA, few portions of a cell's internal machinery are explicitly digital in nature—with the result that signaling in a cell must take place in a highly noisy environment.

It is also interesting that biological function often relies on what might be called exploration with selection—the production of many intermediate products resulting from stochastic subprocesses that are then refined to unique and appropriate solutions.[11] Taken across the entire population, exploration with selection exploits the difference between creating a solution and testing a solution for correctness—the first being in general a much more difficult computational task than the second.[12] Random processes are used to explore the space of possible solutions,[13] and other machinery culls these possible solutions. As Hartwell et al. argue, "Similar messy and probabilistic intermediates appear in engineering systems based on artificial neural networks—mathematical characterizations of information processing that are directly inspired by biology. A neural network can usefully describe complicated deterministic input-output relationships, even though the intermediate calculations through which it proceeds lack any obvious meaning and their choice depends on random noise in a training process."[14]

6.2 AN ENGINEERING PERSPECTIVE ON BIOLOGICAL ORGANISMS

6.2.1 Biological Organisms as Engineered Entities

Engineering insights can be useful in understanding biological organisms as engineered entities, and the rationale for seeking insights from engineering is based on three notions. First, although the physical scales may differ in some cases, human technology and natural systems operate in the same world and must obey the same physical rules. Knowledge that engineering fields have accumulated about what techniques work and the limits of those techniques can serve as a potentially valuable guide in investigating the physical basis of the operations of natural systems. This is especially true for biomechanical feats, such as structural support, locomotion, circulation, and so on.

The second rationale is that because evolution and a long history of environmental accidents have driven processes of natural selection, biological systems are more properly regarded as engineered artifacts than as objects whose existence might be predicted on the basis of the first principles of physics, although the evolutionary context means that an artifact is never "finished" and is rather evaluated on a continuous basis.[15] Both engineered artifacts and biological organisms demonstrate function, embody

[11] For example, the immune system relies on the random generation of pathogen detectors, which are then eliminated when they match some definition of "self." In single molecules, kinetic funnels direct different molecules of the same protein through multiple, different paths from the denatured state to a unique folded structure (K.A. Dill and H.S. Chan, "From Levinthal to Pathways to Funnels," *Nature Structural Biology* 4:10-19, 1997). Within cells, the shape of the mitotic spindle is due partly to selective stabilization of randomly generated microtubules whose ends happen to be close to a chromosome (R. Heald, R. Tournebize, T. Blank, R. Sandaltzopoulos, P. Becker, A. Hyman, and E. Karsenti, "Self-organization of Microtubules into Bipolar Spindles Around Artificial Chromosomes in *Xenopus* Egg Extracts," *Nature* 382(6590):420-425, 1996). Within the brain, the patterning of the nervous system is refined by the death of nerve cells and the decay of synapses that fail to connect to an appropriate target.

[12] This point can be formalized in the language of theoretical computer science. See J. Hartmanis, "Computational Complexity and Mathematical Proofs," pp. 251-256 in *Informatics: 10 Years Back, 10 Years Ahead, 2000,* Lecture Notes in Computer Science, Springer-Verlag, Berlin, Heidelberg, 2001.

[13] For example, random processes are at the heart of stochastic optimization methods that can be used for protein structure prediction and receptor ligand docking, including simulated annealing, basin hopping, and parallel tempering. (An interesting introduction to stochastic optimization methods can be found at W. Wenzel, "Stochastic Optimization Methods," available at http://iwrwww1.fzk.de/biostruct/Opti/opti.htm.) Also, the systematic exploration of ecological models discussed in Section 5.4.8 is also based on the use of random processes.

[14] The quote is taken from L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl.):C47-C52, 1999. Hartwell et al. credit Sejnowski and Rosenberg with the neural network example (T.J. Sejnowski and C.R. Rosenberg, "Parallel Networks That Learn to Pronounce English Text," *Complex Systems* 1:145-168, 1987).

[15] A classic paper on this subject is F. Jacob, "Evolution and Tinkering," *Science* 196(4295):1161-1166, 1977.

behavior, and manifest an evolutionary history.[16] Engineered artifacts serve the purposes of their human designers, and biological organisms serve the purposes of nature—that is, to survive and reproduce.[17] Thus, the concepts needed to understand biological function may have some resemblance to some of the concepts already developed for "synthetic" disciplines, of which engineering and computer science are prime examples.

A third rationale is that the engineering disciplines have already had a long history of systems-level thinking and, indeed, have produced artifacts that are arguably approaching biological levels of complexity. For example, a Boeing 777 jetliner contains about 150,000 subsystem modules, including 1,000 computers, a number of the same order of magnitude as the estimated 300,000 different proteins in a typical human cell. Just as in the cell, moreover, these aeronautical subsystems are linked into an immensely complex "network of networks"—a control system that just happens to fly.[18]

A related point, and a key lesson from engineering, is that large systems are built out of smaller systems that are stable. Decomposition of a complex structure into an assembly of simpler structures whose operation is coordinated tends to be a much more successful strategy that building the complex structure from scratch, and this approach can be seen in the structure of the cell. Consider that a human cell has many physical structures within it—nucleus, mitochondria, and so on; each of these can be regarded as a device, many of which compose the cell. Further, many and perhaps even most cellular functions (e.g., genetic regulatory networks, metabolic pathways, signaling cascades) are implemented in a manner that is highly robust against single-point failure (i.e., the function will continue to operate properly even when one element is missing). Section 6.2.3 addresses this point in more detail.

A second view of biological organisms as engineered entities—as novel entities to be constructed by human beings rather than as existing organisms to be understood by human beings—is discussed in Section 8.4.2 on synthetic biology.

6.2.2 Biology as Reverse Engineering

Biological organisms are generally presented to scientists as completed entities, so the challenge of achieving an engineering understanding of them is in fact a challenge of *reverse engineering*. One definition of reverse engineering is "the process of analyzing a subject system with two goals in mind: (1) to

[16] While it is generally recognized that biology and evolution are intimately linked, the analogous connection between engineering and evolution is less well understood. Nevertheless, most human-engineered objects have a lot of historicity in them as well. Most human objects are designs based as improvements on previous designs, not de novo, and this can complicate the understanding of the relationship between functionality and design of a human artifact. One reason is a desire for backward compatibility—consider the fact that two-prong electric plugs and sockets are much more hazardous than some alternative designs and yet they are ubiquitous in appliances today. The same is true for operating systems—later versions of an operating system often incorporate large amounts of code from previous versions to facilitate backward compatibility. A second reason is that previous designs may have solved a design problem in a particularly effective way, and these solutions from the past are ignored today at the designer's peril. For example, consider the evolution of the rotary phone into today's push-button phones. Donald Norman observes that the cradle of the phone handset and the button-switch in it had two distinct functions: the cradle provided a place for the user to put the phone and the button-switch turned the phone on and off. Norman notes that whether deliberately or by accident, the particular design of the rotary phone that placed the on-off switch in a protected spot in the cradle also protected the on-off switch from the user accidentally hanging up the phone. However, the designers of newer push-button phones did not pick up on that feature; many push-button phones are designed so that the on-off switch and the hang-up cradle are separate—thus making the on-off switch much easier to bump and thereby to accidentally disconnect a phone call. See D. Norman, *The Design of Everyday Things*, Basic Books, New York, 1998.

[17] See for example L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Muray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl):C47-52, 1999, available at http://cgr.harvard.edu/publications/modular.pdf. Hartwell et al. further argue that it is notions of function and purpose that differentiate biology from other natural sciences such as chemistry or physics, and hence that reductionist biology—inquiry that seeks to explain biological phenomena only in chemical or physical terms—is inherently incomplete.

[18] M.E. Csete and J.C. Doyle, "Reverse Engineering of Biological Complexity," *Science* 295(5560):1664-1669, 2002, available at http://www.sciencemag.org/cgi/content/abstract/295/5560/1664.

identify the system's components and their interrelationships and (2) to create representations of the system in another form or at a higher level of abstraction."[19]

A better description could not be developed for the goal of systems biology, even without having to change any words in this definition. And yet reverse engineering, despite being a fairly standard engineering topic, is not taught to biologists.[20] One drawback is that the metaphor itself is foreign to biologists; if they wanted to do engineering of any kind, they would have been engineers. Second, reverse engineering is generally a more difficult task than forward engineering (i.e., the fabrication of a device to implement some specific functionality), and reverse engineering of a biological organism is a particularly difficult endeavor.

One important reason is that reverse engineering is often underdetermined, in the sense that multiple solutions can be developed to account for a given behavior. In such cases, choosing among them thus requires either more data or a priori assumptions about the true nature of the system being reverse-engineered. For example, in dealing with the reverse-engineering task of building detailed kinetic models of intracellular processes from time-series data, Rice and Stolovitzky note that assumptions such as linearity or sparseness or the use of predetermined model structures (e.g., reactions limited in the number of possible reactants and substrates) can help to reduce the non-uniqueness.[21]

A second and even more important reason for the difficulty of reverse engineering is that because of their evolutionary history, the organisms of interest are constructed in a highly nonoptimal manner. When engineers seek to understand how an artifact has been constructed, the basic question they ask is, Why? Why is this structure here? Why was that material used? By asking such questions of a human-engineered artifact, the engineer can often divine a reason that answers them. The reason is that engineers can be expected to design artifacts using principles such as modularity and separation of function (i.e., to minimize unnecessary links between subsystems with different purposes). These principles guard human designs against unforeseen side effects that would arise if components were not deliberately assembled in such a way as to minimize undesired or unanticipated interactions.

However, the same is not true of biological organisms. In many cases, the only answer for biological systems is, "That's the way it was built." Nature builds from accidents that happen to work and creates new mechanisms on top of old ones. While some evolved systems are quite elegant (e.g., the sensory and the motor components of the *Escherichia coli* chemotaxis mechanism), many if not most such systems at least appear to a human as inelegant, redundant, "kludgy," and inefficient—some of them extremely so. Systems engineered by humans, even very poorly engineered ones and even though they too often show their historical origins, are seldom if ever as arcane and kludgy as evolved biological organisms.

Finally, it is helpful to distinguish between two different approaches to reverse engineering. One approach to reverse engineering of biological systems—a "top-down" approach—begins with its observable behavior and characteristics, and seeks to decompose the system into components or subsystems that collectively exhibit the macroscopic behavior in question. That is, the top-down approach is based on a successive decomposition down to the system's most elemental components.

A second approach is based on a "bottom-up" approach, which begins with an understanding of the constituent parts at the lowest level, e.g., the macromolecules and the genetic regulatory networks of the

[19] E.J. Chikofsky and J.H. Cross, "Reverse Engineering and Design Recovery: A Taxonomy," *IEEE Software* 13-17, 1990.

[20] Indeed, the BIO2010 report on undergraduate education in biology (National Research Council, *Bio 2010: Undergraduate Education to Prepare Biomedical Research Scientists*, National Academies Press, Washington, DC, 2003) noted that "one approach to the study of biology is as a problem in reverse engineering. Manufactured systems are easier to understand than biological systems, because they have no unknown components, and their design principles can be explicitly stated. It is easiest to learn how to analyze systems through investigating how manufactured systems achieve their designed purpose, how their function depends on properties of their components, and how function can be reliable even with imperfect components." Also, underscoring the point that engineering is not a part of biology education today, the report emphasized the importance of exposing biology students to engineering principles and analysis in the course of their undergraduate educations. Chapter 10 has more discussion of this point.

[21] J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2(2):70-77, 2004.

Box 6.3
Functional Modules in Biology

An important theme in systems biology has been to look for functional modules that have been conserved and reused. The idea of breaking biological systems into small functional blocks has obvious appeal; the parts can be divided and conquered so that the most complex of machines become readily understood in terms of block diagrams or sets of subroutines. Clearly, some conserved modules exist such as the ribosome and the tricarboxylic acid cycle. One method to search for modules involves looking for higher-order structures or recurring sub-networks (often termed "motifs") in metabolic or gene regulatory networks. Another approach mentioned earlier is clustering expression profiles to produce groups of genes that appear to be co-regulated that should ideally reveal the functional modules. However, this assumption does not appear to generalize to all functional groups under all conditions, as some functional groups show well-correlated expression profiles whereas others do not. The low correlation of genes observed within some functional groups has been attributed to the fact that some of these genes belong to multiple functional classes. In another analysis in *E. coli*, 99 cases were found where one reaction existed in multiple pathways in EcoCyc. These observations suggest potential pitfalls with anticipating too much functional modularity in terms of biology being neatly partitioned into non-overlapping modules. Moreover, the tissue- or species-specific differences mentioned earlier may prevent simplistic transfer of modules from one biological system to another. It remains to be seen if biology is as modular as the system biologist might like it to be.

Biological modules may turn out be more interconnected and overlapping than independent in many systems. In addition, the experiences with pathway reconstruction suggest that the combinations of data source produce a more accurate if not more complete characterization of the system under study. These observations point to an eventual need to develop large-scale, predictive models based on a multitude of data sources. For example, metabolic models may combine data from many sources into a quantitative set of equations that can make predictions amenable to experimental verification. In another system, cardiac models can bridge data at multiple levels (i.e. molecular, cellular, organ, etc.) and their corresponding characteristic timescales. In this system, modeling efforts at the single-cell level in the heart suggested a mechanism of increased contraction force that was later confirmed in experimental studies of whole heart.

SOURCE: Reprinted by permission from J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2(2):70-77. Copyright 2004 Elsevier.

cells that make up the system. The philosophical notion embedded in the bottom-up approach is that a component is likely to be easier to understand than the system in which it is embedded. By successive assembly of component parts, one is able to create ever-larger assemblies whose operation is understood.

Both approaches seek as their underlying ultimate goal an understanding of how a biological system works in all of its complexity. But they require different strategies for acquiring data at different levels of scale (top-down entails data acquisition at ever-smaller scales, while bottom-up entails data acquisition at ever-larger scales). And also, it should be expected that they will generate different intermediate outputs and products along the way to this ultimate goal.

6.2.3 Modularity in Biological Entities[22]

A functional perspective on biology is centrally based on the notion that biological function is separable, into what might be called modules. The essence of a module—well known in engineering disciplines as well as computer science—is that of an entity whose function is separable from other modules. In the computer science context, a module might be a subroutine upon which various programs can build. These various programs would interact with the subroutine only through the programming interface—the set of arguments to the subroutine that parameterize its behavior. Box 6.3 describes how the search for functional modules plays into systems biology.

[22] Section 6.2.3 is based largely on L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl.):C47-C52, 1999.

Important insights into biological organisms can be gained by seeking to identify general principles that govern the structure and function of modules (Box 6.4). In a biological context, a module might be an entity that performs some biochemical function apart from other modules, isolated from those other modules by spatial localization (i.e., it is physically separated from those other modules) or by chemical specificity (i.e., its biochemical processes are sensitive only to the specific chemical signals of that module and not to others that may be present). Furthermore, modules must be able to interact with each other selectively. Specific connectivity enables module A to influence the functional behavior of module B, but not to affect the operation of modules C through Z. Also, the particular pattern of connectivity can account for some emergent properties of these modules, such as an ability to integrate information from multiple sources.

As noted by Hartwell et al., "Higher-level functions can be built by connecting modules together. For example, the super-module whose function is the accurate distribution of chromosomes to daughter cells at mitosis contains modules that assemble the mitotic spindle, a module that monitors chromosome alignment on the spindle, and a cell-cycle oscillator that regulates transitions between interphase and mitosis." When a function of a protein is restricted to one module, and the connections of that module to other modules are through such proteins, it becomes much easier to alter connections to other modules without global consequences for the entire organism.

Modular structures have many advantages. For example, the imposition of modular design on an entity allows a module to be used repeatedly by different parts of the entity. Furthermore, changes internal to the module do not have global impact if those changes do not affect its functional behavior. Modules can be combined and recombined in ways that alter the functionality of the complete system—

Box 6.4
Some Mechanisms Underlying the Structure and Function of Modules

1. Positive feedback loops can drive rapid transitions between two different stable states of a system. For example, positive feedback drives cells rapidly into mitosis, and another makes the exit from mitosis a rapid and irreversible event.[1]

2. Negative feedback loops can maintain an output parameter within a narrow range, despite widely fluctuating input. For example, negative feedback in bacterial chemotaxis[2] allows the sensory system to detect subtle variations in an input signal whose absolute size can vary by several orders of magnitude.[3] (This topic—robustness against noise—is described in more detail in Section 6.2.5.)

3. Coincidence detection systems require two or more events to occur simultaneously in order to activate an output. For example, coincidence detection is central in eukaryotic gene transcription, in which several different transcription factors must be present simultaneously at a promoter site before transcription can occur. (Note the similarity to a multi-input AND gate.)

4. Parallel circuits allow devices to survive failures in one of the circuits. For example, DNA replication involves proofreading by the DNA polymerase backed up by a mismatch repair process that removes incorrect bases after the polymerase has moved on. Both of these must fail before a cell cannot produce viable progeny, and these two mechanisms, combined with a system for killing potentially cancerous cells, reduce the frequency at which individual cells give rise to cancer to about 1 in 10^{15}.

5. Quality control systems monitor the output of many biological processes to ensure that the processes have executed correctly. Such systems can be seen in cell-cycle checkpoints, DNA replication and repair, choices between cell survival and death after insults to cells, or quality control in protein folding and/or sorting events.

[1] D.O. Morgan, "Cyclin-dependent Kinases: Engines, Clocks, and Microprocessors," *Annual Review of Cell and Developmental Biology* 13:261-291, 1997.

[2] Chemotaxis is the propensity of certain bacteria, such as *E. coli*, to swim toward higher concentrations of nutrients.

[3] H.C. Berg, "A Physicist Looks at Bacterial Chemotaxis," *Cold Spring Harbor Symposium on Quantitative Biology* 53(1):1-9, 1988.

SOURCE: Items 1-4 adapted from L. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(Suppl.):C47-C52, 1999.

the building blocks remain more or less stable, while the connectivity among them determines the character of the system.

If biological modules really do exist, one might expect to find them reused in different cellular contexts, performing the same function but to different ends. Understanding the function and behavior of a cellular pathway would entail the discovery and characterization of such modular building blocks, tasks that should be simpler than trying to understand biological networks of different organisms as an irreducible whole.

Several independent pieces of evidence have emerged supporting the modularity hypothesis. For example, evidence is accruing that certain regions of DNA are "conserved" from one species to another. These regions may be associated with genes coding for proteins or with regulatory and structural functionality. Caenepeel et al. found that the human and mouse kinomes (i.e., the collection of protein kinases in an organism) are 99 percent identical, although the percentage of identity between orthologues (i.e., genes or proteins from different organisms that have the same function) ranges from 70 percent to 99 percent (with single nucleotide insertions or deletions in many cases).[23] Dermitzakis et al. found that perhaps a third of the highly conserved DNA regions between mouse and human code for proteins, while much of the rest probably codes for regulatory and structural functionality.[24]

Genetic expression networks may also display regular patterns of interconnections (motifs) recurring in many different parts of a network at frequencies much higher than those found in randomized networks.[25] Such motifs might be regarded as building blocks that can be used to assemble entities of more complex functionality.[26] For example, Shen-Orr et al. discovered a series of simple, recurring network motifs in the gene interaction map of the bacterium *E. coli*.[27] Shortly afterwards, Richard Young and colleagues found the same motifs to recur at statistically surprising frequencies in yeast.[28] Milo et al. found that these motifs were also overrepresented in a neuronal connectivity network of *Caenorhabditis elegans* as well as the connectivity networks in the ISCAS89 benchmark set of sequential logic electronic circuits, but not in ecosystem food webs.[29] Milo et al. speculate that these motifs reflect the underlying processes that generated each type of network, in this case one set of motifs for those that process information (the genetic regulation, neuronal connectivity, and electronic logic networks) and another set of motifs for those that process and carry energy.

Finally, a collaborative project led by Eric Davidson and his group at the California Institute of Technology, and involving Bolouri and Hood at the Institute for Systems Biology, also suggests simple design principles and building blocks in genetic networks. Figure 6.1 is a map of the interactions among

[23] S. Caenepeel, G. Charydezak, S. Sudarsanam, T. Hunter, and G. Manning, "The Mouse Kinome: Discovery and Comparative Genomics of All Mouse Protein Kinases," *Proceedings of the National Academy of Sciences* 101(32):11707-11712, 2004.

[24] E.T. Dermitzakis, A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, S. Deutsch, B.J. Stevenson, et al., "Numerous Potentially Functional But Non-genic Conserved Sequences on Human Chromosome 21," *Nature* 420(6915):578-582, 2002.

[25] R. Milo, S. Shen-Or, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," *Science* 298(5594):824-827, 2002.

[26] Alon refines the notion of module as building block to suggest that modules and motifs are related but separate concepts. In Alon's view, a module in a network is a set of nodes that have strong interactions and a common function. Some nodes are internal and do not interact significantly with nodes outside the module. Other nodes accept inputs and produce outputs that control the module's interactions with the rest of the network. Alon argues that one reason modules evolve in biology is that new devices or entities can be constructed out of existing, well-tested modules; thus, adaptation to new conditions (and new forces of natural selection) is more easily accomplished. If modules are to be swapped in and out, they must possess the property that their input-output response is approximately independent of what is connected to them—that is, that the module is functionally encapsulated. By contrast, a motif is an overrepresented patterns of interconnections in a network that is likely to perform some useful behavior. However, it may not be functionally encapsulated, in which case it is not a module. For more discussion, see U. Alon, "Biological Networks: The Tinkerer as an Engineer," *Science* 301(5641):1866-1867, 2003.

[27] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network Motifs in the Transcriptional Regulation Network of *Escherichia coli*," *Nature Genetics* 31(1):64-68, 2002.

[28] T.I. Lee, H.J. Yang, S.Y. Lin, M.T. Lee, H.D. Lin, L.E. Braverman, and K.T. Tang, "Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*," *Science* 298(5594):799-804, 2002.

[29] R. Milo et al., "Network Motifs," 2002.

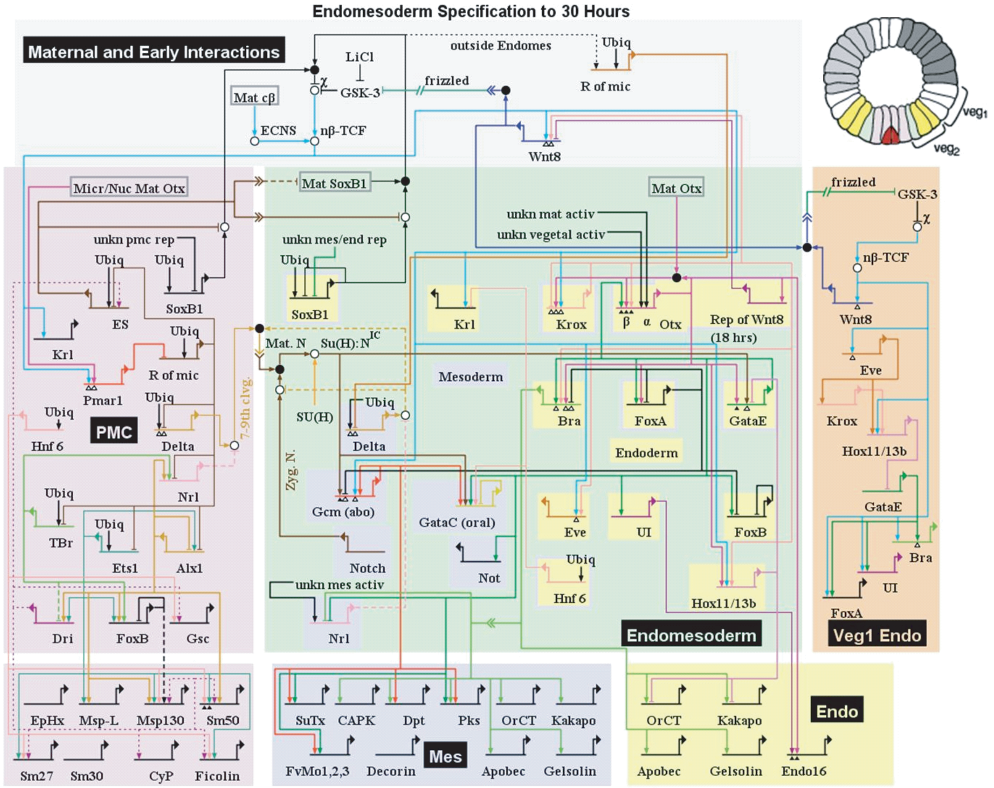


FIGURE 6.1 The endomesoderm specification network in the sea urchin species *Strongylocentrotus purpuratus*.

The period of activity represented spans embryonic growth from single cell to gastrulation (approximately 600 cells). The different background colors denote different cell types, as indicated on the cartoon of an early blastula-stage embryo on the top right. The short, thick horizontal lines represent regulatory DNA of a particular gene in the network, to which transcription factors bind to activate or repress transcription. The bent arrow emanating from each regulatory domain represents the basal transcription apparatus of the gene, and the line(s) emerging from it represent the interactions of the product of the gene with other proteins (via the white and black interaction boxes) or *cis*-regulatory DNA.

The architecture of the network is based on perturbation and expression data, on data from *cis*-regulatory analyses for several genes, and on other experiments discussed in the references below. For quantitative results of perturbation experiments and temporal details and the latest view of the network, see http://sugp.caltech.edu/endomes/.

The repression cascade motif referred to in the text is indicated by the thick black (upstream gene) and gray (downstream genes) arrows. This work is described in the following:

1. E.H. Davidson, J.P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.H. Yuh, T. Minokawa, et al., "A Genomic Regulatory Network for Development," *Science* 295(5560):1669-1678, 2002.
2. H. Bolouri and E.H. Davidson, "Modeling DNA Sequence-based *cis*-Regulatory Gene Networks," *Developmental Biology* 246(1):2-13, 2002.
3. C.T. Brown, A.G. Rust, P.J.C. Clarke, Z. Pan, M.J. Schilstra, T. De Buysscher, G. Griffin, et al., "New Computational Approaches for Analysis of *cis*-Regulatory Networks," *Developmental Biology* 246(1):86-102, 2002.
4. A. Ransick, J.P. Rast, T. Minokawa, C. Calestani, and E.H. Davidson, "New Early Zygotic Regulators of Endomesoderm Specification in Sea Urchin Embryos Discovered by Differential Array Hybridization," *Developmental Biology* 246(1):132-147, 2002.
5. C.H. Yuh, C.T. Brown, C.B. Livi, L. Rowen, P.J.C. Clarke, and E.H. Davidson, "Patchy Interspecific Sequence Similarities Efficiently Identify Positive *cis*-Regulatory Elements in the Sea Urchin," *Developmental Biology* 246(1):148-161, 2002.

6. E.H. Davidson, J.P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.H. Yuh, T. Minokawa, et al., "A Provisional Regulatory Gene Network for Specification of Endomesoderm in the Sea Urchin Embryo," *Developmental Biology* 246(1):162-190, 2002.

7. J.P. Rast, R.A. Cameron, A.J. Poustka, and E.H. Davidson, "Brachyury Target Genes in the Early Sea Urchin Embryo Isolated by Differential Macroarray Screening," *Developmental Biology* 246(1):191-208, 2002.

8. P. Oliveri, D.M. Carrick, and E.H. Davidson, "A Regulatory Gene Network That Directs Micromere Specification in the Sea Urchin Embryo," *Developmental Biology* 246(1):209-228, 2002.

SOURCE: Figure from M. Levine and E.H. Davidson, "Gene Regulatory Networks for Development," *Proceedings of the National Academy of Sciences* 102(14):4936-4942, 2005, available at http://www.pnas.org/cgi/content/full/102/14/4936. Copyright 2005 National Academy of Sciences.

approximately 50 genes underlying an early cell-type specification event in sea urchin embryos that includes several recurring interaction motifs. For example, there are several cases in which a gene (thick black arrow), instead of activating another gene directly, represses a repressor of the target gene (thick gray arrows). Such an arrangement can provide a number of possible advantages, including a sharper activation profile for the target gene, important in defining spatial boundaries between cell types.

Modularity and conservation suggest a potential for comparative studies across species (e.g., pufferfish, mice, humans) to contribute to an understanding of biological function. That is, understanding the role of a certain protein in mice, for example, may suggest a similar role for that same protein if it is found in humans.

These comments should not be taken to mean that functional modules in biological entities are necessarily simple or static. Biological systems are often made up of elements with multiple functions interacting in ways that are complex and difficult to separate, and nature exploits multiple linkages that a human engineer would not tolerate in the design of an artifact.[30] For example, a component of one module may (or may not) play a role in a different module at a different time. A module's functional behavior may be quantitatively regulated or switched between qualitatively different functions by chemical signals from other modules. Despite these important differences between biological modules and the modules that constitute humanly engineered artifacts, the notion of a collection of parts that can be counted on to perform a given function—that is, a module—is meaningful from an analytical perspective and our understanding of that function.

6.2.4 Robustness in Biological Entities

Robustness is one of the characteristics of biological systems that is most admired and most desired for engineered systems. Especially as compared to software and information systems, which are notoriously brittle, biological systems maintain functionality in the face of a range of perturbations. More traditional hardware engineering, however, has studied the questions of robustness (under various names including fault-tolerance and control systems). Applying the analytical techniques developed in engineering to studying the mechanics of robustness in biology, the logic goes, might reveal new insights not only about biology, but about robust system design.

In biology, the term robustness is used in many different ways in different subfields, including the preservation of species diversity, a measure of healing, comprehensibility in the face of incomplete information, continuity of evolutionary lineages, phenotypic stability in development, cell metabolic stability in the face of stochastic events, or resistance to point mutations.[31] Its most general usage,

[30] This is not to say that human-engineered artifacts are not affected by their origins. "Capture by history" characterizes many human artifacts as well, but likely not as strongly. For more discussion of these points, see D. Norman, 1998, cited in Footnote 16.

[31] D.C. Krakauer, "Robustness in Biological Systems—A Provisional Taxonomy," *Complex Systems Science in Biomedicine*, T.S. Deisboeck, J.Y. Kresh, and T.B. Kepler, eds., Kluwer, New York, 2003.

however, refers to the ability of a structure or process to persist in the face of perturbations of internal components or the environment. Those perturbations might include outright component failure, unexpected behavior from components or other cooperating systems, stochastic changes in chemical concentrations or reaction rates, mutations, or the motion of external biochemical parameters. These sorts of perturbations, such as stochastic changes of molecular concentrations, are intrinsic to the nature of biology, from the molecular scale to the ecological.

A robust response to these perturbations generally consists of one of three types: (1) parameter insensitivity, meaning that a robust process does not depend on a single ideal value of an input; (2) graceful degradation, in which the level of functionality of the system is indeed lessened by component failures, but it continues to function; and (3) adaptation, in which internal components reconfigure to react to a change to maintain the same level of functionality.[32]

Kitano notes that robustness is attained in biological systems by using mechanisms well known to human engineers. He describes four mechanisms or approaches to biological robustness:[33]

1. System control mechanisms such as negative-feedback and feed-forward control;
2. Redundancy, whereby multiple components with equivalent functions are introduced for backup;
3. Structural stability, where intrinsic mechanisms are built to promote stability; and
4. Modularity, where subsystems are physically or functionally insulated so that failure in one module does not spread to other parts and lead to system-wide catastrophe.

Kitano then notes that these approaches used in engineering systems are also found in biological systems, pointing out that "redundancy is seen at the gene level, where it functions in control of the cell cycle and circadian rhythms, and at the circuit level, where it operates in alternative metabolic pathways in *E. coli*." Furthermore, engineering approaches have proven to be a useful lens when investigating biological robustness.

For example, Barkai and Leibler[34] established a model (later confirmed experimentally) to explain perfect robust adaptation in bacterial chemotaxis, or the ability of bacteria to move toward increased concentrations of certain ligands. It had long been known that the mechanism responsible for this ability had several key attributes, among them a high sensitivity to changes in chemical concentration, together with an ability to adapt to the absolute level of that concentration. Working with the known molecular makeup of these cells (e.g., the receptors, kinases, and diffusible messenger proteins), Barkai and Leibler showed that when varied separately, many of the rate constants (such as molecular concentrations of elements of the signaling network or reaction rates) could be varied by orders of magnitude without affecting the magnitude of the response.[35]

Later work by Yi et al. used the mathematics of control systems to show how the Barkai-Leibler model was a special case of integral feedback control, a well-studied approach of control theory.[36] In addition to control theory (including feedback and feed-forward control), many other engineering approaches are found in biological systems, including redundancy, modularity, purging (quickly eliminating failing components), and spatial compartmentalization.[37]

[32] H. Kitano, "Systems Biology: A Brief Overview," *Science* 295(5560):1662-1664, 2002. Available at http://www.sciencemag.org/cgi/content/abstract/295/5560/1662.

[33] H. Kitano, "Systems Biology," 2002.

[34] N. Barkai and S. Leibler, "Robustness in Simple Biochemical Networks," *Nature* 387(6636):913-917, 1997.

[35] However, the mechanism does not account for the full dynamic range of the sensor patches at a molecular level. (It may be that some sort of emergent property of the sensor patch as a whole, as opposed to some property of the individual sensor complexes, is necessary to obtain the full dynamic range. See, for example, T.S. Shimizu, S.V. Aksenov, and D. Bray, "A Spatially Extended Stochastic Model of the Bacterial Chemotaxis Signaling Pathway," *Journal of Molecular Biology* 329(2):291-309, 2003.)

[36] T.M. Yi, Y. Huang, M.I. Simon, and J. Doyle, "Robust Perfect Adaptation in Bacterial Chemotaxis Through Integral Feedback Control," *Proceedings of the National Academy of Sciences* 97(9):4649-4653, 2000.

[37] D.C. Krakauer, "Robustness in Biological Systems," 2003.

Kitano makes the point that robustness is a property of an entire system;[38] it may be that no individual component or process within a system would be robust, but the system-wide architecture still provides robust behavior. This presents a challenge for analysis, since elucidating such behaviors can be counterintuitive and computationally demanding.[39] In one such example, von Dassow and colleagues investigated the development of striped patterns in *Drosophila*.[40] They computationally modeled a network of interactions between genes and regulatory proteins active during embryogenesis and explored the parameter space to see which sets of parameters produced stable striping. In their first attempt, they were unable to reproduce such behavior computationally. However, once they added two more molecular events and their interactions to the network, a surprisingly high proportion of the randomly chosen parameters produced the desired results. This strongly implies that such a network, taken as a whole, is a robust developmental module, able to produce a particular effect despite wide variation in reaction parameters.

In a refinement to that work, Ingolia investigated the architecture of that network to attempt to determine the structural sources of such robust behavior.[41] He determined that the source of the robustness at the network level was a pair of positive feedback loops of gene expression, which led to cells being forced to one of two stable states (bistability). That is, small perturbations or changes in certain parameters would necessarily result in individual cells reaching one of two states. Ingolia showed that such bistability, at both an individual cell level and a network level, is an important architectural property leading to robust behavior and that the latter is in fact a consequence of the former. Moreover, it is this bistability that is responsible for the ability of the network to maintain a fixed pattern of gene expression even in the face of cell division and growth.[42]

Robustness comes at a cost of increased complexity. The simplest bacteria can survive only within narrow ranges of environmental parameters, while more complex bacteria, such as *E. coli* (with a genome an order of magnitude larger than mycoplasma), can withstand more severe environmental fluctuations.[43] This increased complexity can in turn be the root of cascading failures, if the elements of the network responsible for the adaptive response fail. This implies that increased robustness of a certain aspect or element of a system with respect to a certain perturbation may come at the cost of increased vulnerability in a different aspect or element or to a different attack.

Robustness can also serve as a signpost for discovering the details of biological function. Although there may be a prohibitively large number of ways that a genetic network could produce a given result, for example, only a few of those ways are likely to do so robustly. Knowledge of the robust qualities of a biological system, coupled with theoretical or simulated analysis of networks, could aid in reverse engineering the system to determine its actual configuration.[44]

An open and intriguing question is the relationship between robustness and evolution. Because robustness is the quality of maintaining stability, in some sense it stands as a potential inhibitor to evolution, for example, by masking the effects of point mutations. And yet robust modules or organisms are more likely to survive, and thus pass on into succeeding generations. How does robustness evolve? How do robust systems evolve? One engineering approach to this problem is to consider biological systems as sets of components interacting through protocols,[45] with one critical measure of a

[38] H. Kitano, "Systems Biology," 2002. Available at http://www.sciencemag.org/cgi/content/abstract/295/5560/1662.

[39] A.D. Lander, "A Calculus of Purpose," *PLoS Biology* 2(6):e164, 2004.

[40] G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, "The Segment Polarity Network Is a Robust Developmental Module," *Nature* 406(6792):188-192, 2000.

[41] N.T. Ingolia, "Topology and Robustness in the *Drosophila* Segment Polarity Network," PLoS Biology 2(6):e123, 2004.

[42] A.D. Lander, "A Calculus of Purpose," 2004.

[43] J.M. Carlson and J. Doyle, "Complexity and Robustness," *Proceedings of the National Academy of Sciences* 99(Suppl. 1):2538-2545, 2002.

[44] U. Alon, "Biological Networks: The Tinkerer as an Engineer," *Science* 301:1866-1867, 2003.

[45] M.E. Csete and J.C. Doyle, "Reverse Engineering of Biological Complexity," *Science* 295:1664-1669, 2002.

good protocol being its ability to support both robustness and evolvability, a key consideration in technical protocols of human engineering such as TCP/IP.

6.2.5 Noise in Biological Phenomena[46]

As one illustration of how engineering disciplines might shed light on biological mechanism, consider the opposition of robustness and noise in biological phenomena. Biological organisms exhibit high degrees of robustness in the face of changing environments. Engineered artifacts designed by human beings have used mechanisms such as negative feedback to provide stability, redundancy to provide backup, and modularity for the isolation of failures to enhance robustness. As the discussion below indicates, these mechanisms are used for these purposes in biological organisms, as well.[47]

In a biological context, noise can take the form of fluctuations in quantities such as reaction rates, concentrations, spatial distributions, and fluxes. In addition, fluctuations may also occur at the molecular level. However, despite the noise inherent in the internal environment of a cell, cells operate—often robustly and quite stably—within strict parameters, and robustness has been hypothesized as an intrinsic property of intracellular networks. (For instance, the chemotaxis pathway in *E. coli* functions over a wide range of enzymatic activities and protein concentrations.[48] Robustness is also illustrated in some developmental processes[49] and phage lambda regulation.[50]) This robustness suggests that cells use and reject noise in a systematic manner.

For the analysis of biological noise, much of the analysis originally derived from signal processing and control theory is applicable.[51] Indeed, pathways can be regarded as analog filters and classified in terms of frequency response, where the differences between filtering electronic noise and filtering biological noise are reflected only in the details of the underlying mechanisms rather than in high-level abstractions of filtering theory.

Cascades and relays such as two-component systems and the mitogen-activated protein kinase pathway function as low-pass filters (i.e., filters that attenuate high-frequency noise).[52] As a general rule, longer cascades are more effective at reducing noise. However, because noise arises in the pathway itself, the amount of internally generated noise increases with cascade length—suggesting that there is an optimal cascade length for attenuating noise.[53]

It is not surprising that low-pass filters are components of biological systems. As noted above, biological systems operate homeostatically,[54] and the essential principle underlying homeostasis is that of negative feedback. From the standpoint of signal processing, a negative feedback loop functions as a low-pass filter.

[46] Section 6.2.5 is based on and incorporates several excerpts from C.V. Rao, D.M. Wolf, and A.P. Arkin, “Control, Exploitation and Tolerance of Intracellular Noise,” *Nature* 420(6912):231-237, 2002.

[47] H. Kitano, “Systems Biology: A Brief Overview,” Science 295(5560):1662-1664, 2002. Available at http://www.sciencemag.org/cgi/content/abstract/295/5560/1662.

[48] N. Barkai and S. Leibler, “Robustness in Simple Biochemical Networks,” *Nature* 387:913-917, 1997; U. Alon, M.G. Surette, N. Barkai and S. Leibler, “Robustness in Bacterial Chemotaxis,” *Nature* 397:168-171, 1999. (Cited in Rao et al., 2002.)

[49] G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, “The Segment Polarity Network Is a Robust Developmental Module,” *Nature* 406:188-192, 2000; E. Meir, G. von Dassow, E. Munro, and G.M. Odell, “Robustness, Flexibility, and the Role of Lateral Inhibition in the Neurogenic Network,” *Current Biology* 12:778-786, 2002. (Cited in Rao et al., 2002.)

[50] J.W. Little, D.P. Shepley, and D.W. Wert, “Robustness of a Gene Regulatory Circuit,” *EMBO Journal* 18:4299-4307, 1999.

[51] A.P. Arkin, “Signal Processing by Biochemical Reaction Networks,” pp. 112-144, *Self-organized Biological Dynamics and Nonlinear Control*, J. Walleczek, ed., Cambridge University Press, London, 2000; M. Samoilov, A. Arkin, and J. Ross, “Signal Processing by Simple Chemical Systems,” *Journal of Physical Chemistry* 106:10205-10221, 2002. (Cited in Rao et al., 2002.)

[52] P.B. Detwiler, S.A. Ramanathan, A. Sengupta, and B.I. Shraiman, “Engineering Aspects of Enzymatic Signal Transduction: Photoreceptors in the Retina,” *Biophysical Journal* 79(6):2801-2817, 2000. (Cited in Rao et al., 2002.)

[53] M. Thattai and A.Van Oudenaarden, “Attenuation of Noise in Ultrasensitive Signaling Cascades,” *Biophysical Journal* 82(6):2943-2950, 2002. (Cited in Rao et al., 2002.)

[54] Homeostasis is the property of a system that enables it to respond to changes in its environment in such a way that it tends to maintain its original state.

A second useful construct from signal processing is the bandpass filter, which is based on the control theory notion of integral feedback. Integral feedback is a kind of negative feedback that amplifies intermediate frequencies and attenuates low and high frequencies. A biological instantiation of integral feedback is contained in bacterial chemotaxis.[55]

In addition to the filters described above, other mechanisms attenuate noise in systems. These include the following:

- *Redundancy*. Noise in a single channel might be misinterpreted as a genuine signal. However, redundancy—in the form of multiple channels serving the same function—can help to minimize the likelihood of such an occurrence. In a biological context, redundancy has been demonstrated in mechanisms such as gene dosage and parallel cascades,[56] which attenuate the effects of noise by increasing the likelihood of gene expression or establishing a consensus from multiple signals.
- *Checkpointing*. Noise can interfere with the successful completion of various biological operations that are essential in a pathway. However, a checkpoint can ensure that each step in a pathway is completed successfully before proceeding with the next step. Such checkpoints have been characterized in the cell cycle and flagellar biosynthesis.[57]
- *Proofreading*. Noise can introduce errors into a process. But error-correcting mechanisms can reduce this effect of noise, as is the case of kinetic proofreading in protein translation.[58]

A final, and surprising, mechanism is that complexity itself in some cases can be implicated in the robustness of an organism against noise. In 1942, Waddington noted the stability of phenotypes (from the same species) against a backdrop of considerable genetic variation, a phenomenon known as canalization.[59] In principle, such stability could result from explicit genetic control of phenotype features, such as the number of fingers on a hand or the placement of wings on an insect's body. However, Siegal and Bergman modeled the developmental process responsible for the emergence of such features as a network of interacting transcriptional regulators and found that the network constrains the genetic system to produce canalization.[60] Furthermore, the extent of canalization, measured as the insensitivity of a phenotype to changes in the genotype (i.e., to mutations), depends on the complexity of the network, such that more highly connected (i.e., more complex) networks evolve to be more canalized. (Box 6.5 provides more details.)

Consider that noise can also make positive contributions to biological systems. For example, it is well known from the agricultural context that monocultures are less robust than ecosystems that involve multiple species—the first can be wiped out by a disease that targets the specific crop in question, whereas the second cannot. Thus, some degree of variation in a populating species is desirable, and noise is one mechanism for introducing variation that results in population heteroge-

[55] The size of a single bacterium is so small that the bacterium is unable to sense a spatial gradient across the length of its body. Thus, to sense a spatial gradient, the bacterium moves around and senses chemical concentrations in different locations at different times; the result is a motion bias toward attractants. See T.M. Yi, Y. Huang, M.I. Simon, and J. Doyle, "Robust Perfect Adaptation in Bacterial Chemotaxis Through Integral Feedback Control," *Proceedings of the National Academy of Sciences* 97(9):4649-4653, 2000. (Cited in Rao et al., 2002.)

[56] H.H. McAdams and A. Arkin, "It's a Noisy Business! Genetic Regulation at the Nanomolar Scale," *Trends in Genetics* 15(2):65-69, 1999; D.L. Cook, A.N. Gerber, and S.J. Tapscott, "Modeling Stochastic Gene Expression: Implications for Haploinsufficiency," *Proceedings of the National Academy of Sciences* 95(26):15641-15646, 1998. (Cited in Rao et al., 2002.)

[57] L.H. Hartwell and T.A. Weinert, "Checkpoints: Controls That Ensure the Order of Cell Cycle Events," *Science* 246(4930):629-634, 1989. (Cited in Rao et al., 2002.)

[58] M.V. Rodnina and W. Wintermeyer, "Ribosome Fidelity: tRNA Discrimination, Proofreading and Enduced Fit," *Trends in Biochemical Science* 26(2):124-130, 2001. (Cited in Rao et al., 2002.)

[59] C.H. Waddington, "Canalization of Development and the Inheritance of Acquired Characters," *Nature* 150:563-565, 1942.

[60] M.L. Siegal and A. Bergman, "Waddington's Canalization Revisited: Developmental Stability and Evolution," *Proceedings of the National Academy of Sciences* 99(16):10528-10532, 2002.

Box 6.5
Canalization and the Connectivity of Transcriptional Regulatory Networks

To explore the possibility that genetic canalization may be a by-product of other selective forces, . . . [we start with] the model of A. Wagner, who treats development as the interaction of a network of transcriptional regulatory genes, phenotype as the equilibrium state of this network, and fitness as a function of the distance between an individual's equilibrium state and the optimum state. . . . Evolution in the model [a generalized version of Wagner's] consists of three phases: mating, development, and selection. Mating and selection are modeled in accord with traditional population-genetic approaches. . . . [To handle development] one can represent a network of transcriptional regulators by a state vector containing the concentration of each gene product and a matrix, the entries of which represent the effects of each gene product on the expression of each gene. Entries may be either positive (activating) or negative (repressing) and may differ in magnitude. Zero elements in the matrix represent the absence of interaction between the given gene product and gene. The developmental process is then fully described by a set of nonlinear coupled difference equations. . . . Wagner draws an analogy between the rows of the interaction matrix and the enhancer regions of the genes in the network and further justifies the biological realism of this type of model by reference to data from actual genetic networks. An important assumption in the model, also justified by A. Wagner, is that functional genetic networks will reach a stable equilibrium gene-expression state, and that unstable networks reflect, in a sense, the failure of development. Thus, in his model and ours, development itself enforces a kind of selection, because we require that the network of regulatory interactions produce a stable equilibrium gene-expression state (its "phenotype"), whose distance to an optimum state can then be measured during the selection phase.

. . . We report here the results of numerical simulations of our model of an evolving developmental-genetic system. We demonstrate an important, perhaps primary, role for the developmental process itself in creating canalization, in that insensitivity to mutation evolves even when stabilizing selection is absent. We go on to demonstrate that the complexity of the network is a key factor in this evolutionary process, in that networks with a greater proportion of connections evolve greater insensitivity to mutation.

. . . One is led to wonder whether the evolution of canalization under no stabilizing selection on the gene-expression pattern is an artifact of the modeling framework or whether it represents a finding of real biological significance. We argue that the latter is true on a number of counts. To begin, we acknowledge that it is difficult to envision a scenario in nature in which the stability of a developmental module is required, but the phenotype produced by that module is not subject to selection. One situation in which this condition may hold is when a species colonizes a new territory with virtually unlimited resources, so selection is only for those that develop to reproduce. Furthermore, even if such a scenario does not pertain, the conceptual decomposition of stabilizing selection into selection for an optimum and selection for developmental stability is important. Thus, even in scenarios in which members of a population are subject to selection for an optimum, the evolution of canalization may proceed because of the underlying selection for stability of the developmental outcome. Our results suggest that this underlying selection can occur very fast. Because others have argued that the evolution of canalization under stabilizing selection may be slow, developmental stability may therefore be the dominant force in the evolution of canalization.

SOURCE: Reprinted by permission from M.L. Siegal and A. Bergman, "Waddington's Canalization Revisited: Developmental Stability and Evolution," *Proceedings of the National Academy of Sciences* 99(16):10528-10532, 2002. Copyright 2002 National Academy of Sciences. (References and figures are omitted above and can be found in the original article.)

neity and diversity. For example, noise (in the form of molecular fluctuations) introduced into the genetic circuit governing development in phage lambda can cause an initially homogeneous population to separate into lytic and lysogenic populations.[61] (In this case, the basic mechanism involves

[61] A. Arkin, J. Ross, and H.H. McAdams, "Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage Lambda-infected *Escherichia coli* Cells," *Genetics* 149(4):1633-1648, 1998. (Cited in Rao et al., 2002.)

two antagonistic feedback loops that create a switch and molecular fluctuations that partition the initial population stochastically.)

Noise can be used to enhance a signal when certain nonlinear effects are present, as demonstrated by the phenomenon of stochastic resonance.[62] Stochastic resonance is found in many biological systems, including the electroreceptors of paddlefish,[63] mechanoreceptors in the tail fins of crayfish,[64] and hair cells in crickets.[65] A similar phenomenon can potentially increase sensitivity in certain signaling cascades.[66]

Finally, noise can be useful for introducing stability. The network that controls circadian rhythms consists of multiple, complex, interlocking feedback loops. Both deterministic and stochastic mechanisms for noise resistance in circadian rhythms have been explored,[67] and it turns out that stochastic models are able to produce regular oscillations when the deterministic models do not,[68] suggesting that the regulatory networks may utilize molecular fluctuations to their advantage.

The discussion above suggests that biological robustness is in some ways a problem of controlling the effects of noise and in other ways one of exploiting those effects. Considerations of noise and robustness thus offer insight into the design and function of intracellular networks.[69] That is, the function of an intracellular network may require specific regulatory and information structures, and certain design features are necessary for a stable network phenotype.

Finally, note that mechanisms of the sorts described above do not generally function in isolation, but rather interact in complex networks involving multiple feedback loops, and the resulting networks can produce diverse phenomena, including switches, memory, and oscillators.[70] Such coupling also has an important analytical consequence—namely, that the composite behavior of multiple coupled mechanisms is much more difficult to predict than the behavior of individual components. To analyze multiple coupled systems, computational models are highly useful.

6.3 A COMPUTATIONAL METAPHOR FOR BIOLOGY

In addition to the abstractions described above, computing and computer science can also provide life scientists with a rich source of language, metaphors, and analogies with which to describe biological phenomena and insights from a computational perspective. These linguistic and cognitive aspects may well make it easier for insights originating in computing to be made relevant to biology, and thus

[62] L. Gammaitoni, P. Hanggi, P. Jung, and F. Marchesoni, "Stochastic Resonance," *Reviews of Modern Physics* 70:223-287, 1998. (Cited in Rao et al., 2002.)

[63] D.F. Russell, L.A. Wilkens, and F. Moss, "Use of Behavioural Stochastic Resonance by Paddle Fish for Feeding," *Nature* 402(6759):291-294, 1999. (Cited in Rao et al., 2002.)

[64] J.K. Douglass, L. Wilkens, E. Pantazelou, and F. Moss, "Noise Enhancement of Information Transfer in Crayfish Mechanoreceptors by Stochastic Resonance," *Nature* 365(6444):337-340, 1993. (Cited in Rao et al., 2002.)

[65] J.E. Levin and J.P. Miller, "Broadband Neural Encoding in the Cricket Cercal Sensory System Enhanced by Stochastic Rresonance," *Nature* 380(6570):165-168, 1996. (Cited in Rao et al., 2002.)

[66] J. Paulsson, O.G. Berg, and M. Ehrenberg, "Stochastic Focusing: Fluctuation-enhanced Sensitivity of Intracellular Regulation," *Proceedings of the National Academy of Sciences* 97(13):7148-7153, 2000. (Cited in Rao et al., 2002.)

[67] N. Barkai and S. Leibler, "Circadian Clocks Limited by Noise," *Nature* 403(6767):267-268, 2000; D. Gonze, J. Halloy, and A. Goldbeter, "Robustness of Circadian Rhythms with Respect to Molecular Noise," *Proceedings of the National Academy of Sciences* 99(2):673-678, 2002; P. Smolen, D.A. Baxter, and J.H. Byrne, "Modeling Circadian Oscillations with Interlocking Positive and Negative Feedback Loops," *Journal of Neuroscience* 21(17):6644-6656, 2001. (Cited in Rao et al., 2002.)

[68] J.M. Vilar, H.Y. Kueh, N. Barkai, and S. Leibler, "Mechanisms of Noise Resistance in Genetic Oscillators," *Proceedings of the National Academy of Sciences* 99(9):5988-5992, 2002. (Cited in Rao et al., 2002.)

[69] M.E. Csete and J.C. Doyle, "Reverse Engineering of Biological Complexity," *Science* 295(5560):1664-1669, 2002; M. Morohashi, et al., "Robustness as a Measure of Plausibility in Models of Biochemical Networks," *Journal of Theoretical Biology* 216(1):19-30, 2002; L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl):C47-C52, 1999. (Cited in Rao et al., 2002.)

[70] M.B. Elowitz and S. Leibler, "A Synthetic Oscillatory Network of Transcriptional Regulators," *Nature* 403(6767):335-338, 2000; T.S. Gardner, C.R. Cantor, and J.J. Collins, "Construction of a Genetic Toggle Switch in *Escherichia coli*," *Nature* 403(6767):339-342, 2000. (Cited in Rao et al., 2002.)

information abstractions can be used to communicate about or to explain biological processes and concepts. Consider, for example, the Jacob and Monod description of the genome as a "genetic program," capable of controlling its own execution.[71] (Conversely, biological metaphors and language might offer analogous benefits to computing, which is the subject of Chapter 8.) At the same time, poorly chosen metaphors can limit understanding by carrying over misleading or irrelevant details. For example, the "genetic program" metaphor described above might lead one to think of protein synthesis as being executed one instruction at a time (as most computer programs would be), obscuring the parallel and interconnected nature of the genetic protein synthesis network.[72]

The use of a metaphor (to look at a problem in field A through the lens of field B) invites one to apply insights from field B to the problem in field A. Metaphors are often (indeed, almost always) imprecise and somewhat vague, because they are not specific about which insights from field B are relevant to field A. They can nevertheless be useful, because they constitute an additional source of insight and new ways of thinking to be brought to bear on field A that might not otherwise be available in the absence of those metaphors. Moreover, field B—as a discipline—constitutes an existence proof that the insights in question can in fact be part of an intellectually coherent whole.

Consider, for example, extending the notion of the "genetic program." In some sense, the DNA sequence can be analogized to the binary code of a program. However, in many real computer programs, a program structure or architecture or individual components may be apparent from representing the program in its source code form, where things such as variable declarations and subroutines make manifestly obvious what is obscured in the binary representation. Calling sequences between program and subprogram define program interfaces and protocols for how different components of a program may communicate—data definitions, formats, and semantics, for instance. Thus, it may be meaningful to inquire about the analogous things in biology, and indeed, a gene contained in DNA might well be one analogue of a subprogram or the action potential in neuroscience one analogue of a communications protocol.

Another analogy can be drawn between the evolution of computing and the biological transition from single-cell organisms to multicell organisms. Multicellular life exploits four broad strategies: collaboration between highly specialized cells; communication by polymorphic messages; self, defined by a stigmergic structure; and self, protected by programmed cell death. These strategies are rare in single-cell organisms but nearly universal in multicellular organisms, and evolved before or coincident with the emergence of multicellular life. As described in Table 6.1, each of these strategies may be analogous to trends seen in computing today.

To illustrate how the use of a computational metaphor can provide insight and lead to deeper exploration, note that cellular processes are concurrent (i.e., changes in the surrounding environment can trigger the execution of many parallel processes); operate at many levels including the submolecular, molecular, subcellular, and cellular; and involve relationships among many subcellular and molecular objects. Computer scientists have devised a number of formalisms that are capable of representing such processes, and Kam et al.[73] modeled aspects of T-cell activation using the formalism of Statecharts,[74] as they have been adapted to the framework of object-oriented modeling.[75] Because the object-oriented Statechart approach supports

[71] F. Jacob and J. Monod, "Genetic Regulatory Mechanisms in the Synthesis of Proteins," *Journal of Molecular Biology* 3:318-356, 1961.

[72] E.F. Keller, *Making Sense of Life—Explaining Biological Developments with Models, Metaphors, and Machines*, Harvard University Press, Cambridge, MA, 2003.

[73] N. Kam, I.R. Cohen, and D. Harel, "The Immune System as a Reactive System: Modeling T Cell Activation with Statecharts," *Proceedings of a Symposium on Visual Languages and Formal Methods* (VLFM'01), part of IEEE Symposium on Human-centric Computing (HCC'01), 2001, pp. 15-22.

[74] D. Harel, "Statecharts: A Visual Formalism for Complex Systems," *Science of Computer Programming* 8:231-274, 1987. (Cited in Kam et al., " The Immune System as a Reactive System," 2001.)

[75] G. Booch, *Object-Oriented Analysis and Design, with Applications*, Addison-Wesley, Menlo Park, CA, 1994; D. Harel and E. Gery, "Executable Object Modeling with Statecharts," *Computer*, 31-42, 1997; J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, and W. Lorensen, *Object-Oriented Modeling and Design*, Prentice Hall, Englewood Cliffs, NJ, 1991. (Cited in Kam et al., 2001.)

TABLE 6.1 Principles of Operation for Multicellular Organisms and Networked Computing

Principle	Multicellular Organisms	Networked Computing
Collaboration between highly specialized cells	Cells in biofilms specialize temporarily according to "quorum" cues from neighbors. Cells in "true" multicellular organisms permanently specialize (differentiate) during development. Loss of differentiation is an early sign of cancer.	Today most computers retain a large repertoire of unused general behavior susceptible to viral or worm attack. Biology suggests that more specialization and less monoculture would be advantageous (although market forces may oppose this).
Communication by polymorphic messages	Cells in multicelled organisms communicate with each other via messenger molecules, *never* DNA. The "meaning" of cell-to-cell messages is determined by the receiving cell, not the sender.	Executable code is the analogue of DNA. Most PCs permit easy, and hidden, download of executable code (Active-X or even exe). However, importing executable code is well known to create security risks, and secure systems minimize or eliminate this capability.
"Self" defined by a stigmergic structure	Multicelled organisms and biofilms build extracellular stigmergic structures (bone, shell, or just slime) that define the persistent self. "Selfness" resides as much in the extracellular matrix as in the cells.	Determination of self is largely ad hoc in today's systems. However, an organization's intranet is a stigmergic structure, as are its persistent databases.
"Self" protected by programmed cell death (PCD)	Every healthy cell in a multicelled organism is prepared to commit suicide. PCD evolved to deal with DNA replication errors, viral infection, and rogue undifferentiated cells. PCD reflects a multicellular perspective—sacrificing the individual cell for the good of the multicellular organism.	A familiar example in computing is the Blue Screen of Death, which is a programmed response to an unrecoverable error. An analogous computer should sense its own rogue behavior (e.g., download of uncertified code) and disconnect itself from the network or reboot itself periodically to give itself a clean initial state.

SOURCE: Steve Burbeck, IBM, personal communication, October 11, 2004.

concurrency, multilevel description, and object orientation, Kam et al. constructed a T-cell simulation that presents its results by displaying animated versions of the model's Statecharts.

A second example is provided by the work of Searls. It is a common, if not inescapable, metaphor that DNA represents the language of life. In the late 1980s and early 1990s, David B. Searls and collaborators made the metaphor much more concrete, applying formal language theory to the analysis of nucleic acid sequences.[76] Linguistics theory considers four levels of interpretation of text: lexical (the

[76] D.B. Searls, "The Linguistics of DNA," *American Scientist* 80:579-591, 1992. Formal language theory is a major subfield of computer science theory; it is based on Noam Chomsky's work on linguistics in the 1950s and 1960s, especially the Chomsky hierarchy, a categorization of languages by their inherent complexity. Formal languages are at the heart of parsers and compilers, and there exists a wide range of both theoretic analysis and practical software tools for the production, transformation, and analysis of text. The main algorithmic tool of language theory is the generative grammar, a series of rules that transforms higher-level abstract units of meaning (such as "sentence" or "noun phrase") into more concrete potential statements in a given language. Grammars can be categorized into regular, context-free, context-sensitive, and recursively enumerable, each of which requires more algorithmic complexity to recognize than the level before it.

identification of specific words), syntactic (the grouping of words into grammatically correct phrases), semantic (the assignment of meaning to words and phrases), and pragmatic (the role of a piece of text in the larger context). These match entirely well to genomic analysis: grouping bases into codons, genes, the function of the resulting protein, and the role of that protein in the larger molecular system.[77]

Linguistic analyses can reveal or explain relationships between bases that are far apart in a sequence. For example, an RNA structure called a stem-loop has a palindrome-like sequence, with Watson-Crick pairs at equal distances away from the center. Traditional probabilistic or pattern-searching approaches would have some difficulty recognizing this structure, but it is quite simple with a grammar that produces palindromes. Some sequences of nucleic acids result in ambiguous linguistic interpretations; while this is a difficulty for computer languages, it represents a strength of biological linguistic analysis, because these ambiguities correctly represent alternative secondary structures.[78]

This approach has been fruitful for analyzing genetic sequences and characterizing the complexity and structure of genes. GenLang, a software system that employs linguistic approaches, has successfully identified tRNA genes, group I introns, protein-encoding genes, and the specification of gene regulatory elements.[79] Other important findings include placing RNA in the Chomsky hierarchy as at least beyond context-free languages. Finally, the approach provides a powerful tool for understanding the evolution of nucleic acid sequences; since the first sequences were most likely random (and thus regular languages), there must be a mechanism that somehow promoted sequence language into more powerful linguistic categories. This can be seen as an algebraic problem of operational closure, and the question is, For which string operations are regular languages and context-free languages not closed?[80]

[77] D.B. Searls, "Reading the Book of Life," *Bioinformatics* 17(7):579-580, 2001.

[78] D.B. Searls, "The Language of Genes," *Nature* 420(6912):211-217, 2002.

[79] D.B. Searls, and S. Dong, "A Syntactic Pattern Recognition System for DNA Sequences" in *Proceedings of the Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*, H.A. Lim, J. Fickett, C.R. Cantor, and R.J. Robbins, eds., World Scientific Publishing Co., pp. 89-101, 1993.

[80] D.B. Searls, "Formal Language Theory and Biological Macromolecules," *Series in Discrete Mathematics and Theoretical Computer Science* 47:117-140, 1999.

7

Cyberinfrastructure and Data Acquisition

7.1 CYBERINFRASTRUCTURE FOR 21ST CENTURY BIOLOGY

Twenty-first century biology seeks to integrate scientific understanding at multiple levels of biological abstraction, and it is holistic in the sense that it seeks an integrated understanding of biological systems through studying the set of interactions between components. Because such an enormous, data-intensive effort is necessarily and inherently distributed over multiple laboratories and investigators, an infrastructure is necessary that facilitates the integration of experimental data, enables collaboration, and promotes communication among the various actors involved.

7.1.1 What Is Cyberinfrastructure?

Cyberinfrastructure for science and engineering is a term coined by the National Science Foundation (NSF) to refer to distributed computer, information, and communication technologies and the associated organizational facilities to support modern scientific and engineering research conducted on a global scale. As articulated by the Atkins panel,[1] the technology substrate of cyberinfrastructure involves the following:

- *High-end general-purpose computing centers* that provide supercomputing capabilities to the community at large. In the biological context, such capabilities might be used to undertake, for example, calculations to determine the three-dimensional structure of proteins given their genetic sequence. In some cases, these computing capabilities could be provided by local clusters of computers; in other cases, special-purpose hardware; and in still others, computing capabilities on demand from a computing grid environment.
- *Data repositories* that are well curated and that store and make available to all researchers large volumes and many types of biological data, both in raw form and as associated derived products. Such repositories must store data, of course, but they must also organize, manage, and document these

[1] "Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the NSF Blue-Ribbon Advisory Panel on Cyberinfrastructure," 2003, available at http://www.communitytechnology.org/nsf_ci_report/report.pdf.

datasets dynamically. They must provide robust search capabilities so that researchers can find the datasets they need easily. Also, they are likely to have a major role in ensuring the data interoperability necessary when data collected in one context are made available for use in another.

- *Digital libraries* that contain the intellectual legacy of biological researchers and provide mechanisms for sharing, annotating, reviewing, and disseminating knowledge in a collaborative context. Where print journals were once the standard mechanism through which scientific knowledge was validated, modern information technologies allow the circumvention of many of the weaknesses of print. Knowledge can be shared much more broadly, with much shorter lag time between publication and availability. Different forms of information can be conveyed more easily (e.g., multimedia presentations rich in biological imagery). One researcher's annotations to an article can be disseminated to a broader audience.
- *High-speed networks* that connect large-scale, geographically distributed computing resources, data repositories, and digital libraries. Because of the large volumes of data involved in biological datasets, today's commodity Internet is inadequate for high-end scientific applications, especially where there is a real-time element (e.g., remote instrumentation and collaboration). Network connections ten to a hundred times faster than those generally available today are a lower bound on what will be necessary.

In addition to these components, cyberinfrastructure must provide software and services to the biological community. For example, cyberinfrastructure will involve many software tools, system software components (e.g., for grid computing, compilers and runtime systems, visualization, program development environments, distributed scalable and parallel file systems, human computer interfaces, highly scalable operating systems, system management software, parallelizing compilers for a variety of machine architectures, sophisticated schedulers), and other software building blocks that researchers can use to build their own cyberinfrastructure-enabled applications. Services, such as those needed to maintain software on multiple platforms and provide for authentication and access control, must be supported through the equivalent of help-desk facilities.

From the committee's perspective, the primary value of cyberinfrastructure resides in what it enables with respect to data management and analysis. Thus, in a biological context, machine-readable terminologies, vocabularies, ontologies, and structured grammars for constructing biological sentences are all necessary higher-level components of cyberinfrastructure as tools to help manage and analyze data (discussed in Section 4.2). High-end computing is useful in specialized applications but, by comparison to tools for data management and analysis, lacks broad applicability across multiple fields of biology.

7.1.2 Why Is Cyberinfrastructure Relevant?

The Atkins panel noted that the lack of a ubiquitous cyberinfrastructure for science and engineering research carries with it some major risks and costs. For example, when coordination is difficult, researchers in different fields and at different sites tend to adopt different formats and representations of key information. As a result, their reconciliation or combination becomes difficult to achieve—and hence disciplinary (or subdisciplinary) boundaries become more difficult to break down. Without systematic archiving and curation of intermediate research results (as well as the polished and reduced publications), useful data and information are often lost. Without common building blocks, research groups build their own application and middleware software, leading to wasted effort and time.

As a field, biology faces all of these costs and risks. Indeed, for much of its history, the organization of biological research could reasonably be regarded as a group of more or less autonomous fiefdoms. Unifying biological research into larger units of aggregation is not a plausible strategy today, and so the federation and loose coordination enabled by cyberinfrastructure seem well suited to provide the major advantages of integration while maintaining a reasonably stable large-scale organizational structure.

Furthermore, well-organized, integrated, synthesized information is increasingly valuable to biological research (Box 7.1). In an era characterized by data-intensive research observations, collecting,

Box 7.1
A Cyberinfrastructure View: Envisioning and Empowering Successes for 21st Century Biological Sciences

Creating and sustaining a comprehensive cyberinfrastructure (CI; the pervasive applications of all domains of scientific computing and information technology) are as relevant and as required for biology as for any science or intellectual endeavor; in the advances that led to today's opportunity, the National Science Foundation's Directorate for Biological Sciences (NSF BIO) made numerous, ad hoc contributions, and now can integrate its efforts to build the complete platforms needed for 21st century biology. Doing so will accelerate progress in extraordinary ways.

The time has arrived for creating a CI for all of the sciences, for research and education, and NSF will lead the way. NSF BIO must co-define the extent and fine details of the NSF structure for CI, which will involve major internal NSF partnerships and external partnerships with other agencies, and will be fully international in scope.

Only the biological sciences have seen advances as remarkable, sustained, and revolutionary as those in computer and information sciences. Only in the past few years has the world of computing and information technology reached the level of being fully applicable to the wide range of cutting-edge themes characteristic of biological research. Multiplying the exponentials (of continuing advances in computing and bioscience) through deep partnerships will inevitably be exciting beyond any anticipation.

The stretch goals for the biological sciences community include both community-level involvement and realization of the complete spectrum of CI, namely, people and training, instrumentation, collaborations, advanced computing and networking, databases and knowledge management; and analytical methods (modeling and simulation).

NSF BIO must:

- Invest in people;
- Ensure science pull, technology push;
- Stay the course;
- Prepare for the data deluge;
- Enable science targets of opportunity;
- Select and direct the technology contributions; and
- Establish national and international partnerships.

The biology community must decide how it can best interact with the quantitative science community, where and when to intersect with computational sciences and technologies, how to cooperate on and contribute to infrastructure projects, and how NSF BIO should partner administratively. An implementation meeting, as well as briefings to the community through professional societies, will be essential.

For NSF BIO to underestimate the importance of cyberinfrastructure for biology, or fail to provide fuel over the entire journey, would severely retard progress and be very damaging for the entire national and international biological sciences community.

SOURCE: Adapted from Subcommittee on 21st Century Biology, NSF Directorate for Biological Sciences Advisory Committee, *Building a Cyberinfrastructure for the Biological Sciences 2005 and Beyond: A Roadmap for Consolidation and Exponentiation,* a workshop report, July 14-15, 2003.

managing, and connecting data from various modalities and on multiple scales of biological systems, from molecules to ecosystems, are essential to turn that data into information. Each biological subdiscipline also now requires the tools of information technology to probe that information, to interconnect experimental observations and modeling, and to contribute to an enriched understanding or knowledge. The expansion of biology into discovery and synthetic analysis, that is, genome-enabled biology and systems biology as well as the hardening of many biological research tools into high-throughput pipelines, serves also to drive the need for cyberinfrastructure in biology.

Box 7.2 illustrates existing efforts in the development of cyberinfrastructure for biology that are relevant. Note that the examples span a wide range of subfields within biology, including proteomics (PDB), ecology (NEON and LTER), neuroscience (BIRN), and biomedicine (NBCR).

Data repositories and digital libraries are discussed in Chapter 3. The discussion below focuses primarily on computing and networking.

Box 7.2
Examples of Possible Elements of a Cyberinfrastructure for Biology

Pacific Rim Application and Grid Middleware Assembly

The Pacific Rim Application and Grid Middleware Assembly (PRAGMA) is a collaborative effort of 15 institutions around the Pacific Rim. PRAGMA's mission is to establish sustained collaborations and advance the use of grid technologies among a community of investigators working with leading institutions around the Pacific Rim. To fulfill this mission, PRAGMA hosts a series of workshop for members to focus on developing applications and on developing a testbed for these applications. Current applications include workflows in biology (protein annotation); linking via Web services climate data (working with some Long-Term Ecological Research [LTER] Network sites in the United States and East Asia Pacific region [ILTER]); running solvation models; and extending telescience application to more institutions.

The Protein Data Bank

The Protein Data Bank (PDB) was established in 1971 as a computer-based archival resource for macromolecular structures. The purpose of the PDB was to collect, standardize, and distribute atomic coordinates and other data from crystallographic studies. In 1977 the PDB listed atomic coordinates for 47 macromolecules. In 1987, the number began to increase rapidly at a rate of about 10 percent per year due to the development of area detectors and widespread use of synchrotron radiation; by April 1990, atomic coordinate entries existed for 535 macromolecules. Commenting on the state of the art in 1990, Holbrook and colleagues [citation omitted] noted that crystal determination could require one or more man-years. As of 1999, the Biological Macromolecule Crystallization Database (BMCD) of the PDB contain[ed] entries for 2,526 biological macromolecules for which diffraction quality crystals had been obtained. These include proteins, protein-protein complexes, nucleic acids, nucleic acid-nucleic acid complexes, protein-nucleic acid complexes, and viruses. In July 2004, the PDB held information on 26,144 structures (23,676 proteins, peptides, and viruses; 1,338 nucleic acids; 1,112 protein/nucleic acid complexes; and 18 carbohydrates).

The National Center for Biotechnology Information

The National Center for Biotechnology Information (NCBI), part of NIH's National Library of Medicine, has been charged with creating automated systems for storing, analyzing, and facilitating the use of knowledge about molecular biology, biochemistry, and genetics. In addition to GenBank, NCBI curates the Online Mendelian Inheritance in Man (OMIM), the Molecular Modeling Database (MMDB) of three-dimensional protein structures, the Unique Human Gene Sequence Collection (UniGene), the Taxonomy Browser, and the Cancer Genome Anatomy Project (CGAP), in collaboration with the National Cancer Institute. NCBI's retrieval system, Entrez, permits linked searches of the databases, while a variety of tools have been developed for data mining, sequence analysis, and three-dimensional structure display and similarity searching. NCBI's senior investigators and extended staff collaborate with the external research community to develop novel algorithms and research approaches that have transformed computational biology and will enable further genomic discoveries.

EUROGRID's Bio GRID

Funded by the European Commission, Bio GRID is intended to help biologists and chemists who are not familiar with high-performance computing (HPC) execution systems by developing intuitive user interfaces for selected biomolecular modeling packages and creating compatibility interfaces between the packages and their databases through Bio GRID's UNICORE platform. The UNICORE system will allow investigators to streamline their work processes, connect to Internet-accessible databases, and run a number of quantum chemistry and molecular dynamics software programs developed as plug-ins by Bio GRID's staff.

The NSF National Ecological Observatory Network (NEON)

NEON is a continental-scale research instrument consisting of geographically distributed networked infrastructure, with lab and field instrumentation; site-based experimental infrastructure; natural history archive facilities; and computational, analytical, and modeling capabilities. NEON is intended to transform ecological research by enabling studies on major environmental challenges at regional to continental scales. Scientists and engineers use NEON to conduct real-time ecological studies spanning all levels of biological organization and many temporal and geographical scales. NEON's synthesis, computation, and visualization infrastructure constitutes a virtual laboratory that enables the development of a predictive understanding of the direct effects and feedbacks between environmental change and biological processes.

The NSF Long-Term Ecological Research Network (LTER)

Since 1980, NSF has supported the Long-Term Ecological Research (LTER) Network. The LTER program is characterized by long temporal and broad spatial scales of research and fosters ecological comparisons among 26 U.S. sites that illustrate the importance of comprehensive analyses of ecosystems and of distinguishing system features across multiple scales of time and space. Data collected at each site are accessible to other scientists and the general public, and the LTER network works with other research institutions to standardize information management practices to achieve network- and community-wide data integration, facilitating data exchange and advancing data analysis and synthesis. LTER-supported work has included efforts in climate variability and ecosystem response, standardization of protocols for measuring soil properties for long-term ecological research, synthesis of global data on winter ice duration on lakes and rivers, and comparisons of ecosystem productivity, among others.

SOURCES: *PRAGMA:* material adapted from http://www.pragma-grid.net.

PDB: material pre-2004 excerpted from T. Lenoir, "Shaping Biomedicine as an Information Science," pp. 27-45 in *Proceedings of the 1998 Conference on the History and Heritage of Science Information Systems*, M. Bowden, T. Hahn, and R. Williams, eds., ASIS Monograph Series, Information Today, Inc., Medford, NJ, 1999, available at http://www.stanford.edu/dept/HPST/TimLenoir/Publications/Lenoir_BioAsInfoScience.pdf. Information for 2004 taken from Protein Data Bank Annual Report 2004, available at http://www.rcsb.org/pdb/annual_report04.pdf.

NCBI: material adapted from http://www.ncbi.nlm.nih.gov.

Bio GRID: material adapted from http://www.eurogrid.org.

NEON: material adapted from http://www.nsf.gov/bio/neon/.

LTER: material adapted from the LTER brochure, available at http://intranet.lternet.edu/archives/documents/Publications/brochures/lter_brochure.pdf.

7.1.3 The Role of High-performance Computing

Loosely speaking, processing capability refers to the speed with which a computational solution to a problem can be delivered. High processing capability is generally delivered by computing units operating in parallel and is generally dependent on two factors—the speed with which individual units compute (usually measured in operations per second) and the communications bandwidth between individual units. If a problem can be partitioned so that each subcomponent can be processed independently, then no communication at all is needed between individual computing units. On the other hand, as the dependence of one subcomponent on others increases, so does the amount of communications required between computing units.

Many biological applications must access large amounts of data. Furthermore, because of the combinatorial nature of the exploration required in these applications (i.e., the relationships between different pieces of data is not known in advance and thus all possible combinations are a priori possible), assumptions of locality that can be used to partition problems with relative ease (e.g., in computational fluid dynamics problems) do not apply, and thus the amount of data exchange increases. One estimate of the magnitude of the data-intensive nature of a biological problem is that a comparison of two of the smallest human chromosomes using the best available dynamic programming algorithm allowing for substitutions and gaps would require hundreds of petabytes of memory and hundred-petaflop processors.[2]

Thus, in supercomputers intended for biological applications, speed in computation and in communication are both necessary—and many of today's supercomputing architectures are thus inadequate for these applications.[3] Note that communications issues deal both with interprocessor communications (e.g., comparing sequences between processors, dividing long sequences among multiple processors) and traditional input-output (e.g., searching large sequence libraries on disk, receiving many requests at a time from the outside world). When problems involve large amounts of data exchange, communications become increasingly important.

Greater processing capability would enable the attack of many biologically significant problems. Today, processing capability is adequate to sequence and assemble data from a known organism. To some extent, it is possible to find genes computationally (as discussed in Chapter 4), but the accuracy of today's computationally limited techniques is modest. Simulations of interesting biomolecular systems can be carried out routinely for about hundreds of thousands of atoms for tens of nanoseconds. Order-of-magnitude increases (perhaps even two or three orders of magnitude) in processing capability would enable great progress in problem domains such as protein folding (ab initio prediction of three-dimensional structure from one-dimensional sequence information), simulation methods based on quantum mechanics that can provide more accurate predictions of the detailed behavior of interesting biomolecules in solution,[4] simulations of large numbers of interacting macromolecules for times of biological interest (i.e., for microseconds and involving millions of atoms), comparative genomics (i.e., finding similar genetic sequences across the genomes of different organisms—the multiple sequence alignment problem), proteomics (i.e., understanding the combinatorially large number of interactions between gene products), predictive and realistic simulations of biological systems ranging from cells to ecosystems), and phylogenetics (the reconstruction of historical relationships between species or individuals). Box 7.3 provides some illustrative applications of high-performance computing in life sciences research.

Any such estimate of the computing power needed to solve a given problem depends on assumptions about how a solution to that problem might be structured. Different ways of structuring a problem

[2] Shankar Subramanian, University of California, San Diego, personal communication, September 24, 2003.

[3] This discussion of communications issues is based on G.S. Heffelfinger, "Supercomputing and the New Biology," PowerPoint presentation at the AAAS Annual Meeting, Denver, CO, February 13-18, 2003.

[4] A typical problem might be the question of enzymes that exhibit high selectivity and high catalytic efficiency, and a detailed simulation might well provide insight into the related problem of designing an enzyme with novel catalytic activity. Simulations based on classical mechanics treat molecules essentially as charged masses on springs. These simulations (so-called molecular dynamics simulations) have had some degree of success, but lead to seriously inaccurate results where ions must interact in water or when the breaking or forming of bonds must be taken into account. Simulations based on quantum mechanics model molecules as collections of nuclei and electrons and entail solving of quantum mechanical equations governing the motion of such particles; these simulations offer the promise of much more accurate simulations of these processes, although at a much higher computational cost. These comments are based on excerpts from a white paper by M. Colvin, "Quantum Mechanical Simulations of Biochemical Processes," presented at the National Research Council's Workshop on the Future of Super-computing, Lawrence Livermore National Laboratory, Santa Fe, NM, September 26-28, 2003. See also "Biophysical Simulations Enabled by the Ultrasimulation Facility," available at http://www.ultrasim.info/doe_docs/Biophysics_Ultrasimulation_White_Paper_4-1-03.pdf.

Box 7.3
Grand Challenges in Computational Structural and Systems Biology

The Onset of Cancer

It is well known that cancer develops when cells receive inappropriate signals to multiply, but the details of cell signaling are not well understood. For example, activation of the epidermal growth factor signaling pathway is under the control of growth factors that bind to a receptor site on the exterior of a cell. Binding of the receptor initiates a cascade of protein conformational changes through the cell membrane, involving a complex rearrangement of many different proteins, including the Ras enzyme. The Ras enzyme is a molecular switch that can initiate a cascade of protein kinases that in turn transfer the external signal to the cell nucleus where it controls cell proliferation and differentiation. Disruption of this signaling pathway can have dire consequences as illustrated by the finding that mutations of the Ras enzyme have been found in 30 percent of human tumors. Because computer simulations can provide atomic-level detail that is difficult or impossible to obtain from experimental studies, computational studies are essential. However, this requires the modeling of an extremely large complex of biomolecules, including bilayer lipid membranes, transmembrane proteins, and a complex of many intercellular kinases, and thousands of molecules of waters of solvation.

Environmental Remediation

Microbes may be able to contribute to the cleanup of polluted sites by concentrating waste materials or degrading them into nontoxic form. Understanding the role of gram-negative bacteria in moderating subsurface reduction-oxidation chemistry and the role of such systems in bioremediation technologies requires the study of how cell walls, including many transmembrane protein substituents, interact with extracellular mineral surfaces and solvated atomic and molecular species in the environment. Simulations of these processes requires that many millions of atoms be included.

Degradation of Toxic Chemical Weapons

Computational approaches can be used for the rational redesign of enzymes to degrade chemical agents. An example is the enzyme phosphotriesterase (PTE), which could be used to degrade nerve gases. Combined experimental and computational efforts can be used to develop a series of highly specific PTE analogues, redesigned for optimum activity at specific temperatures, or for optimum stability and activity in nonaqueous, low-humidity environments or in foams, for improved degradation of warfare neurotoxins. Advanced computations can also facilitate the design of better reactivators of the enzyme acetylcholinesterase (AChE) that can be used as more efficient therapeutic agents against highly toxic phosphoester compounds such as the nerve warfare agents DFP (diisopropyl fluorophosphate), sarin, and soman and insecticides such as paraoxon. AChE is a key protein in the hydrolysis of acetylcholine, and inhibition of AChE through a phosphorylation reaction with such phosphoesters can rapidly lead to severe intoxication and death.

Multiscale Physiological Modeling of the Heart

The heart has a characteristic volume of around 60 cm^3. At a resolution of 0.1 mm, a grid of some 6×10^7 cells is required. If 100 variables are associated with each cell, 10 floating point operations are needed for each time step in a simulation, and the time resolution is around 1 ms (a single heartbeat has a duration around 1 second), a computing throughput of 6×10^{13} floating point operations per second (60 teraflops) is necessary. In addition, a flexible and composable simulation infrastructure is required. For example, for a spatially distributed system, only a representative and relatively small subset of substructures can be represented in the model explicitly, because it is not feasible to model all of them. Contributions of the substructures missing from the model are inferred by an interpolative process. For practical purposes, it will not be known in advance how much and what kinds of detail will be necessary for a useful simulation; the same a priori ignorance also characterizes the nature and extent of the communications required between different levels of the simulation. Thus, the infrastructure must support easy experimentation in which different amounts of detail and different degrees of communication can be explored.

SOURCE: The first three examples are adapted with minimal change from D.A. Dixon, T.P. Straatsma, and T. Head-Gordon, "Grand Challenges in Computational Structural and Systems Biology," available at http://www.ultrasim.info/doe_docs/ESC-response.bio.dad.pdf.

solution often result in different estimates for the required computing power, and for any complex problem, the "best" structuring may well not be known. (Different ways of structuring a problem may involve different algorithms for its solution, or different assumptions about the nature of the biologically relevant information.) Furthermore, the advantage gained through algorithm advances, conceptual reformulations of the problem, or different notions about the answers being sought is often comparable to advantages from hardware advances, and sometimes greater. On the other hand, for decades computational scientists have been able to count on regular advances in computing power that accrued "for free," and whether or not scientists are able to develop new ways of looking at a given problem, hardware-based advances in computing are likely to continue.

Three types of computational problem in biology must be distinguished.[5] Problems such as protein folding and the simulation of biological systems are similar to other simulation problems that involve substantial amounts of "number crunching." A second type of problem entails large-scale comparisons or searches in which a very large corpus of data—for example, a genomic sequence or a protein database—is compared against another corpus, such as another genome or a large set of unclassified protein sequences. In this kind of problem, the difficult technical issues involve the lack of good software for broadcast and parallel access disk storage subsystems. The third type of problem involves single instances of large combinatorial problems, for example, finding a particular path in a very large graph. In these problems, computing time is often not an issue if the object can be modeled in the memory of the machine. When memory is too small, the user must write code that allows for efficient random access to a very large object—a task that significantly increases development time and even under the best of circumstances can degrade performance by an order of magnitude.

The latter two types of problem often entail the consideration of large numbers of biological objects (cells, organs, organisms) characterized by high degrees of individuality, contingency, and historicity. Such problems are typically found in investigations involving comparative and functional genomics and proteomics, which generally involve issues such as discrete combinatorial optimization (e.g., the multiple sequence alignment problem) or pattern inference (e.g., finding clusters or other patterns in high-dimensional datasets). Algorithms for discrete optimization and pattern inference are often NP-hard, meaning that the time to find an optimal solution is far too long (e.g., longer than the age of the universe) for a problem of meaningful size, regardless of the computer that might be used or that can be foreseen. Since optimal solutions are not in general possible, heuristic approaches are needed that can come reasonably close to being optimal—and a considerable degree of creativity is involved in developing these approaches.

Historically, another important point has been that the character of biological data is different from that of data in fields such as climate modeling. Many simulations of nonbiological systems can be composed of multiple repeating volume elements (i.e., a mesh that is well suited for finding floating point solutions of partial differential equations that govern the temporal and spatial evolution of various field quantities). By contrast, some important biological data (e.g., genomic sequence data) are characterized by quantities that are better suited to integer representations, and biological simulations are generally composed of heterogeneous objects. However, today, the difference in speed between integer operations and floating point operations is relatively small, and thus the difference between floating point and integer representations is not particularly significant from the standpoint of supercomputer design.

Finally, it is important to realize that many problems in computational biology will never be solved by increased computational capability alone. For example, some problems in systems biology are combinatorial in nature, in the sense that they seek to compare "everything against everything" in a search for previously unknown correlations. Search spaces that are combinatorially large are so large that even

[5] The description of problem types in this paragraph draws heavily from G. Myers, "Supercomputing and Computational Molecular Biology," presented at the NRC Workshop on the Future of Supercomputing, Santa Fe, NM, September 26-28, 2003.

with exponential improvements in computational speed, methods other than exhaustive search must be employed as well to yield useful results in reasonable times.[6]

The preceding discussion for the life sciences focuses on the large-scale computing needs of the field. Yet these are hardly the only important applications of computing, and rapid innovation is likely to require information technology on many scales. For example, researchers need to be able to explore ideas on local computers, albeit for scaled-down problems. Only after smaller-scale explorations are conducted do researchers have the savvy, the motivation, and the insight needed for meaningful use of high-end cyberinfrastructure. Researchers also need tools that can facilitate quick and dirty tasks, and working knowledge of spreadsheets or Perl programming can be quite helpful. For this reason, biologists working at all scales of problem size will be able to benefit from advances in and knowledge of information technology.

7.1.4 The Role of Networking

As noted in Chapter 3, biological data come in large quantities. High-speed networking (e.g., one or two orders of magnitude faster than that available today) would greatly facilitate the exchange of certain types of biological data such as high-resolution imaging as well as enable real-time remote operation of expensive instrumentation. High-speed networking is critical for life science applications in which large volumes of data change or are created rapidly, such as those involving imaging or remote operation of instrumentation.[7]

The Internet2 effort also includes the Middleware Initiative (I2-MI), intended to facilitate the creation of interoperable middleware infrastructures among the membership of Internet2 and related communities.[8] Middleware generally consists of sets of tools and data that help applications use networked resources and services. The availability of middleware contributes greatly to the interoperability of applications and reduces the expense involved in developing those applications. I2-MI develops middleware to provide services such as identifiers (labels that connect a real-world subject to a set of computerized data); authentication of identity; directories that index elements that applications must access; authorization of services for users; secure multicasting; bandwidth brokering and quality of service; and coscheduling of resources, coupling data, networking, and computing together.

7.1.5 An Example of Using Cyberinfrastructure for Neuroscience Research

The Biomedical Informatics Research Network (BIRN) project is a nationwide effort by National Institutes of Health (NIH)-supported research sites to merge data grid and computer grid cyberinfrastructure into the workflows of biomedical research. The Brain Morphometry BIRN, one of the testbeds driving the development of BIRN, has undertaken a project that uses the new technology by integrating data and analysis methodology drawn from the participating sites. The Multi-site Imaging Research in the Analysis of Depression (MIRIAD) project (Figure 7.1) applies sophisticated image processing of a dataset of magnetic resonance imaging (MRI) scans of a longitudinal study of elderly subjects. The subjects include patients who enroll in the study with symptoms of clinical depressions

[6] Consider the following example. The human genome is estimated to have around 30,000 genes. If the exploration of interest is assumed to be 5 genes operating together, there are approximately 3×10^{20} possible combinations of 30,000 genes in sets of 5. If the assumption is that 6 genes may operate together, there are on the order of 10^{26} possible combinations (the number of possible combinations of n genes in groups of k is given by $n!/(k!(n-k)!)$, which for large n and small k reduces to $n^k/k!$).

[7] In the opposite extreme case, in which enormous volumes of data never change, it is convenient rather than essential to use electronic or fiber links to transmit the information—for a small fraction of the cost of high-speed networks, media (or even entire servers!) can be sent by Federal Express more quickly than a high-speed network could transmit the comparable volume of information. See, for example, Jim Gray et al., *TeraScale SneakerNet: Using Inexpensive Disks for Backup, Archiving, and Data Exchange*, Microsoft Technical Report, MS-TR-02-54, May 2002, available at ftp://ftp.research.microsoft.com/pub/tr/tr-2002-54.pdf.

[8] See http://middleware.internet2.edu/overview/.

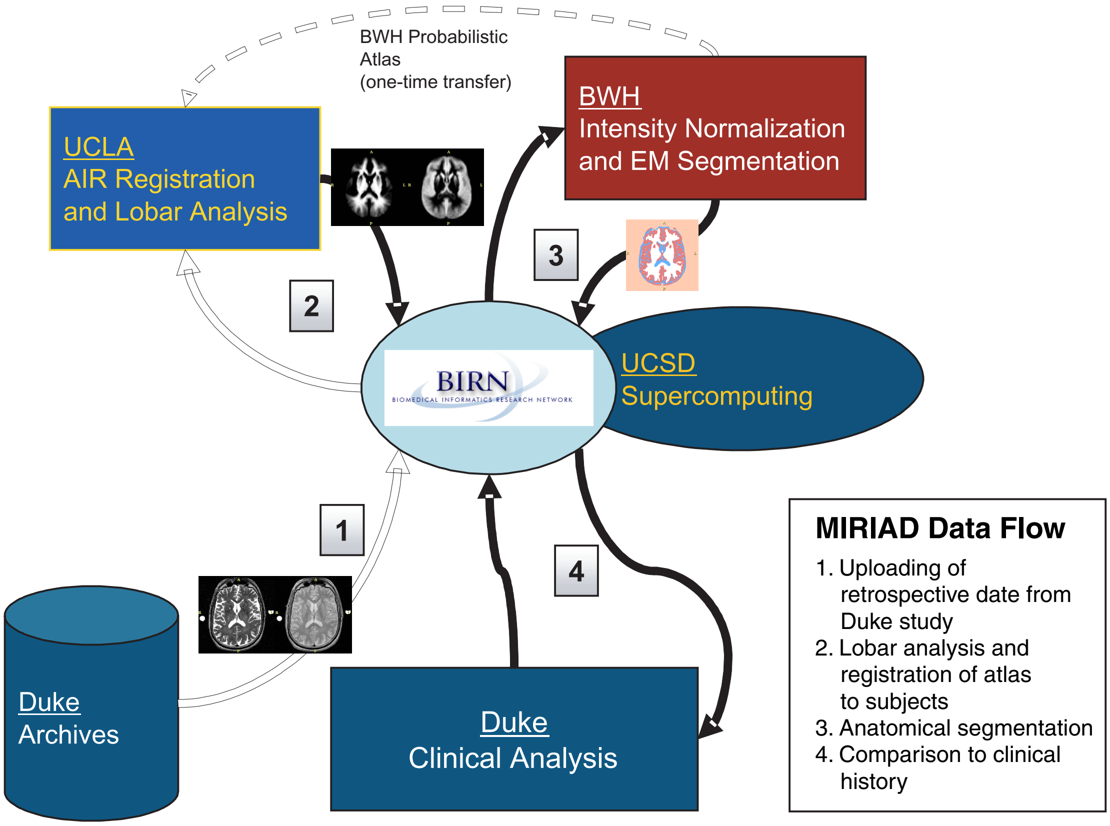

BWH Probabilistic
Atlas
(one-time transfer)
UCLA
AIR Registration
and Lobar Analysis
BWH
Intensity Normalization
and EM Segmentation
3
2
BIRN
BIOMEDICAL INFORMATICS RESEARCH NETWORK
UCSD
Supercomputing
1
4
Duke
Archives
Duke
Clinical Analysis
MIRIAD Data Flow
1. Uploading of retrospective date from Duke study
2. Lobar analysis and registration of atlas to subjects
3. Anatomical segmentation
4. Comparison to clinical history

FIGURE 7.1 Steps in data processing in the BIRN MIRIAD project.

1. T2-weighted and proton density (PD) MRI scans from the Duke University longitudinal study are loaded into the BIRN data archive (data grid), accessible by members of the MIRIAD group for analysis using the computer resources at the University of California, San Diego (UCSD) and the San Diego Supercomputer Center (compute grid).

2. The Laboratory of Neuro Imaging at the University of California, Los Angeles (UCLA) performs a nonlinear registration to define the three-dimensional geometric mapping between each subject and a standard brain atlas that encodes the probabilities of each tissue class at each location in the brain.

3. The Surgical Planning Laboratory at Brigham and Women's Hospital (BWH) then applies an intensity normalization and expectation-maximization algorithm to combine the original image pixel intensities (T2 and PD) and the tissue probabilities to label each point in the images and to calculate the overall volumes of tissue classes.

4. Duke performs statistical tests on the image-processing results to assess the predictive value of the brain morphometry measurements with respect to clinical outcome.

and age-matched controls. Some of the depression patients go on to develop Alzheimer's disease (AD) and the goal of the MIRIAD project is to measure the changes in brain images, specifically volume changes in cortical and subcortical gray matter, that correlate with clinical outcome.

Of particular significance from the standpoint of cyberinfrastructure, the MIRIAD project is distributed among four separate sites: Duke University Neuropsychiatric Imaging Research Laboratory, Brigham and Women's Hospital Surgical Planning Laboratory, University of California, Los Angeles Laboratory of Neuro Imaging, and University of California, San Diego BIRN. Each of these sites has responsibility for some substantive part of the work, and the work would not be possible without the BIRN infrastructure to coordinate it.

7.2 DATA ACQUISITION AND LABORATORY AUTOMATION

As noted in Chapter 3, the biology of the 21st century will be data-intensive across a wide range of spatial and temporal scales. Today's high-throughput data acquisition technologies depend on parallelization rather than on reducing the time needed to take individual data points. These technologies are capable of carrying out global (or nearly global) analyses, and as such they are well suited for the rapid and comprehensive assessment of biological system properties and dynamics. Indeed, in 21st century biology, many questions are asked because relevant data can be obtained to answer them. Whereas earlier researchers automated existing manual techniques, today's approach is more oriented toward techniques that match existing automation.

7.2.1 Today's Technologies for Data Acquisition[9]

Some of today's data acquisition technologies include the following:[10]

- *DNA microarrays*. Microarray technology enables the simultaneous interrogations of a human genomic sample for complete human transcriptomes, provided that the arrays do not contain only putative protein coding regions. The oligonucleotide microarray can identify single-nucleotide differences and distinguish mRNAs from individual members of multigene families, characterize alternatively spliced genes, and identify and type alternative forms of single-nucleotide polymorphisms. Microarrays are also used to observe in vitro protein-DNA binding events and to do comparative genome hybridization (CGH) studies. Box 7.4 provides a close-up of microarrays.
- *Automated DNA sequencers*. Prior to automated sequencing, the sequencing of DNA was performed manually, at many tens (up to a few hundred) of bases per day.[11] In the 1970s, the development of restriction enzymes, recombinant DNA techniques, gene cloning techniques, and polymerase chain reaction (PCR) contributed to increasing amounts of data on DNA, RNA, and protein sequences. More than 140,000 genes were cloned and sequenced in the 20 years from 1974 to 1994, many of which were human genes. In 1986, an automated DNA sequencer was first demonstrated that sequenced 250 bases per day.[12] By the late 1980s, the NIH GenBank database (release 70) contained more than 74,000 sequences, while the Swiss Protein database (Swiss-Prot) included nearly 23,000 sequences. In addition, protein databases were doubling in size every 12 months. Since 1999, more advanced models of automated DNA sequencer have come into widespread use.[13] Today, a state-of-the-art automated sequencer can produce on the order of a million base pairs of raw DNA sequence data per day. (In addition, technologies are available that allow the parallel processing of 16 to 20 residues at a time.[14] These enable the determination of complete transcriptomes in individual cell types from organisms whose genome is known.)
- *Mass spectroscopy*. Mass spectroscopy (MS) enables the in-quantity identification and quantification of large numbers of proteins.[15] Used in conjunction with genomic information, MS information can be used to identify and type single-nucleotide polymorphisms. Some implementations of mass spec-

[9] Section 7.2.1 is adapted from T. Ideker, T. Galitski, and L. Hood, "A New Approach to Decoding Life: Systems Biology," *Annual Review of Genomics and Human Genetics* 2:343, 2001.

[10] Adapted from T. Ideker et al., "A New Approach to Decoding Life," 2001.

[11] L. Hood and D.J. Galas, "The Digital Code of DNA," *Nature* 421(6921):444-448, 2003.

[12] L.M. Smith, J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, C. Heiner, et al., "Fluorescence Detection in Automated DNA Sequence Analysis," *Nature* 321(6071):674-679, 1986. (Cited in Ideker et al., 2001.)

[13] L. Rowen, S. Lasky, and L. Hood, "Deciphering Genomes Through Automated Large Scale Sequencing," *Methods in Microbiology*, A.G. Craig and J.D. Hoheisel, eds., Academic Press, San Diego, CA, 1999, pp. 155-191. (Cited in Ideker et al., 2001.)

[14] S. Brenner, M. Johnson, J. Bridgham, G. Golda, D.H. Lloyd, D. Johnson, S. Luo, et al., "Gene Expression Analysis by Massively Parallel Signature Sequencing (MPSS) on Microbead Arrays," *Nature Biotechnology* 18(6):630-634, 2000. (Cited in Ideker et al., 2001.)

[15] J.K. Eng, A.L. McCormack, and J.R.I. Yates, "An Approach to Correlate Tandem Mass Spectral Data of Peptides with Amino Acid Sequences in a Protein Database," *Journal of the American Society for Mass Spectrometry* 5:976-989, 1994. (Cited in Ideker et al., 2001.)

Box 7.4
Microarrays: A Close-up

A "classical" microarray typically consists of single-stranded pieces of DNA from virtually an entire genome placed physically in tiny dots on a flat surface and labeled with a fluorescent dye. (Lithographic techniques used to develop semiconductor chips are now used to deposit the DNA on a silicon chip that can later be read optically.) In a microarray experiment, messenger RNA (mRNA) from a cell of interest is extracted and placed in contact with the prepared surface. If the sample contains mRNA corresponding to the DNA on one or more of the dots on the surface, the molecules will bind and the dye will fluoresce. Because the mRNA represents the fraction of genes from the sample that have been transcribed from DNA into mRNA, the resulting fluorescent dots on the surface are a visual indicator of gene expression (or transcription) in the cell's genome. Different intensities of the dots reflect greater or lesser levels of transcription of particular genes.

Obtaining the maximum value from a microarray experiment depends on the ability to correlate the data from a microarray experiment per se with extensive data that identify or classify the genes by other characteristics. In the absence of such data, any given microarray experiment merely points out the fact that some genes are expressed to a greater extent than others in a particular experimental situation.

Protein microarrays can identify protein-protein (and protein-drug) interactions among some 10,000 proteins at once.[1] As described by Templin,[2]

> [protein] microarray technology allows the simultaneous analysis of thousands of parameters within a single experiment. Microspots of capture molecules are immobilized in rows and columns onto a solid support and exposed to samples containing the corresponding binding molecules. Readout systems based on fluorescence, chemiluminescence, mass spectrometry, radioactivity or electrochemistry can be used to detect complex formation within each microspot. Such miniaturized and parallelized binding assays can be highly sensitive, and the extraordinary power of the method is exemplified by array-based gene expression analysis. In these systems, arrays containing immobilized DNA probes are exposed to complementary targets and the degree of hybridization is measured. Recent developments in the field of protein microarrays show applications for enzyme-substrate, DNA-protein and different types of protein-protein interactions. Here, we discuss theoretical advantages and limitations of any miniaturized capture-molecule-ligand assay system and discuss how the use of protein microarrays will change diagnostic methods and genome and proteome research.

[1] See G. MacBeath and S.L. Schreiber, "Printing Proteins as Microarrays for High-Throughput Function Determination," *Science* 289(5485): 1760-1763, 2000.

[2] Reprinted by permission from M.F. Templin, D. Stoll, M. Schrenk, P.C. Traub, C.F. Vohringer, and T.O. Joos, "Protein Microarray Technology," *Trends in Biotechnology* 20(4):160-166, 2002. Copyright 2002 Elsevier.

NOTE: An overview of microarray technology is available on a private Web site created by Leming Shi: http://www.gene-chips.com/. See also http://www.genome.gov/10000533 and P. Gwynne and G. Page, "Microarray Analysis: The Next Revolution in Molecular Biology," special advertising supplement, *Science* 285, August 6, 1999, available at http://www.sciencemag.org/feature/e-market/benchtop/micro.shl.

troscopy today allow 1,000 proteins per day to be analyzed in an automated fashion, and there is hope that a next-generation facility will be able to analyze up to 1 million proteins per day.[16]

- *Cell sorters*. Cell sorters separate different cell types at high speed on the basis of multiple parameters. While microarray experiments provide information on average levels of mRNA or protein within a cell population, the reality is that these levels vary from cell to cell. Knowing the distribution of expression levels across cell types provides important information about the underlying control mechanisms and regulatory network structure. A state-of-the-art cell sorter can separate 30,000 elements per second according to 32 different parameters.[17]

[16] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, and R. Aebersold, "Quantitative Analysis of Complex Protein Mixtures Using Isotope-coded Affinity Tags," *Nature Biotechnology* 17(10):994-999, 1999. (Cited in Ideker et al., 2001.)

[17] See, for example, http://www.systemsbiology.org/Default.aspx?pagename=cellsorting.

Box 7.5
Applications of Embedded Network Sensor Systems

Marine Microorganisms[1]

Marine microorganisms such as viruses, bacteria, microalgae, and protozoa have a major impact on the ecology of the coastal ocean; present public health issues for coastal human populations as a consequence of the introduction of pathogenic microorganisms into these waters from land runoff, storm drains, and sewage outflow; and have the potential to contaminate drinking water supplies with harmful, pathogenic, or nuisance microbial species.

Today, the environmental factors that stimulate the growth of such microorganisms are still poorly understood. To understand these factors, scientists need to correlate environmental conditions with microorganismal abundances at the small spatial and temporal scales that are relevant to these organisms. For a variety of technological and methodological reasons, sampling the environment at the necessary high resolution and identifying microorganisms in situ in near-real time has not been possible in the past.

Habitat Sensing[2]

Understanding in detail the environmental, organismal, and cultural conditions, and the interactions between them, in natural and managed habitats is a problem of considerable biological complexity. Data must be captured and integrated across a wide range of spatial and temporal scales for chemical, physiological, ecological, and environmental purposes. For example, data of interest might include microclimate data; a video of avian behavioral activities related to climate, nesting, and reproduction; and data on soil moisture, nitrate, CO_2, temperature, and root-fungi activities in response to weather.

[1] Adapted from http://www.cens.ucla.edu/portal/aquatic_microbial_observing_syst.html.
[2] Adapted from http://deerhound.ats.ucla.edu:7777/portal/page?_pageid=54,42365,54_42372&_dad=portal&_schema=PORTAL.

- *Microfluidic systems*. Microfluidic systems, also known as micro-TAS (total analysis system), allow the rapid and precise measurement of sample volumes of picoliter size. These systems put onto a single integrated circuit all stages of chemical analysis, including sample preparation, analyte purification, microliquid handling, analyte detection, and data analysis.[18] These "lab-on-a-chip" systems provide portability, higher-quality and higher-quantity data, faster kinetics, automation, and reduction of sample and reagent volumes.
- *Embedded networked sensor (ENS) systems*. ENS systems are large-scale, distributed systems, composed of smart sensors embedded in the physical world, that can provide data about the physical world at unprecedented granularity. These systems can monitor and collect large volumes of information at low cost on such diverse subjects as plankton colonies, endangered species, and soil and air contaminants. Across a wide range of large-scale biological applications broadly cast, these systems promise to reveal previously unobservable phenomena. Box 7.5 describes some applications of ENS systems.

Finally, a specialized type of data acquisition technology is the hybrid measurement device that interacts directly with a biological sample to record data from it or to interact with it. As one illustration, contemporary tools for studying neuronal signaling and information processing include implantable probe arrays that record extracellularly or intracellularly from multiple neurons simultaneously.

[18] See, for example, http://www.eurobiochips.com/euro2002/html/agenda.asp. To illustrate the difficulty, consider the handling of liquids. Dilution ratios required for a process may vary by three or four orders of magnitude, and so an early challenge (now largely resolved successfully) is the difficulty of engineering an automated system that can dispense both 0.1-microliter and 1-milliliter volumes with high accuracy and in reasonable time periods.

Such arrays have been used in moths (*Manduca sexta)* and sea slugs (*Tritonia diomeda)* and, when linked directly to the electronic signals of a computer, essentially record and simulate the neural signaling activity occurring in the organism. Box 7.6 describes the dynamic clamp, a hybrid measurement device that has been invaluable in probing the behavior of neurons. Research on this interface will serve both to reveal more about the biological system and to represent that system in a format that can be computed.

Box 7.6
The Dynamic Clamp

The dynamic clamp is a device that mimics the presence of a membrane or synapse proximate to a neuron. That is, the clamp essentially simulates the electrical conductances in the network to which a neuron is ostensibly connected. During clamp operation, the membrane potential of a neuron is continuously measured and fed into a computer. The dynamic clamp program contains a mathematical model of the conductance to be simulated and computes the current that would flow through the conductance as a function of time. This current is injected into the neuron, and the cycle of membrane potential measurement, current computation, and current injection continues. This cycle enables researchers to study the effects of a membrane current or synaptic input in a biological cell (the neuron) by generating a hybrid system in which the artificial conductance interacts with the natural dynamic properties of the neuron.

The dynamic clamp can be used to mimic any voltage-dependent conductance that can be expressed in a mathematical model. Depending on the type of conductance, most applications can be grouped in one of the following categories:

1. *Generating artificial membrane conductances*. These may be voltage dependent or independent.
2. *Simulating natural stimuli*. The dynamic clamp can mimic natural conditions such as barrages of synaptic inputs to neurons in silent brain slices. Here, an artificial synaptic conductance trace is used to compute an artificial synaptic current from the momentary membrane potential of the postsynaptic neuron. That current is continuously injected into the neuron, and the effect of the artificial input on the activity of the neuron is assessed.
3. *Generating artificial synapses*. In a configuration where the dynamic clamp computer monitors the membrane potential of several neurons and computes and injects current through several output channels, the dynamic clamp can be used to create artificial chemical or electrotonic synaptic connections between neurons that are not connected in vivo or to modify the strength or dynamics of existing synaptic connections.
4. *Coupling of biological and model neurons*. The dynamic clamp can also be used to create hybrid circuits by coupling model and biological neurons through artificial synapses. In this application, the dynamic clamp computer continuously solves the differential equations that describe the model neuron and the synapses that connect it to the biological neuron.

The first application of the dynamic clamp involved the stimulation of a gamma-aminobutyric acid (GABA) response in a cultured stomatogastric ganglion neuron. This application illustrated that the dynamic clamp effectively introduces a conductance into the target neuron. Demonstration of an artificial voltage-dependent conductance resulted from simulation of the action of a voltage-dependent proctolin response on a neuron in the intact stomatogastric ganglion, which showed that shifts in the activation curve and the maximal conductance of the response produced different effects on the target neuron. Lastly, the dynamic clamp was used to construct reciprocal inhibitory synapses between two stomatogastric ganglion neurons that were not coupled naturally, illustrating that the dynamic clamp could be used to simulate new networks at will.

SOURCE: The description of a dynamic clamp is based heavily on A.A. Prinz, "The Dynamic Clamp a Decade After Its Invention," *Axon Instruments Newsletter* 40, February 2004, available at http://www.axon.com/axobits/AxoBits40.pdf. The description of the first application of the dynamic clamp is nearly verbatim from A.A. Sharp, M.B. O'Neil, L.F. Abbott, and E. Marder, "Dynamic Clamp: Computer-generated Conductances in Real Neurons," *Journal of Neurophysiology* 69(3):992-995, 1993.

7.2.2 Examples of Future Technologies

As powerful as these technologies are, new instrumentation and methodology will be needed in the future. These technical advances will have to reduce the cost of data acquisition by several orders of magnitude.

Consider, for example, the promise of genomically individualized medical care, which is based on the notion that treatment and/or prevention strategies for disease can be customized to groups of individuals smaller than the entire population, and perhaps ultimately groups as small as one. Because these groups will be identified in part by particular sets of genomic characteristics, it will be necessary to undertake the genomic sequencing of these individuals. The first complete sequencing of the human genome took 13 years and $2.7 billion. For broad use in the population at large, the cost of assembling and sequencing a human genome must drop to hundreds or thousands of dollars—a reduction in cost of 10^5 or 10^6 that would enable the completion of a human genome at such cost in a matter of days.[19]

Computation per se is expected to continue to drop in cost in accordance with Moore's law at least over the next decade. But automation of data acquisition will also play an enormous role in facilitating such cost reductions. For example, the laboratory of Richard Mathies at the University of California, Berkeley, has developed a 96-lane microfabricated DNA sequencer capable of sequencing at a rate of 1,700 bases per minute.[20] Using this technology, the complete sequencing of an individual 3-billion base genome would take 1,000 sequencer-days. Future versions will incorporate higher degrees of parallelism.

Similar advances in technology will help to reduce the cost of other kinds of biological research as well. A number of biological signatures useful for functional genomics have been susceptible to significantly greater degrees of automation, miniaturization, and multiplexing; these signatures are associated with electrophoresis, molecular microarrays, mass spectrometry, and microscopy.[21] Electrophoresis, molecular microarrays, and mass spectrometry provide more opportunities for multiplexed measurement (i.e., the simultaneous measurement of signatures from many molecules from the same source). Such multiplexing can reduce errors due to misalignment of unmultiplexed measures in space and/or time.

In general, the biggest payoffs in laboratory automation are those efforts that can address processes that involve physical material. Much work in biology involves multiple laboratory procedures that each call for multiple fluid transfers, heating and cooling cycles, and mechanical operations such as centrifuging, waiting, and imaging. When these procedures can be undertaken "on-chip," they reduce the amount of human interaction involved and thus the associated time and cost.

In addition, the feasibility of lab automation is closely tied to the extent to which human craft can be taken out of lab work. That is, because so much lab work must be performed by humans, the skills of the particular individuals involved matter a great deal to the outcomes of the work. A particular individual may be the only one in a laboratory with a "knack" for performing some essential laboratory procedure (e.g., interpretation of certain types of image, preparation or certain types of sample) with high reliability, accuracy, and repeatability.

[19] L.M. Smith, J.Z. Sanders, R.J. Kaiser, P. Hughes, C. Dodd, C.R. Connell, C. Heiner, et al., "Fluorescence Detection in Automated DNA Sequence Analysis," *Nature* 321(6071):674-679, 1986; L. Hood and D. Galas, "The Digital Code of DNA," *Nature* 421(6921):444-448, 2003. Note that done properly, the second complete sequencing of a human being would be considerably less difficult. The reason is that every member of a biological species has a DNA that is almost identical to that of every other member. In humans, the difference between DNA sequences of different individuals is about one base pair per thousand. (See special issues on the human genome: *Science* 291(5507) February 16, 2001; *Nature* 409(6822), February 15, 2001.) So, assuming it is known where to check for every difference, a reduction in effort of at least a factor of 10^3 is obtainable in principle.

[20] B.M. Paegel, R.G. Blazej, and R.A. Mathies, "Microfluidic Devices for DNA Sequencing: Sample Preparation and Electrophoretic Analysis," *Current Opinion in Biotechnology* 14(1):42-50, 2003, available at http://www.wtec.org/robotics/us_workshop/June22/paper_mathies_microfluidics_sample_prep_2003.pdf.

[21] G. Church, "Hunger for New Technologies, Metrics, and Spatiotemporal Models in Functional Genomics," available at http://recomb2001.gmd.de/ABSTRACTS/Church.html.

While reliance on individuals with specialized technical skills is often a workable strategy for an academic laboratory, it makes much less sense for any organization interested in large-scale production. For large-scale, cost-effective production, process automation is a sine qua non. When a process can be automated, it is generally faster to perform, more free from errors, more accurate, and less expensive.[22]

Some of the clearest success stories involve genomic technologies. For example, DNA sequencing was a craft at the start of the 1990s—today, automated DNA sequencing is common, with instruments to undertake such sequencing in high volume (a million or more base pairs per day) and even a commercial infrastructure to which sequencing tasks can be outsourced. Nevertheless, a variety of advanced sequencing technologies are being developed, primarily with the intent of lowering the cost of sequencing by another several orders of magnitude.[23]

An example of such a technology is pyrosequencing, which has also been called "sequencing by synthesis."[24] With pyrosequencing, the DNA to be sequenced is denatured to form a single strand and then placed in solution with a set of selected enzymes. In a cycle of individual steps, the DNA-enzyme solution is mixed with deoxynucleotide triphosphate molecules containing each of the four bases. When the base that is the complement to the next base on the target strand is added, the added base joins a forming complement strand and releases a pyrophosphate molecule. That molecule starts a reaction that ends with luciferin emitting a detectable amount of light. Thus, by monitoring the light output of the reaction (for example, with a CCD camera), it is possible to observe in real time which of the four bases has successfully matched.

454 Life Sciences has applied pyrosequencing to whole-genome analyses by taking advantage of its high parallelizability. Using a PicoTiter plate, a microfluidic system performs pyrosequencing on hundreds of thousands of DNA fragments simultaneously. Custom software analyzes the light emitted and stitches together the complete sequence. This approach has been used successfully to sequence the genome of an adenovirus,[25] and the company is expected to produce commercial hardware to perform whole-genome analysis in 2005.

A second success story is microarray technology, which historically has relied heavily on electrophoretic techniques.[26] More recently, techniques have been developed that do away entirely with electrophoresis. One approach relies instead on microbeads with different messenger RNAs on their surfaces (serving as probes to which targets bind selectively) and a novel sequencing procedure to

[22] The same can be said for many other aspects of lab work. In 1991, Walter Gilbert noted, "The march of science devises ever newer and more powerful techniques. Widely used techniques begin as breakthroughs in a single laboratory, move to being used by many researchers, then by technicians, then to being taught in undergraduate courses and then to being supplied as purchased services—or, in their turn, superseded. . . . Fifteen years ago, nobody could work out DNA sequences, today every molecular scientists does so and, five years from now, it will all be purchased from an outside supplier. Just this happened with restriction enzymes. In 1970, each of my graduate students had to make restriction enzymes in order to work with DNA molecules; by 1976 the enzymes were all purchased and today no graduate student knows how to make them. Once one had to synthesize triphosphates to do experiments; still earlier, of course, one blew one's own glassware." See W. Gilbert, "Towards a Paradigm Shift in Biology," *Nature* 349(6305):99, 1991.

[23] A review by Shendure et al. classifies emerging ultralow-cost sequencing technologies into one of five groups: microelectrophoretic methods (which extend and incrementally improve today's mainstream sequencing technologies first developed by Frederick Sanger); sequencing by hybridization; cyclic array sequencing on amplified molecules; cyclic array sequencing on single molecules; and noncyclical, single-molecule, real-time methods. The article notes that most of these technologies are still in the relatively early stages of development, but that they each have great potential. See J. Shendure, R.D. Mitra, C. Varma, and G.M. Church, "Advanced Sequencing Technologies: Methods and Goals," *Nature Reviews: Genetics* 5(5):335-344, 2004, available at http://arep.med.harvard.edu/pdf/Shendure04.pdf. Pyrosequencing, provided as an example of one new sequencing technology, is an example of cyclic array sequencing on amplified molecules.

[24] M. Ronaghi, "Pyrosequencing Sheds Light on DNA Sequencing," *Genome Research* 11(1):3-11, 2001.

[25] A. Strattner, "From Sanger to 'Sequenator'," *Bio-IT World*, October 10, 2003.

[26] Genes are expressed as proteins, and these proteins have different weights. Electrophoresis is a technique that can be used to determine the extent to which proteins of different weights are present in a sample.

Box 7.7
On Optical Mapping

Optical mapping is a single molecule based physical mapping technology, which creates an ordered restriction map by enumerating the locations of occurrences of a specific "restriction pattern" along a genome. Thus, by locating the same patterns in the sequence reads or contigs, optical maps can detect errors in sequence assembly, and determine the phases (i.e., chromosomal location and orientation) of any set of sequence contigs. Since the input genomic data that can be collected from a single DNA molecule by the best chemical and optical methods (such as those used in Optical Mapping) are badly corrupted by many poorly understood noise processes, this type of technology derives its utility through powerful probabilistic modeling used in experiment design and Bayesian algorithms that can recover from errors by using redundant data. In this way, optical mapping with Gentig, a powerful statistical map-assembly algorithm invented and implemented by the authors, has proven instrumental in completing many microbial genomic maps (*Escherichia coli*, *Yersinia pestis*, *Plasmodium falciparum*, *Deinococcus radiodurans*, *Rhodobacter sphaeroides*, etc.) as well as clone maps (DAZ locus of Y chromosome).

SOURCE: T. Anantharaman and B. Mishra, *Genomics via Optical Mapping* (I): 0-1 *Laws for Single Molecules*, S. Yancopoulos, ed., Oxford University Press, Oxford, England, 2005, in press.

identify specific microbeads.[27] Each bead can be interrogated in parallel, and the abundance of a given messenger RNA is determined by counting the number of beads with that mRNA on their surfaces. In addition to greatly simplifying the sample-handling procedure, this technique has two other important advantages: a direct digital readout of relative abundances (i.e., the bead counts) and throughput increases by more than a factor of 10 compared to other techniques.

A second approach to the elimination of electrophoresis is known as optical mapping or sequencing (Box 7.7). Optical mapping eliminates dependence on ensemble-based methods, focusing on the statistics of individual DNA molecules. Although this technique is fragile and, to date, not replicable in multiple laboratories,[28] it may eventually be capable of sequencing entire genomes much more rapidly than is possible today.

A different approach based on magnetic detection of DNA hybridization seeks to lower the cost of performing microarray analysis. Chen et al. have suggested that instead of tagging targets with fluorescent molecules, targets are tagged with microscopic magnetic beads.[29] Probes are implanted on a magnetically sensitive surface, such as a floppy disk, after removing the magnetic coating at the probe

[27] S. Brenner, "Gene Expression Analysis by Massively Parallel Signature Sequencing (MPSS) on Microbead Arrays," *Nature Biotechnology* 18(6):630-634, 2002. The elimination of electrophoresis (a common laboratory technique for separating biological samples by molecular weight) has many practical benefits. Conceptually, electrophoresis is a straightforward process. A tagged biological sample is inserted into a viscous gel and then subjected to an external electric field for some period of time. The sample differentiates in the electric field because the lighter components move farther under the influence of the electric field than the heavier ones. The tag on the biological sample is, for example, a compound that fluoresces when exposed to ultraviolet light. Measuring the intensity of the fluorescence provides an indication of the relative abundances of components of different molecular weight. However, in practice there are difficulties. The gel must undergo appropriate preparation—no small task. For example, the gel must be homogeneous, with no bubbles to interfere with the natural movement of the sample components. The temperature of the gel-sample combination may be important, because the viscosity of the gel may be temperature-sensitive. While the gel is drying (a process that takes a few hours), it must not be physically disturbed in a way that introduces defects into the gel preparation.

[28] Bud Mishra, New York University, personal communication, December 2003.

[29] C.H.W. Chen, V. Golovlev, and S. Allman, "Innovative DNA Microarray Hybridization Detection Technology," poster abstract presented at Human Genome Meeting 2002, April 14-17, 2004, Shanghai, China; also, "Detection of Polynucleotides on Surface of Magnetic Media, available at http://www.scien-tec.com/news1.htm.

location, and different probes are attached to different locations. Readout of hybridized probe-target pairs is accomplished through the detection of a magnetic signal at given locations; locations without such pairs provide no signal because the magnetic coating of the floppy disk has been removed from those locations. Also, the location of any given probe-target pair is treated simply as a physical address on the floppy disk. Preliminary data suggest that with the spatial resolution currently achieved, a single floppy diskette can carry up to 45,000 probes, a figure that compares favorably to that of most glass microarrays (of order 10,000 probes or less). Chen et al. argue that this approach has two advantages: greater sensitivity and significantly lower cost. The increased sensitivity is due to the fact that signal strength is controlled by the strength of the beads rather than the amount of hybridizing DNA per se; and so, in principle, this approach could detect even a single hybridization event. Lower costs arguably result from the fact that the most of the components for magnetic detection are mass-produced in quantity for the personal computer industry today.

Laboratory robotics is another area that offers promise of reduced labor costs. For example, the minimization of human intervention is illustrated by the introduction of compact, user-programmable robot arms in the early 1980s.[30] One version, patented by the Zymark Corporation, equipped a robot arm with interchangeable hands. This arm was the foundation of robotic laboratory workstations that could be programmed to carry out multistep sample manipulations, thus allowing them to be adapted for different assays and sample-handling approaches.

Building on the promise offered by such robot arms, a testbed laboratory formed in the 1980s by Dr. Masahide Sasaki at the Kochi Medical School in Japan demonstrated the feasibility of a high degree of laboratory automation: robots carried test tube racks, and conveyor belts transported patient samples to various analytical workstations. Automated pipettors drew serum from samples for the required assays. One-armed stationary robots performed pipetting and dispensing steps to accomplish preanalytical processing of higher complexity. The laboratory was able to perform all clinical laboratory testing for a 600-bed hospital with a staff of 19 employees. By comparison, hospitals in the United States of similar size required up to 10 times as many skilled clinical laboratory technologists.

Adoption of the "total laboratory automation" approach was mixed. Many clinical laboratories in particular found that it provided excess capacity whose costs could not be recovered easily. Midsized hospital laboratories had a hard time justifying the purchase of multimillion-dollar systems. By contrast, pharmaceutical firms invested heavily in robotic laboratory automation, and automated facilities to synthesize candidate drugs and to screen their biological effects provided three- to fivefold increases in the number of new compounds screened per unit time.

In recent years, manufacturers have marketed "modular" laboratory automation products, including modules for specimen centrifugation and aliquoting, specimen analysis, and postanalytical storage and retrieval. While such modules can be assembled like building blocks into a system that provides very high degrees of automation, they also enable a laboratory to select the module or modules that best address its needs.

Even mundane but human-intensive tasks are susceptible to some degree of automation. Consider that much of biological experimentation depends on the availability of mice as test subjects. Mice need to be housed and fed, and thus require considerable human labor. The Stowers Institute for Medical Research in Kansas City has approached this problem with the installation of an automated mouse care facility involving two robots, one of which dumps used mouse bedding and feeds it to a conveyor washing machine and the other of which fills the clean cages with bedding and places them on a rack.[31] These robots can process 350 cages per hour and reduce the labor needs of cleaning cages by a factor of three (from six technicians to two). At a cost of $860,000, the institute expects to recoup its investment in

[30] J. Boyd, "Robotic Laboratory Automation," *Science* 295(5554):517-518, 2002. Much of the discussion of laboratory automation is based on this article.

[31] C. Holden, ed., "High-tech Mousekeeping," *Science* 300(5618):421, 2003.

6 years, with much of the savings coming from reduced repetitive motion injuries and fewer health problems caused by allergen exposure.

In the future, modularization is likely to continue. In addition, fewer stand-alone robot arms are being used because the robotics necessary for sampling from conveyor belts are often integrated directly into clinical analyzers. Attention is turning from the development of hardware to the design of process control software to control and integrate the various automation components; to manage the transport, storage, and retrieval of specimens; and to support automatic repeat and follow-up testing strategies.

7.2.3 Future Challenges

From a conceptual standpoint, automation for speed depends on two things—speeding up an individual process and processing many samples in parallel. Individual processes can be speeded up to some extent, but because they are limited by physical time constants (e.g., the time needed to mix a solution uniformly, the time needed to dry, the time needed to incubate), the speedups possible are limited—perhaps factors of a few or even ten can be possible. By contrast, parallel processing is a much bigger winner, and it is easy to imagine processing hundreds or even thousands of samples simultaneously.

In addition to quantitative speedups, qualitatively new data acquisition techniques are needed as well. The difficulty of collecting meaningful data from biological systems has often constrained the level of complexity at which to collect data. Biologists often must use indirect or surrogate measures that imply activity. For example, oxygen consumption can be used as a surrogate for breathing.

There is a need to develop new mechanisms to collect data, particularly mechanisms that can form a bridge from the living system to a computer system, in other words, tools that detect and monitor biological events and directly collect and store information about those events for later analysis. Challenges in this area include the connection of cellular material, cells, tissues, and humans to computers for rapid diagnostics and data download, bio-aided computation, laboratory study, or human-computer interactivity, and how to perform "smart" experiments that use models of the biological systems to probe the biology dynamically so that measurements of the spatiotemporal dynamics of living cells at many scales become possible.

A good example of future data acquisition challenges is provided by single-cell assays and single-molecule detection. Traditional assays can involve thousands or tens of thousands of cells and produce datasets that reflect the aggregate behavior of the entire sample. While for many types of experiments this is an appropriate approach, there are current and future biological research issues for which this does not provide sufficient resolution. For example, cells within a population may be in different stages of their life cycle, may be experiencing local variations of environmental conditions, or may be of entirely different types. Alternatively, a probe might not touch the cell type of interest, due to inadequate purification of a sample drawn from a subject that contains many cell types.[32] For some biological questions, there is simply not a sufficient supply of cells of interest; for example, certain human nervous system tissue is highly specialized, and a biological inquiry may concern only a few cells. Similarly, in attempts to isolate some diseases, there may be only a few, or even only one, affected cell—for example, in attempts to detect cancerous cells before they develop into a tumor.

Many technologies offer approaches to analyzing and characterizing the behavior of single cells, including the use of mass spectrometry, microdissection, laser-induced fluorescence, and electrophoresis. Ideally, it would be possible to monitor the behavior of a living cell over time with sufficient resolution to determine the functioning of subcellular components at different stages of the life cycle and in response to differing environmental stimuli.

[32] Today, this issue is addressed by the very labor-intensive process of "plucking" individual cells from a sample and aggregating them—a process that typically requires 10^4 to 10^5 cells when today's assays are used.

A further challenge in ultrasensitive data acquisition in living cells is that the substances of interest, particularly proteins, occur at a wide range of concentrations (varying by many orders of magnitude). For many important proteins, this may be as few as hundreds of individual molecules. Detection and analysis at such low levels must work even in the face of wide statistical fluctuation, transient modifications, and a wide range of physical and chemical properties.[33]

At the finest grain, detection and analysis of single molecules could provide further understanding of cellular mechanisms. Again, although there are current techniques to analyze molecular structure (such as nuclear magnetic resonance and X-ray crystallography), these work on large, static samples. To achieve more precise understanding of cellular mechanisms, it is necessary to detect the presence and activity of very small concentrations, even single molecules, dynamically within living cells. Making progress in this field will require advances in chemistry, instrumentation, sensors, and image analysis algorithms.[34]

Embedded networked sensor (ENS) systems will ride the cost reduction curve that characterizes much of modern electronic systems. Based on microsensors, on-board processing, and wireless communications, ENS systems can monitor phenomena "up close." Nevertheless, taken as a whole, ENS systems present challenges with respect to longevity, autonomy, scalability, performance, and resilience. For example, off-the-shelf sensors embedded in heterogeneous soil for monitoring soil moisture and nitrate levels raise issues related to calibration when embedded in a previously unknown environment. In addition, the uncertainty in the data they provide must be characterized. Interesting theoretical issues arise with respect to the statistical and information-theoretic foundations for adaptive sampling and data fusion. Also, of course, programming abstractions, common services, and tools for programming the network must be developed.

To illustrate a specific application, consider some of the computing challenges in deploying ENS systems for marine microorganisms. The ultimate goal is to deploy large groups of autonomous, mobile microrobots capable of identifying and tracking microorganisms in real time in the marine environment, while measuring the relevant environmental conditions at the required temporal and spatial scales. Sensors must be mobile to track microorganisms and assess their abundance with a reasonable number of sensors. They must be small, so that they are able to gather information at a spatial scale comparable to the size of the microorganisms and to avoid disturbing them. They must operate in a liquid environment—combined with small sensor size, operation in such an environment raises many difficult issues of mobility, communications, and power, which in turn strongly impact network algorithms and strategies. Also, sensors must be capable of in situ, real-time identification of microorganisms, which requires the development of new sensors with considerable on-board processing capability. Progress in this application—monitoring marine environments and single-cell identification—is expected to be applicable to other liquid environments, such as the circulatory system of higher organisms, including humans.

[33] R.D. Smith et al., "Application of New Technologies for Comprehensive, Quantitative and High Throughput Microbial Proteomics," abstracts of the Department of Energy's (DOE) Genomes to Life Systems-Biology Projects on Microbes Sequenced by the U.S. DOE's Microbial Genome Program, available at http://doegenomestolife.org/pubs/2004abstracts/html/Tech_Dev.shtml#_VPID_289.

[34] See, for example, the text of the NIH Program Announcement PA-01-049, "Single Molecule Detection and Manipulation," released February 12, 2001, available at http://grants.nih.gov/grants/guide/pa-files/PA-01-049.html.

8

Biological Inspiration for Computing

Chapters 4-7 address ways in which computer science and engineering can assist in the pursuit of a broadly defined research agenda in biology. This chapter suggests how insights from the biological sciences may have a positive impact on certain research areas in computing, although the impact of this reversed direction is at present much more speculative.[1]

8.1 THE IMPACT OF BIOLOGY ON COMPUTING

8.1.1 Biology and Computing: Promise and Skepticism

Today's computer systems are highly complex and often fragile. Although they provide high degrees of functionality to their users, many of today's systems are also subject to catastrophic failure, difficult to maintain, and full of vulnerabilities to outside attack. An important goal of computing is to be able to build systems that can function with high degrees of autonomy, robustly handle data with large amounts of noise, configure themselves automatically into networks (and reconfigure themselves when parts are damaged or destroyed), rapidly process large amounts of data in a massively parallel fashion, learn from their environment with minimal human intervention, and "evolve" to become better adapted to what they are supposed to do.

There is little doubt that such computer systems with these properties would be highly desirable. Although the development of such systems is an active area of computer science research today (indeed, the Internet itself is an example of a system that is capable of operating without centralized authority and reconfiguring itself when parts are damaged), computer science researchers are working to develop new such systems, and the prospect of looking outside the existing computer science toolbox for new types of hardware, software, algorithms, or something entirely different (and unknown) is increasingly attractive.

One possible area of research focuses on a set of techniques inspired by the biological sciences, because biological organisms often exhibit properties that would be desirable in computer systems.

[1] A popularized account of biological inspiration for computing is N. Forbes, *Imitation of Life: How Biology Is Inspiring Computing,* MIT Press, Cambridge, MA, 2004.

They function with high degrees of autonomy. Some biological entities—such as neurons in a brain—can configure themselves automatically into networks (and reconfigure themselves to some degree when parts are damaged or destroyed). Sensory systems rapidly pick out salient features buried in large amounts of data. Many animals learn from their environment and become better adapted to what they are supposed to do. All biological organisms have mechanisms for self-repair, and all multicellular organisms grow from an initial state that is much less phenotypically complex than their final states.

Carver Mead once noted that "engineers would be foolish to ignore the lessons of a billion years of evolution." The solutions that nature has evolved to difficult engineering problems are, in many cases, far beyond present-day engineering capability. For example, the human brain is not fast enough to process all of the raw sensory data detected by the optic or auditory nerves into meaningful information. To reduce processing load, the brain uses a strategy we know as "attention" that focuses on certain parts of the available information and discards other parts. Such a strategy might well be useful for an artificial machine processing a large visual field. Studies of the way in which humans limit their attention has led to computational models of the strategy of shifting attention. Such models of biological systems are worth studying even if they appear intuitively less capable than computation, if only for the fact that no machine systems exist that can function as autonomously as a housefly or an ant.

On the other hand, biological organisms operate within a set of constraints that may limit their suitability as sources of inspiration for computing. Perhaps the most important constraint is the fact that biological organisms emerge from natural selection and the evolutionary process. Because selection pressures are multidimensional, biological systems must be multifunctional. For example, a biological system may be able to move, but it has also evolved to be able to feed itself, to reproduce, and to defend itself. The list of desirable functions in a biological system is long, and successfully mimicking biology for one particular function requires the ability to separate nonrelevant parts of the system that do not contribute to the desired function. Furthermore, because biological systems are multifunctional, they cannot be optimized for any one function. That is, their design always represents a compromise between competing goals. Organisms must be adequately (rather than optimally) adapted to their environments. (The notion of "optimal design" is also somewhat problematic in the context of stochastic real-world environments.) Also, optimal adaptation to any one environment is likely to disadvantage an organism in a significantly different environment, and so adequately adapted organisms tend to be more robust across a range of environments.

The evolutionary process constrains biological solutions as well. For example, biological systems inevitably include vestiges of genetic products and organs that are irrelevant to the organism in its current existence. Thus, biological adaptation to a given environment depends not only on the circumstances of the environment but also on its entire evolutionary history—a fact that may well obscure the fundamental mechanisms and principles in play that are relevant to the specific environment of interest. (This point is a specific instantiation of a more general phenomenon, which is that our understanding of biological phenomena will often be inadequate to provide detailed guidance in engineering a computational device or artifact.)

A corollary notion is that nature may evolve different biological mechanisms to solve a given problem. All of these mechanisms may enable the organism to survive and even to prosper in its environment, but it is far from clear how well these mechanisms work relative to one another.[2] Thus, which one of many biological instantiations is the most appropriate model to mimic remains an important question.

Finally, there are only a few examples of successful biologically inspired computing innovations. Thus, the jury is still out on the ultimate value of biology for computing. Rather than biology being helpful across the board to all of computing, the committee believes that biology's primary relevance (at least in the short term) is likely to be to specific problem areas within computing that are poorly

[2] For example, fish and squid use different mechanisms to propel themselves through the water. Which mechanism is better under what circumstances and for what engineered artifacts is a question for research to answer.

understood, or for which the relevant underlying technologies are too complex or unwieldy, and in providing approaches that will address parts of a solution (as described in Section 8.1.2). Nevertheless, the potential benefits that biology might offer to certain problem areas in computing are large, and it is worth exploring different approaches to exploit these benefits; this is the focus of Sections 8.2 to 8.4.

8.1.2 The Meaning of Biological Inspiration

What does it mean for something to be biologically inspired? It is helpful to consider several possible interpretations. One interpretation is that significant progress in computing can occur *only* through the application of principles derived from the study of biology. This interpretation, offered largely as a strawman, is absurd—there are many ways in which computing can progress without the application of biologically derived principles.

A second, somewhat less grandiose and more reasonable interpretation is that significant progress in computing *can* occur through the application of principles derived from the study of biology. That is, a biological system may operate according to principles that have applicability to nonbiological computing problems. By studying the biological system, one may be able to derive or understand the relevant principles and use them to help solve a nonbiological problem. It is this interpretation—that biology is relevant to computing only when principles emerge directly from a study of biological phenomena—that underlies many claims of biological relevance or irrelevance to computing.

A third interpretation is that certain aspects of biology are analogous to aspects of computing, which means that insights from biology are relevant to aspects of computing. This is the case, for instance, when a set of principles or paradigms turns out to have strong applicability both to a biological system or systems and to interesting problems in computing. These principles or paradigms may have had their intellectual origin in the study of a biological or a nonbiological system.

When their origin is in a biological system, this interpretation reduces to the second interpretation above. What makes the case of an origin in a nonbiological system interesting is that the principles in question may be more manifestly obvious in a biological context than in a nonbiological context. That is, the principles and their application may most easily be seen and appreciated in a biological context, even if they did not initially originate in a biological context. Moreover, the biological context may also provide a source of language, concepts, and metaphors that are useful in talking about a nonbiological problem or phenomenon.

For this report, the term "inspiration" will be used in its broadest sense, that is, the third interpretation above, but there are three other points to keep in mind:

- Biological inspiration does not mean that the weaknesses of biology must be adopted along with the strengths. In some cases, it may be possible to overcome problems found in the actual biological system when the principles underlying them are implemented in engineered artifacts.
- As noted in Chapter 1, even when biology cannot provide insight into potential computing solutions, the drive to solve biological problems can still inspire interesting, relevant, and intellectually challenging research in computing—so biology can serve as a useful and challenging problem domain for computing.[3]

[3] For example, IBM used the problem of protein folding to motivate the development of the BlueGene/L supercomputer. Specifically, the problem was formulated in terms of obtaining a microscopic view of the thermodynamics and kinetics of the dynamic protein-folding process over longer time scales than have previously been possible. Because this project involved both computer architecture and the exploration of algorithmic alternatives, the applications architecture was structured in such a way that subject experts in molecular simulation could work on their applications without having to deal with the complexity of the parallel communications environment required by the underlying machine architecture (see BlueGene/L Team, "An Overview

- Incomplete (and sometimes even incorrect) biological understandings help to inspire different and useful approaches to computing problems. Important and valuable insights into possible ways to solve a current problem have been derived from biological models that were incomplete (as in the case of evolutionary programming) or even inaccurate (as in the case of immunologically based computer security).

On the other hand, it must be understood that the use of a biological metaphor to inspire new approaches to computing does not necessarily imply that the biological side is well understood, whether or not the metaphor leads to progress in computing. That is, even if a biological metaphor is applicable and relevant to a computing problem, this does not mean that the corresponding biological phenomena can necessarily be understood in computational terms.

For example, although researchers use the term "genetic algorithms" to describe a class of algorithms using operators that have a similar flavor to evolutionary genetic operators such as mutation or recombination to search a solution space stochastically, the definition and implementation of these genetic operators does not imply a fundamental understanding of biological evolutionary processes. Similarly, although the field of "artificial neural networks" is an information-processing paradigm inspired by the parallel processing capabilities and structure of nerve tissue, and it attempts to mimic learning in biology by learning to adjust "synaptic" connections between artificial processing elements, the extent to which an artificial neural network reflects real neural systems may be tenuous.

8.1.3 Multiple Roles: Biology for Computing Insight

Biological inspiration can play many different roles in computing, and confusion about this multiplicity of meanings accounts for a wide spectrum of belief about the value of biology for developing better computer systems and improved performance of computational tasks. One point of view is that only a detailed "ground-up" understanding of a biological system can result in such advances, and because such understanding is available for only a very small number of biological systems (and "very small" is arguably zero), the potential relevance of biology for computing is small, at least in the near term.

A more expansive view of biology's value for computing acknowledges that detailed understanding is the key for a maximal application of biology to computing, but also holds that biological metaphors, analogies, examples, and phenomenological insights may suggest new and interesting ways of thinking about computational problems that might not have been imagined without the involvement of biology.[4] From this perspective, what matters is performance of a task rather than simulation of what a biological system actually does, though one would not necessarily expect initial performance models

of the BlueGene/L Supercomputer," presented at Supercomputing Conference, November 2002, available at http://sc-2002.org/paperpdfs/pap.pap207.pdf). Other obvious problems inspired by biology include computer vision and artificial intelligence. It is also interesting to note this historical precedent of biological problems being the domain in which major suites of statistical tools were developed. For instance, Galton invented regression analysis (correlation tests) to study the relation of phenotypes between parents and progeny (see F. Galton, *Natural Inheritance*, 5th Edition, Macmillan and Company, New York, 1894). Pearson invented the chi-square and other discrete tests to study the distribution of different morphs in natural populations (see K. Pearson, "Mathematical Contributions to the Theory of Evolution, VIII. On the Inheritance of Characters Not Capable of Exact Quantitative Measurement," *Philosophical Transactions of the Royal Society of London*, *Series A* 195:79-150, 1900). R.A. Fisher invented analysis of variance to study the partitioning of different effects in inheritance (see R. Fisher, "The Correlation Between Relatives on the Supposition of Mendelian Inheritance," *Transactions of the Royal Society of Edinburgh* 52:399-433, 1918).

[4] An analogy might be drawn to the history of superconducting materials. A mix of quantum principles, phenomenology, and trained experience has led to superconducting materials with ever-higher transition temperatures. (Indeed, the discovery of superconducting materials preceded quantum mechanics by more than a decade.)

based on biological systems to function more effectively than models constructed using more traditional techniques.

One of biology's most important roles is that it can serve as an existence proof of performance—that some desirable behavior is possible. The reasoning is that if a biological system can do something interesting, why can't an artificial system to the same thing? Birds fly, so why shouldn't people or constructed artifacts be able to fly? Many biological behaviors and functions would be desirable in a computing context, and biological systems that exhibit such behavior demonstrate that this behavior is possible.[5]

Existence proofs are important in engineering. For example, in the view of many nuclear scientists associated with the Manhattan Project, the information that was most critical to the Soviet development effort was not a secret gained through espionage, but rather the fact that a nuclear explosion was possible at all—and that fact was reported in every major newspaper in the world.[6] In other words, it is one thing to work toward a goal that may well be impossible to achieve and an entirely different psychological matter to work toward a goal whose achievement is known—with certainty—to be possible.

An example of using a biological metaphor for understanding some dimension of computing relates to computer security. From many centuries of observation, it is well known that an ecology based on a monoculture is highly vulnerable to threats that are introduced from the outside. With this insight in mind, many expert observers have used the term "monoculture" to describe the present-day security environment for desktop computers in which one vendor dominates the operating system market. This report does not take a position on whether such a characterization is necessarily accurate,[7] but the point is that the metaphor, used in this manner, can determine the terms of discussion and thus provide a useful way of looking at the issue.

Despite its conceptual value, an existence proof does not speak directly to how to build the artifact so that it does the same thing. That is, existence proofs do not necessarily provide insight about construction or creation. Diversity as a strategy for survival does not necessarily indicate how much or what kinds of diversity would be helpful in any given instance. Similarly, aerodynamics is a body of theory that explains the flight of birds, and also enables human beings to design airplanes, but a study of birds did not lead to the airplane. For construction or creations, a deeper understanding of biology is required. Knowing what kind of deeper understanding is possible potentially leads to at least three additional roles for biology:

- *Biology as source of principles*. Nature builds systems out of the same atoms that are available to human engineers. If a biological system can demonstrate a particular functionality, it is because that system is built according to principles that enable such functionality. The hope is that upon close examination, the physical, mathematical, and information-processing principles underlying the interesting biological functionality can be applied through human engineering to realize a better artificial system. Note also that in some cases, the actual principles underlying some biological functionality may be difficult to discern. However, plausibility counts for a great deal here, and biology may well provide inspiration for engineered artifacts if human beings propose a set of plausible principles that govern the behavior of interest in an actual organism, even if those principles, as articulated, turn out not to have a biological instantiation in that organism. (Note that in this domain the division between "applying

[5] An accessible and more extended discussion of these ideas can be found in J. Benyus, *Biomimicry: Innovation Inspired by Nature*, William Morrow, New York, 1997.

[6] D. Holloway, *Stalin and the Bomb: The Soviet Union and Atomic Energy, 1939-1956*, Yale University Press, New Haven, 1994.

[7] For example, it may be that even though the number of operating system platforms is small compared to the number of desktop computers in use, different computer configurations and different operational practices might introduce sufficient diversity to mitigate any system-wide instabilities. Furthermore, replication has many other advantages in the computer context, such as easier interoperability.

biological principles to information processing" and "understanding biological information processing" is least meaningful.)

- *Biology as implementer of mechanism*. Nature also implements mechanisms to effect certain functions. For example, a biological organism may implement an algorithm that could be the basis of a solution to a computing problem of interest to people. Or, it may implement an architecture or a way to organize and design the structural and dynamic relationships between elements in a complex system, knowledge of which might greatly improve the design of an engineered artifact. In this category are the neural network architecture as inspired by the activation model of dendrites and axons in the brain, evolutionary computation as driven by genomic changes and selection pressures, and the use of electroactive polymers as actuator mechanisms for robots**,** inspired by the operation of animal muscles (rather than, for example, gears). (Note that implementations of biological mechanisms tend to be easier to identify and extract for later use when they involve physical observables—and so mechanisms underlying sensors and locomotion have had some nontrivial successes in their application to engineered artifacts.)
- *Biology as physical substrate for computing.* Computation can be regarded as an abstract or a physically instantiated form. In the abstract, it is divorced from anything tangible. But all real-world computation requires hardware—a device of some kind, whether artificial or biological—and given that biological organisms are functional physical devices, it makes sense to consider how engineered artifacts might have biological components. For example, biology may provide parts that can be integrated into engineered devices. Thus, a sensitive chemical detection system might use a silk moth as the sensor for chemicals in the air and thus instrument the moth to appropriate readouts. Or a small animal might be used as the locomotive platform for carrying a useful payload (e.g., a camera), and its movements might be teleoperated through electrodes implanted in the animal by a human being viewing the images sent back by a camera.

These three different roles are closely connected to the level(s) of abstraction appropriate for thinking about biological systems. For some systems and phenomena of interest, a very "bottom-up" perspective is warranted. In the same way that one needs to know how to use transistors to build a logic gate for a silicon-based computer, one needs to know how neurons in the brain encode information in order to understand how a neural implant or prosthetic device might be constructed. For other systems and phenomena, architecture provides the appropriate level of abstraction. In this case, understanding how parts of a system are interconnected, the nature of the information that is passed between them, and the responses of those parts to such information flows may be sufficient.

Another way of viewing these three roles is to focus on the differences between computational content, computational representation, and computational hardware. Consider, for example, a catenary curve—the shape that a cable suspended at both ends takes when subjected to gravity.

- The computational content is specified by a differential equation and the appropriate boundary conditions. Although the solution is not directly apparent from the differential equation, the differential equation implies a specific curve that represents the answer.
- The computational representation refers to how the computation is actually represented—in digital form (as bits in a computer), in analog form (as voltages in an analog computer), in neural form (as how a calculus student would solve the problem), or in physical form (as the string or cable being represented).
- The computational hardware refers to the physical device used to solve the equation—the digital computer, the analog computer, the human being, or the cable itself.

These three categories correspond roughly and loosely to the three categories described above: content as source of principles, representation as implementer of mechanism, and hardware as physical substrate. The remaining sections of this chapter describe some biological inspirations for work in computing.

8.2 EXAMPLES OF BIOLOGY AS A SOURCE OF PRINCIPLES FOR COMPUTING

8.2.1 Swarm Intelligence and Particle Swarm Optimization

Swarm intelligence is a property of systems of nonintelligent, independently acting robots that exhibit collectively intelligent behavior in an environment that the robots do sense and can alter.[8] One form of swarm intelligence is particle swarm optimization, based on the flocking of birds.[9]

The canonical example of flocking behavior is a flight of birds wheeling through the sky, or a school of fish darting through a coral reef. Somehow, myriad not-very-bright individuals manage to move, turn, and respond to their surroundings as if they were as a single, fluid organism. Moreover, they seem to do so collectively, without a leader: biologists armed with high-speed video cameras have shown that the natural assumption—that each flock or school has a single, dominant individual that always initiates each turn just a fraction of a second before the others follow—is simply not true.

The first known explanation of the leaderless, collective quality of flocking or schooling behavior emerged in 1986. This explanation used swarms of simulated creatures—"boids"—that could form surprisingly realistic flocks if each one simply sought to maintain an optimum distance from its neighbors. The steering rules of the so-called Reynolds simulation were simple:[10]

- *Separation:* steer to avoid crowding local flock mates.
- *Alignment:* steer toward the average heading of local flock mates.
- *Cohesion:* steer toward the average position of local flock mates.

These rules were entirely local, referring only to what an individual boid could see and do in its immediate vicinity;[11] none of them said, "Form a flock." Yet the flocks formed every time, regardless of the starting positions of the boids. These flocks were able to fly around obstacles in a very fluid and natural manner. Sometimes the flock would even break into subflocks that flowed around both sides of an obstacle, rejoining on the other side as if the boids had planned it all along. In one run, a boid accidentally hit a pole, fluttered around for a moment, and then darted forward to rejoin the flock as it moved on.

Today, the Reynolds simulation is regarded as one of the best and most evocative demonstrations of *emergent behavior*, in which complex global behavior arises from the interaction of simple local rules. The approach embodied in the simple-rule/complex-behavior approach has become a widely used technique in computer animation—which was Reynolds' primary interest in the first place.[12]

[8] T. White, "Swarm Intelligence: A Gentle Introduction with Applications," PowerPoint presentation, available at http://www.sce.carleton.ca/netmanage/tony/swarm-presentation/tsld001.htm.

[9] Bird flocks are an example of complex, adaptive systems. Among the many other examples that scientists have studied are the world economy, brains, rain forests, traffic jams, corporations, and the prehistoric Anasazi civilization of the Four Corners area. Complex adaptive systems are similar in structure and behavior even if they differ in their superficial manifestations. For example, complex adaptive systems are massively parallel and involve many quasi-independent "agents" interacting at once. (An agent might be a single firm in an economy, a single driver on a crowded freeway, and so on.) Each of them is adaptive, meaning that the agents that constitute them are constantly responding and adapting to each other. And each of them is decentralized, meaning that no one agent is in charge. Instead, a complex system's overall behavior tends to emerge spontaneously from myriad low-level interactions.

[10] C.W. Reynolds, "Flocks, Herds, and Schools: A Distributed Behavioral Model," *Computer Graphics* 21(4):25-34, 1987, available at http://www.cs.toronto.edu/~dt/siggraph97-course/cwr87/ and http://www.red3d.com/cwr/papers/1987/SIGGRAPH87.pdf. An updated discussion, with many pictures and references to modern applications, can be found in C.W. Reynolds, "Boids: Background and Update," 2001, available at http://www.red3d.com/cwr/boids/.

[11] More precisely, each boid had global information about the physical layout of its environment, including any obstacles, but it had no information about its flock mates, except for those that happened to come within a certain distance that defined its local neighborhood.

[12] The first Hollywood film to use a version of Reynolds' boids software was Tim Burton's *Batman Returns* (1992), which featured swarms of animated bats and flocks of animated penguins. Since then it has been used in films such as *The Lion King* (1994) and many others (see http://www.red3d.com/cwr/boids/).

A second simulation of flocking behavior, developed in 1990, employed the Reynolds' rules (though they were independently developed) and also incorporated the influence of "dynamic forces" on the behavior of the simulated creatures.[13] These dynamic forces would allow the creatures to be attracted toward a convenient roosting point, say, or a particularly rich cornfield. As a result, the flock would turn and head in the direction of a cornfield as soon as it was placed into view, with various subgroups swinging out and in again until finally the whole group had landed right on target.

These two models are direct ancestors of the particle swarm optimization (PSO) algorithm, first published in 1995.[14] The algorithm substitutes a mathematical function for the original roosts and cornfields, and employs a conceptual swarm of bird-like particles that swoop down on the function's maximum value, even when the function has many local maxima that might confound more standard optimization algorithms.

The essential innovation of the PSO algorithm is to scatter particles at random locations throughout a multidimensional phase space that represents all the arguments to the function to be maximized. Then the algorithm sets the particles in motion. Each particle evaluates the function as it flies through phase space and keeps trying to turn back toward the best value that it has found so far. However, it is attracted even more toward the best value that any of its neighboring particles have found. So it inexorably begins to move in that direction—albeit with a little built-in randomness that allows it to explore other values of the function along the way. The upshot is that the particles quickly form a flock that flows toward a point that is one of the highest function values available, if not *the* highest.

The PSO algorithm is appealing for both its simplicity—the key steps can be written in just a few lines of computer code—and its effectiveness. In the original publication of the PSO algorithm, the algorithm was applied to a variety of neural network problems, and it was found to be a very efficient way to choose the optimum set of connection weights for the network.[15] Since then, the basic technique has been refined and extended to systems that have discrete variables, say, or that change with time. It also has been applied to a wide variety of engineering problems,[16] such as the automatic adjustment of power systems.[17]

The PSO algorithm is biologically inspired in the sense that it is a plausible account of bird flocking behavior. However, it is not known whether birds, in fact, use the PSO algorithm to fly in formation.

Swarm algorithms have the virtues of simplicity and robustness, not to mention an ability to function without the need for centralized control. For this reason, they may find their most important applications in, say, self-healing and self-organizing communications networks or in electrical power networks that could protect themselves from line faults and reroute current around a broken link "on the fly."[18]

On the other hand, simple rules are not automatically good. Witness army ants, which are such obsessive self-organizers that the members of an isolated group will often form a "circular mill," follow-

[13] F.H. Heppner and U. Grenander, "A Stochastic Nonlinear Model for Coordinated Bird Flocks," *The Ubiquity of Chaos*, S. Krasner, ed., AAAS Publications, Washington, DC, 1990.

[14] J. Kennedy and R.C. Eberhart, "Particle Swarm Optimization," pp. 1942-1948 in *Proceedings of the IEEE International Conference on Neural Networks*, IEEE Service Center, Piscataway, NJ, 1995; R. Eberhart, Y. Shi, and J. Kennedy, *Swarm Intelligence*, Morgan Kaufman, San Francisco, CA, 2001.

[15] See Section 8.3.3.2 for further discussion.

[16] A good sense of current activity in the field can be gleaned from the programs and talks at the 2003 IEEE Swarm Intelligence Symposium, April 24-26, 2003, available at http://www.computelligence.org/sis/index.html. Extensive references to PSO can be found at "Welcome to Particle Swarm Central," 2003, available at http://www.particleswarm.info. This site also contains a number of links to online tutorials and downloadable PSO code.

[17] K.Y. Lee and M.A. El-Sharkawi, eds., *Modern Heuristic Optimization Techniques with Applications to Power Systems*, John Wiley and IEEE Press, New York, March 2003.

[18] E. Bonabeau, "Swarm Intelligence," presented at the O'Reilly Emerging Technology Conference, April 22-25, 2005, Santa Clara, CA. Powerpoint presentation available at http://conferences.oreillynet.com/presentations/et2003/Bonabeau_eric.ppt.

ing one another around and around and around until they die from starvation.[19] Such blind-leading-the-blind behaviors are an ever-present possibility in swarm intelligence; the trick is to find simple rules that minimize the chances of that happening.

A closely related challenge is to find ways of designing emergent behavior, so that the swarm will produce predictable and desirable results. Today, swarm algorithms are based on the loose and imprecise specification of a relatively small number of parameters—but it is almost certainly true that engineered artifacts that exhibit complex designed behavior will require the tight specification of many parameters.

This point is perhaps most obvious in the cooperative construction problem, where the rule sets that produce interesting, complex structures are actually very rare; most self-organized structures look more like random blobs.[20] The same problem is common to all collective behaviors; finding the right rules is still largely a matter of trial and error—not least because it is in the very nature of emergence for a simple-seeming change in the rules to produce a huge change in the outcome. Thus, in their efforts to find the right rules, researchers may well seek to develop procedures that will find in the right rules rather than trying to find them directly themselves. This point is discussed further in Section 8.3.1.

8.2.2 Robotics 1: The Subsumption Architecture

One approach to robotic design is based on the notion that complex and highly capable systems are inherently expensive, and hence fewer can be built. Instead, this approach asserts the superiority of using large numbers of individually smaller, less capable, and inexpensive systems.[21] In 1989, Brooks and Flynn suggested that "gnat robots" might be fabricated using silicon micromachining to fabricate freely movable structures onto silicon wafers. Such an approach potentially allows sensors, actuators, and electronics to be embedded on the same silicon substrate. This arrangement is the basis for Brooks' subsumption architecture, in which low-level functionality can be used as building blocks for higher-level functionality.

Robots fabricated in this manner could be produced by the thousands, just as integrated circuits are produced today—and thus become an inexpensive, disposable system that does its work and need not be retrieved. For applications such as exploration in hostile environments, the elimination of a retrieval requirement is a significant cost savings.

To the best of the committee's knowledge, no self-propelled robots or other operational systems have been built using this approach. Indeed, experience suggests that the actual result of applying the swarm principle is that one highly capable robot is not replaced by many robots of lesser capability, but rather *one* such robot. This suggests that real-world applications are likely to depend on the ability to fabricate many small robots inexpensively.

A key challenge is thus to develop ways of assembling microrobots that are analogous to chip fabrication production lines. One step toward meeting this challenge has been instantiated in a concept known as "smart dust," for which actual prototypes have been developed. Smart dust is a concept for a

[19] B. Hölldobler and E.O. Wilson, *The Ants*, Belknap Press of Harvard University Press, Cambridge, MA, 1990, pp. 585-586. In a famous account published in 1921, the entomologist William Beebe described a mill he saw in the Amazonian rain forest that measured some 360 meters across, with each ant taking about $2^1/_2$ hours to complete a circuit. They kept at it for at least 2 days, stumbling along through an ever-accumulating litter of dead bodies, until a few workers finally straggled far enough from the trail to break the cycle. And from there, recalled Beebe, the group resolutely marched off into the forest. See W. Beebe, *Edge of the Forest*, Henry Holt and Company, New York, 1921.

[20] But then, so do most insect nests. Honeycombs, wasps' nests, and other famous examples are the exception rather than the rule.

[21] R.A. Brooks and A.M. Flynn, "Fast, Cheap and Out of Control: A Robot Invasion of the Solar System," *Journal of the British Interplanetary Society* 42:478-485, 1989.

highly distributed sensor system.[22] Each dust mote has sensors, processors, and wireless communications capabilities and is light enough to be carried by air currents. Sensors could monitor the immediate environment for light, sound, temperature, magnetic or electric fields, acceleration, pressure, humidity, selected chemicals, and other kinds of information, and the motes, when interrogated, would send the data over kilometer-scale ranges to a central base station, as well as communicate with local neighbors.

This architecture was the basis of an experiment that sought to track vehicles with an unmanned aerial vehicle (UAV)-delivered sensor network.[23] The prototype sensors were approximately a cubic inch in volume and contained magnetic sensors for detecting vehicles (at ranges of about 10 meters), a microprocessor, radio-frequency communications, and a battery or solar cell for power. With six to eight air-delivered sensor motes landed diagonally across a road at about 5-meter intervals, the sensor network was able to detect and track vehicles passing through the network, store the information, and then transfer vehicle track information from the ground network to the interrogating UAV and then to the base camp.

The subsumption architecture also asserts that this robust behavior can emerge from the bottom up.[24] For example, in considering the problem of an autonomously functioning vehicle (i.e., one that drives itself), a series of layers can be defined that

- Avoid contact with objects (whether the objects move or are stationary),
- Wander aimlessly around without hitting things, and
- Explore the world by seeing places in the distance that look reachable and heading for them.

Any given level contains as a subset (subsumes) the lower levels of competence, and each level can be built as a completely separate component and added to existing layers to achieve higher levels of competence. In particular, a level 0 machine would be built that simply avoided contact with objects. A level 1 machine could be built by adding another control layer that monitors data paths in the level 0 layer and inserts data onto the level 0 data paths, thereby subsuming the normal data flow of level 0. More complex behavior is thus built on top of simpler behaviors.

Brooks claims that the subsumption architecture is capable of accounting for the behavior of insects, such as a house fly, using a combination of simple machines with no central control, no shared representation, slow switching rates, and low-bandwidth communication. This results in robust and reliable behavior despite its limited sensing capability and an unpredictable environment, because individual behaviors can compensate for each others' failures, resulting in coherent and emergent behavior despite the limitations of the component behaviors. A number of robots have been built using subsumption architectures. Of particular note is Hannibal,[25] a hexapod with more than 100 physical sensors and 1,500 augmented finite-state machines grouped into several dozen behaviors split over eight on-board computers.[26]

8.2.3 Robotics 2: Bacterium-inspired Chemotaxis in Robots[27]

The problem of locating gradient sources and tracking them over time is an important problem in many real-world contexts. For example, fires cause temperature gradients in their immediate vicinity;

[22] See, for example, http://robotics.eecs.berkeley.edu/~pister/SmartDust/.

[23] See http://robotics.eecs.berkeley.edu/~pister/29Palms0103/.

[24] R.A. Brooks and A.M. Flynn, "Fast, Cheap and Out of Control," 1989.

[25] C. Ferrell, "Robust Agent Control of an Autonomous Robot with Many Sensors and Actuators," Ph.D. thesis, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 1993.

[26] A finite-state machine is a machine with a finite number of internal states that transitions from one state to another on the basis of a specified function. That is, the argument of the function is the machine's previous state, and the function's output is its new state. An augmented finite-state machine is a finite-state machine augmented with a timer that forces a transition after a certain time.

[27] Material in Section 8.2.3 is based on excerpts from A. Dhariwal, G.S. Sukhatme, and A.A.G. Requicha, "Bacterium-inspired Robots for Environmental Monitoring," *International Conference on Robotics and Automation*, New Orleans, LA, April 2004.

chemical spills lead to chemical concentration gradients in the soil and/or water; ecosystems host gradients of light, salinity, and pH. In many cases, the source intensity of these gradients varies with time (e.g., because of movement of the source), and there may be multiple sources for any given characteristic (e.g., two fires causing a complex temperature gradient).

Autonomous detection, location, and tracking of gradient sources would be very helpful for those trying to study or respond to the environment. Using robots, an environmental scientist might need to find the source(s) of a given toxic chemical, whereas a firefighter might need to locate the source(s) of a fire in order to extinguish it.

Noting that other approaches for locating and tracking gradient sources were primarily useful in static or quasi-static environments, and inspired by biological studies of how bacteria are attracted to gradient sources of nutrition, Dhariwal et al.[28] sought to develop a strategy for finding gradient sources that worked well with sources that are small, weak, mobile, or time-varying in intensity. Specifically, their algorithm is based on the repetition of a straight-line run for a certain time, followed by a random change in direction that sets up the direction for a new run. If the bacterium senses a higher concentration in its immediate environment, the run length is longer. Thus, although the bacterium still undergoes a random walk, it is a random walk biased in the direction of the gradient source.

This algorithm is also well suited for implementation in a simple robot. That is, only the last sensor reading must be stored, and so memory requirements are lower. Because only one computation has to be done (a comparison between the present and the previous sensor reading), processing requirements are minimal.

Dhariwal et al. compared the performance of this algorithm with a simple gradient descent algorithm. They found that for single, weak sources, the simple gradient algorithm displayed better performance. However, the bacterium-inspired algorithm displayed better performance in locating and tracking multiple and/or dissipative sources and in covering the entire area in which the gradient can be found.

8.2.4 Self-healing Systems

In the past few years, the term "self-healing" has become a fashionable object of study and interest in the academic and research computer science communities[29] and in the marketing materials of information technology (IT) companies such as IBM,[30] Microsoft,[31] Sun,[32] and HP.[33] Despite (or because of?) this level of interest, there is no commonly accepted definition of "self-healing" or agreement of what functionality it encompasses or requires.

[28] A. Dhariwal, G.S. Sukhatme, and A.A.G. Requicha, "Bacterium-inspired Robots for Environmental Monitoring," *IEEE International Conference on Robotics and Automation*, New Orleans, LA, April 25-30, 2004, available at http://www-lmr.usc.edu/~lmr/publications/Icra04bact.pdf.

[29] Workshop on Self-healing, Adaptive and Self-managed Systems (SHAMAN), June 23, 2002, available at http://www.cse.psu.edu/~yyzhang/shaman/proc.html; ICSE 2003 Workshop on Software Architectures for Dependable Systems, May 2003 (for more information, see http://www.cs.kent.ac.uk/events/conf/2003/wads/); David Garlan, Self-healing Systems Course, #17-811, Carnegie Mellon University seminar, Spring 2003 (for more information see http://www-2.cs.cmu.edu/~garlan/17811/); D. Garlan, J. Kramer, and A. Wolf, eds., *Proceedings of the First Workshop on Self-healing Systems*, ACM Press, New York, 2002.

[30] M. Hamblen, "IBM to Boost Self-healing Capabilities in Tivoli Line," *Computerworld*, April 4, 2003, available at http://www.computerworld.com/softwaretopics/software/story/0,10801,80050,00.html.

[31] "Windows 2000 Professional: Most Reliable Windows Ever," December 5, 2000, available at http://www.microsoft.com/windows2000/professional/evaluation/business/overview/reliable/default.asp.

[32] "Sun and Raytheon Create Open, Adaptive, Self-healing Architecture for DD 21," available at http://wwws.sun.com/software/jini/news/Jini-Raytheon.pdf.

[33] "HP Delivers Self-healing and Virtual Server Software to Advance the Adaptive Enterprise," press release, May 6, 2003, available at http://www.hp.com/hpinfo/newsroom/press/2003/030506c.html.

In fact, many of the techniques described as self-healing are familiar to the decades-old hardware field of reliable systems, also known as fault tolerance or high availability. These techniques, such as fault detection, fault masking, and fault tolerance, are in common use when designing hardware to improve the reliability and availability of large systems. This is most likely because hardware designers, unlike software programmers, long ago accepted the unavoidable reality that components of their designs will fail at some point. (It also helps immeasurably that hardware failures are often easier to characterize than software failures.) In areas with extremely high demands for reliability, such as aerospace or power plants, these fault-tolerance techniques have become quite sophisticated, as have mechanisms for testing system operation. The oldest and most accepted use of the term self-healing is found in networking;[34] networks from the original ARPANET (and even the public switched telecommunications network) to modern peer-to-peer embedded networks are self-healing in the sense that traffic is routed around unresponsive nodes.

In contrast, until quite recently, software quality has focused on producing bug-free products, by an intensive effort of careful design, code review, and extensive prerelease testing. However, when bugs do occur, software typically has no ability to detect or react to them, or to continue to operate. This was a workable strategy for much of the history of modern software, but the continuing rise of the complexity of software applications has made formal review or correctness proofs inadequate to provide minimum levels of reliability.[35]

This rise in complexity and the resulting rise in human cost of configuration and maintenance of software applications has spurred interest in self-healing, hoping to shift much of the burden of this configuration and maintenance back to the software. The idea is that, like its biologically analogous namesake, a self-healing system would detect the presence of nonfunctioning (or, more challengingly, malfunctioning) components and initiate some response to continue proper overall functionality, preferably without any centralized or external force (such as a system administrator) required. The most common implementation today seems to be one of reconfiguration: if a fault is detected, a spare hardware component is brought into play. This is "healing" only in the loosest sense, although it certainly is a valid fault tolerance technique. However, it doesn't translate well to software-only failures.

None of the systems that describe themselves as self-healing (such as Microsoft Windows 2000, IBM DB/2, or Sun's Jini) seem to actually employ biological principles, other than in the grossest sense of having redundancy. However, one research project that is inspired very explicitly by biology is Swarm at the University of Virginia.[36] The Swarm programming model defines units as individual cells, which can both reproduce through cellular division and die. Additionally, they can emit signals at various strengths and respond to the aggregate strength of signals in the environment. For example, a system set to grow to a certain size would start with a single cell that emitted a small amount of signal and with a program set to reproduce if the aggregate signal was at a certain threshold. Until the total amount of signal exceeded that threshold, the cells would continue to divide, but they would stop once the threshold was exceeded. If cells were to fail or otherwise be deleted, other cells would respond by dividing again to bring the signal back to the threshold. This is indeed a primitive form of self-healing. However, this programming model is unlikely to catch on for complex tasks without significant higher-level abstractions available.

[34] W.D. Grover, "The Self-healing Network: A Fast Distributed Restoration Technique for Networks Using Digital Cross-connect Machines," *Proceedings of the IEEE Global Telecommunications Conference*, Tokyo, 1987, pp. 1090-1095.

[35] In his lecture on receiving the ACM Turing Award in 1980, C.A.R. Hoare said, "There are two ways of constructing a software design: One way is to make it so simple that there are obviously no deficiencies, and the other way is to make it so complicated that there are no obvious deficiencies." Lecture available at http://www.braithwaite-lee.com/opinions/p75-hoare.pdf.

[36] G. Selvin, D. Evans, and L. Davidson, "A Biologically Inspired Programming Model for Self-healing Systems," *Proceedings of the First Workshop on Self-Healing Systems*, November 2002, available at http://www-2.cs.cmu.edu/~garlan/woss02/.

8.2.5 Immunology and Computer Security[37]

The mammalian immune system is an information processor—this is clear from its ability to distinguish between self and nonself. (Section 5.4.4.3 provides a brief introduction to the immune system.) Some have thus been drawn to the architecture of the immune system as a paradigm of information processing that might be useful in solving a variety of different computational problems. Immunological approaches have been proposed for solving problems in computer security, semantic classification and query, document and e-mail classification, collaborative filtering problem, and optimization.[38] This section concentrates on computer security applications.

8.2.5.1 Why Immunology Might Be Relevant

Computer and network security is intended to keep external threats at bay, and this remains an intellectually challenging problem of the highest order. It is useful to describe two general approaches to such security problems. The first, widely in use today, is based on the notion of what might be called environmental control—the idea that by adequately controlling the environment in which a computer or network functions, better security can be obtained. The computer or network environment is defined broadly, to include security policy (who should have what rights and privileges), resources (e.g., programs that provide users with computing or communications capability), and system configuration. In support of this approach, a number of reports[39] cite security problems that arise from flaws in security policy, bugs in programs, and configuration errors and argue that correcting these flaws, bugs, and errors will result in greater security.

A complementary approach is to take as a given the inability to control the computing or network environment.[40] This approach is based on the idea that computer security can result from the use of system design principles that are more appropriate for the imperfect, uncontrolled, and open environments in which most computers and networks currently exist. Note that there is nothing mutually exclusive about the two approaches—both could be used in the design of an effective overall approach to system or network security.

For inspiration in addressing problems in computer security, some researchers have considered the immune system and the unpredictable and largely hostile environment in which it functions.[41] That is, the unpredictable pathogens to which the immune system must respond are analogous to some of the threats that computer systems face, and the principles underlying the operation of the immune system may provide new approaches to computer security.

8.2.5.2 Some Possible Applications of Immunology-based Computer Security

A variety of loose analogies between computer security and immunology are intuitively obvious, and there is clearly at least a superficial conceptual connection between the protection afforded to

[37] The discussion in Section 8.2.5 owes much to Stephanie Forrest of the University of New Mexico.

[38] For a view of the immune system as information processor, see S. Forrest and S. Hofmeyr, "Immunology as Information Processing," *Design Principles for Immune Systems and Other Distributed Autonomous Systems*, L.A. Segal and I.R. Cohen, eds., Oxford University Press, 2000. For an overview of various applications of an immunological computing paradigm, see www.hpl.hp.com/personal/ Steve_Cayzer/downloads/030213ais.ppt and references therein.

[39] National Research Council, *Cybersecurity Today and Tomorrow: Pay Now or Pay Later*, National Academy Press, Washington, DC, 2002.

[40] This discussion is based on A. Somayaji, S. Hofmeyr, and S. Forrest, "Principles of a Computer Immune System," *Proceedings of the 1997 Workshop on New Security Paradigms*, ACM Press, Langdale, UK, 1998, pp. 75-82.

[41] One of the first papers to suggest that self-nonself discrimination, as used by the immune system might be useful in computer security was by S. Forrest, A.S. Perelson, L. Allen, and R. Cherukuri, "Self-nonself Discrimination in a Computer," *Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy*, IEEE Computer Society Press, Los Alamitos, CA, 1994, pp. 202-212. This paper focused mainly on the issue of protection against computer viruses but set the stage for a great deal of subsequent work.

human beings by the immune system and computer security. The following examples are adapted from Somayaji et al.:[42]

- *Protecting active processes on a single host*. For this application, a computer running multiple processes might be conceptualized as a multicellular organism (in which each process is analogous to a cell). An adaptive immune system could be a detector process that queried other processes to see whether they were functioning normally. If not (i.e., if the detector process found "nonself" in its probes), the adaptive system could slow, suspend, kill, or restart the misbehaving process. One approach to detection (positive detection) is based on the establishment of a profile of observed normal behaviors and using that profile to notice when a program behaves abnormally.[43]
- *Protecting a network of computers.* For this application, each computer in a network might be conceptualized as a cell in an individual. Each process would still be considered as a cell, but now an individual is a network of computers. (Another possible analogy for the network of computers is that each computer represents a single organism and population-level protections are achieved by the collective group through independence, diversity, and sharing of information.) An adaptive detector process could be implemented as described above, with the added feature that these detectors could migrate between computers, thereby enabling all computers on the network to benefit from the detection of a problem on one of them.
- *Protecting a network of disposable computers*. This application is similar to that described above, with the addition that when an anomaly is detected, the problematic machine can be isolated, rebooted, or shut down. If the true source of the anomaly were outside the network, a detector process or system could stand in for the victimized machine, doing battle with the malicious host and potentially sacrificing itself for the good of the network. Note that this application requires that hosts be more or less interchangeable—otherwise the network could not afford the loss of a single host.

8.2.5.3 Immunological Design Principles for Computer Security

The immune system exhibits a number of characteristics—one might call them design principles—that could reasonably describe how effective computer security mechanisms might operate in a computer system or network. (As in Section 5.4.4.3, "immune system" is understood to mean the *adaptive* immune system.) For example, the immune system is:[44]

- *Distributed*, in the sense that it has no central point of control. Instead, the components of the immune system interact locally to mount responses to foreign pathogens (e.g., pathogen detectors [lymphocytes] operate locally to flag the presence of pathogens). By contrast, a computer system based on centralized control is vulnerable to "decapitation"—a successful attack on the point(s) of centralized control renders the system entirely useless.[45]
- *Diverse*, in the sense that because of the ways in which pathogen detectors are produced, each individual human being can detect a somewhat different set of pathogens—a diversity that protects

[42] A. Somayaji, S. Hofmeyr, and S. Forrest, "Principles of a Computer Immune System," *Proceedings of the 1997 Workshop on New Security Paradigms,* ACM Press, Langdale, UK, 1998, pp. 75-82.

[43] An alternative approach is to use a randomly generated detector or set of detectors, living for a limited amount of time, after which it would be replaced by another detector. Detectors that proved particularly useful during their lifetimes (e.g., by detecting new anomalies) could be given a longer life span or allowed to spawn related processes. This approach has been used by Forrest et al. in the development of a network intrusion detection system known as LISYS.

[44] This discussion of the immune system is based on S. Forrest and S. Hofmeyr, "Immunology as Information Processing," *Design Principles for Immune Systems and Other Distributed Autonomous Systems*, L.A. Segal and I.R. Cohen, eds., Oxford University Press, New York, 2001.

[45] A distributed, mobile agent architecture for security was also proposed in M. Crosbie and G. Spafford, "Active Defense of a Computer System Using Autonomous Agents," Technical Report 95-008, Department of Computer Science, Purdue University, 1995.

the species as a whole. By contrast, computer system monoculture (i.e., lack of diversity) implies that systems share vulnerabilities, and a successful attack on one system is likely to succeed on other systems as well.[46]

- *Autonomous*, in the sense that it classifies and eliminates pathogens and repairs itself by replacing damaged cells without the benefit of any centralized control mechanism. Given the growing security burden placed on today's computer systems and networks, it will be increasingly desirable for these system and networks to manage security problems with minimal human intervention.
- *Tolerant of error*, in the sense that some mistakes in identification of pathogens (false positives or false negatives) are not generally fatal and do not cause immune system collapse, although they can cause lingering autoimmune disease. Such tolerance is in part the result of a multilayered design of the immune system, in which multiple, independently architected layers of defense ("defense in depth") operate to provide levels of protection that are not achievable by any single mechanism.[47] Computer systems are often not so tolerant, and small errors or problems in some part of a system can lead to significant malfunctions.
- *Dynamic*, in the sense that pathogen detectors are continually being produced to replace those that are (routinely) destroyed. These detectors, circulated through the body, provide whole-body protection and may be somewhat different in each new generation (in that they respond to different pathogens). Because these detectors turn over, the immune system has a greater potential coverage. By contrast, protection against computer viruses, for example, is based on the notion that all threat viruses are known—and most antiviral systems are unable to cope with a new virus for which no signature is known.
- *Capable* of remembering (adaptable), in the sense that the immune system can learn about new pathogens and "remember" how it coped with one pathogen in order to respond more effectively to a future encounter with the same or a similar pathogen. It can also "forget" about nonself entities that are incorporated into the body (e.g., food gets turned into body parts). Computer systems must also adapt to new environments, as for example, when new software is added legitimately, as well as identify new threats.
- *Imperfect*, in the sense that individual pathogen detectors do not identify pathogens perfectly, but rather respond to a variety of pathogens. Greater specificity is obtained through redundant detection of a pathogen using different detector types. By contrast, computer security systems that look for precise signatures of intruders (e.g., viruses) are easily circumvented.
- *Redundant*, in the sense that multiple and different immune system detectors can recognize a pathogen. Pathogens generally contain many parts, called epitopes, that are recognized by immune system detectors; thus, failure to recognize one epitope is not fatal because many others are available for recognition.
- *Homeostatic*, in the sense that the immune system can be regarded as one mechanism through which the human body seeks to maintain a stable internal state despite a changing environment. A computer system can be designed to autonomously monitor its own activities, routinely making small corrections to maintain itself in a "normal" state, even in the face of wide variations in inputs, such as those caused by intruders.[48]

At a deeper level, it is instructive to ask whether the particular methods by which the immune system achieves these characteristics (implements these design principles) have potential relevance to computer security. To address this issue, deeper and more detailed immunological knowledge is necessary, but some work has been done in this area and is described below.

[46] For more discussion of this point, see Computer Science and Telecommunications Board, National Research Council, *Computers at Risk: Safe Computing in the Information Age*, National Academy Press, Washington, DC, 1991.

[47] This point suggests that detection mechanisms are biased to be more tolerant of false negatives than false positives, because threats that are unaffected by one layer (i.e., false negatives) might well be intercepted by another.

[48] A. Somayaji and S. Forrest, "Automated Response Using System Call Delays," *Journal of Computer Security* 6:151-180, 1998.

8.2.5.4 An Example: Immunology and Intruder Detection

To detect pathogens, the immune system generates detectors that can bind to pathogens, and only to pathogens (i.e., do not bind to self). (A detector binding to a pathogen is the marker of a detection event.) To vastly simplify a complex process, the immune system first generates detectors at random. Through a process known as tolerization, detectors that bind to self are destroyed, leaving only detectors that bind to nonself at the end; these detectors are called mature. Mature detectors are released throughout the body; if they do not bind to a nonself entity in some period of time (several days?), they are destroyed (self-destruct?). Those that do bind to nonself entities are regarded as activated detectors. However, an activated detector must receive a second, independent signal (created by the binding of another type of detector to the same pathogen costimulation) to become capable of surviving for a long period of time. These long-term survivors are memory detectors that enable subsequent immune responses to be generated much more rapidly and are the basis for long-term immunity. (Memory detectors have lifetimes that range from days to the lifetime of an organism, and the underlying mechanisms governing their lifetimes are not well understood.)

In the context of computer security, Forrest and Hofmeyr have described models for network intrusion detection and virus detection.[49] In the network intruder detection example, self is defined through a set of "normal" connections in a local area network. Each connection is defined by a triplet consisting of the addresses of the two parties in communication with each other and the port over which they communicate (a total of 49 bits), and the set of all triplets (normal triplets) generated during a training period represents, by definition, normal operation of the network.

When the network operates outside the training period, the intrusion detection system generates random detector strings that are 49 bits in length. Matches are declared according to an "*r*-contiguous-bit" rule—a match is deemed to exist if a random detector string matches some normal triplet in at least *r* contiguous bit positions. In this phase (the maturation phase), detector strings that match some normal triplet are eliminated, leaving only mature detectors that have not matched any normal triplet.

Mature detectors—which might match an abnormal triplet that arises as the result of a network intrusion—are then exposed to the nontraining network operation. If a mature detector matches some triplet found in the nontraining network operation, such a match is potentially a sign of network intrusion (which would be indicated by an unusual pair of systems communicating over an unusual port). If a mature detector does not match any such triplet in a given period of time, it too is eliminated.[50] The remaining detectors—activated detectors—are now fully capable of signaling the presence of abnormal triplets.

However, as a further guard against false positives, the system invoked a mechanism inspired by immunological *costimulation*. Costimulation reduces the likelihood that a pathogen will be indicated when there is no pathogen. After negative selection of lymphocytes occurs, the remaining now-mature lymphocytes are likely to bind to nonself entities encountered. However, before the lymphocytes are "promoted" to memory cells, they must be activated by a costimulatory signal indicating that the substances to which they bind are in fact pathogens. This costimulatory signal is generated independently and reduces the incidence of pathogen detectors that are overly sensitive (and hence the likelihood of autoimmune reactions).

The intrusion detection system implements a costimulatory mechanism as the requirement of a human confirmation of behavior flagged as potentially anomalous—that is, it presents matches signaled by an activated detector to a human operator for confirmation. If the system receives human confirmation within a fixed amount of time, the activated detector responsible for the warning is made

[49] S. Forrest and S. Hofmeyr, "Immunology as Information Processing," *Design Principles for Immune Systems and Other Distributed Autonomous Systems*, L.A. Segal and I.R. Cohen, eds., Oxford University Press, New York, 2001.

[50] In fact, the mature detector is eliminated if it does not exceed some parametrically set threshold (the activation threshold) for the number of matches to abnormal triplets.

into a memory detector (with an indefinite lifetime and a subsequent activation threshold of 1). However, if human confirmation is not forthcoming, the detector responsible is eliminated.

An intrusion detection product based on this approach was introduced in early 2003.[51] The real-world success of this product remains to be seen.

8.2.5.5 Interesting Questions and Challenges

8.2.5.5.1 Definition of Self Any paradigm for computer security that is based on the differentiation of self from nonself must imply some operational definition of self that represents normal and benign operation. It is clear that a good definition is matched to the signature of the threat being defended against, and hence the designer must be able to answer the question, "How would I know my system were under attack?" Thus, self might be definable in terms of memory access patterns on a single host, TCP/IP packets entering and leaving a single host, the collective behavior of a local network of computers, network traffic through a router, instruction sequences in an executing or stored program, sequences of system calls, user behavior patterns, or keyboard typing patterns.[52]

At the same time, computer security must account for the fact that "self" on a computer system, even one that has not been subject to threat or intrusion, changes over time. New users are added, new software is added, and files are created, deleted, and modified in the course of normal activity, even though all such activities may also occur in the course of an attack. That is, the notion of self must be dynamically modifiable.

These points suggest that better insights into characterizing threat signatures dynamically would be helpful if immunological approaches are to be used to enhance computer security.

8.2.5.5.2 More Immunological Mechanisms Another intellectual challenge is to incorporate more of what is known about immunology into computer security. Thus, it is interesting to consider how a number of immunological mechanisms known today might be useful in making the analogy closer, using the functions and design principles of these specific mechanisms within the general context of an immunologically based approach to computer security. One such mechanism is antigen processing and the major histocompatibility complex (MHC). Some pathogens have the ability to "hide" within cells generally recognized as self. Because lymphocytes can detect antigens only by binding to them, they are unable to detect pathogens inside friendly cells. Molecules from the MHC have the ability to bring key parts of such pathogens to the surface of those cells, thereby enabling the lymphocytes to detect them. Moreover, each individual has a different set of MHC molecules; hence the kinds of hidden pathogens that can be brought to a cell's surface are different for different individuals, providing an important immunological diversity in the population as a whole.

An analogous mechanism was implemented in the intrusion detection system described above. Just as certain pathogens are able to hide within cell interiors to avoid detection, the use of detectors that can match a number of subsets of nonself patterns (so that fewer detectors are needed) implies that there exist some nonself patterns for which no detectors can be generated. In other words, a detector capable of matching such nonself patterns would also match some patterns found in self. Furthermore, as the number of nonself patterns that can be recognized by a single detector increases, the number of problematic nonself patterns also increases. Because they result from the structure of the set of self patterns, dynamic change in the detectors cannot find them.

A solution that proved to be effective at reducing the overall number of holes (i.e., gaps in coverage) is multirepresentation—different representations are used for different detectors. One way of achieving

[51] See http://www.sanasecurity.com.

[52] S. Forrest, S.A. Hofmeyr, and A. Somayaji, "Computer Immunology," *Communications of the ACM* 40(10):88-96, 1997.

this is for each detector to have a randomly generated permutation rule, according to which all data path triples are permuted before being matched against the detector. This effectively changes the structure of the self set for each detector, with the result that different detectors will be subject to different holes. Consequently, where one detector fails to detect a nonself triple, another may succeed. Multirepresentation was particularly effective at reducing the number of holes when the nonself patterns were similar to self patterns. To deal with this problem, the bits in a given triplet of connection triplets were randomly permuted before presentation to detectors, just as the specific MHC molecules that are operating to bring pathogens to the surface are probabilistically determined (with respect to an averaging over the population).

8.2.5.6 Some Possible Difficulties with an Immunological Approach

Although these analogies have appeal, it remains to be seen how far they can be pushed. Given that the immune system is a very complex entity whose operation is not fully understood, a bottom-up development of a computer security system based on the immune system is not possible today. The human immune system has evolved to its present state due to many evolutionary accidents as well as the constraints imposed by biology and chemistry—much of which is likely to be artifactual and mostly irrelevant to the underlying principles that the system embodies and also to the design of a computer security system. Further, the immune system is oriented toward problems of survival. By contrast, computer security is traditionally concerned with confidentiality, accountability, and trustworthiness—and the relevance of immunological processes to confidentiality and accountability is entirely unclear today.

8.2.6 Amorphous Computing

An area of research known as amorphous computing seeks to understand how to obtain "coherent behavior from the cooperation of large numbers of unreliable parts that are interconnected in unknown, irregular, and time-varying ways."[53] This work, inspired by observations of cooperative and self-organizing biological phenomena, seeks to identify the engineering principles that can be used to observe, control, organize, and exploit the behavior of cooperating multitudes for human purposes such as the design of engineered artifacts.

An individual entity in a collection of cooperating multitudes has the following characteristics:

- It is inexpensive, in the sense that it is easy to create large numbers of them. For all practical purposes, each entity is identical to every other one.
- It is locally guided or programmed. That is, the guidance or programming is carried by the entity "on-board" rather than being resident elsewhere in the overall system. As a consequence of fabrication, the guidance or programming aboard any given entity is identical to that aboard every other entity.
- It communicates with nearby entities, but in a stochastic manner without the need for precise interconnections and testing. Note also that the ability to function in a stochastically connective environment implies that the overall macrosystem is robust in the face of damaged or nonoperational components. Furthermore, by eliminating the need for precision interconnections, these entities can reduce the enormous costs usually associated with interconnection in traditional forms of assembly, costs that are generally higher than those associated with individual elements.
- It interacts with its environment locally, so that the entity is directly knowledgeable about some aspect of its immediate environment but not about anything more global. To the extent that an individual entity gains global knowledge about the environment, it is as the result of a self-organizing process that develops such information and transmits it to all entities in the system. Similarly, any on-board effectors affect only the immediate environment.

[53] See http://www.swiss.ai.mit.edu/projects/amorphous/.

These characteristics are easily obtained by biology and are increasingly true for certain artifacts that result from today's chip fabrication technologies. A metaphor with some resonance is that of "paintable" computers—a paint that can be applied to a surface, in which are suspended millions of computing and MEMS-like entities that communicate with each other and interact with the surface on which they are painted. (MEMS is an acronym for microelectromechanical systems.)

The vision presented by Abelson et al.[54] is that smart materials may reduce the need for strength and precision in mechanical and electrical apparatus, through the application of computation. For example, coating a building or a bridge with "smart paint'' may enable it to report on traffic loads and wind loads, to monitor the integrity of the structure, to resist buckling, or to heal small cracks by shifting material around. A different kind of smart paint may make it possible to build clean rooms with "active skins," lined with cilia that can push dirt and dust into a corner for removal. Still other paints may enable walls that sense vibration or actively cancel noise. "Smart dust," with light sensors in each particle, could be spread over a wide area to recognize shadows or other traffic passing overhead.

In short, the hope is to create systems with unprecedented responsiveness to their environment. Abelson et al. further argue that the study of amorphous computing has implications for software design in a more general sense. Specifically, a software problem has long been recognized—the dependence of greater functionality of software on increasingly complex software packages and systems. Today, software is mostly developed "by hand," and each line is individually coded. One obtains a high degree of detailed control in this manner, but reliably abstracting the higher-level behavior of a software system so developed is highly problematic. Principles of amorphous computing may enable a more top-down specification of systems that more closely tracks how humans define the functionality they wish to obtain from software.

Amorphous computing may be applicable to fabrication as well. For example, consider amorphous computing entities that are capable of some mechanical interaction with the substrate on which they are painted (e.g., they might expand or contract in certain directions). Nagpal has demonstrated the feasibility of an amorphous computing substrate that is capable of pattern formation (Box 8.1); if the entities making up this formation have the mechanical property described, it is conceivable that they might be able to warp a sheet onto which they were painted into a three-dimensional structure.

It is also conceivable that the vision described in amorphous computing and other approaches to that area could be extended so that appropriately configured microentities could be programmed to self-assemble into useful physical structures on the nanoscale. These structures might be useful to end users in and of themselves, or might serve as nanofabrication machinery that could construct other structures useful to the end user. In particular, the large macromolecules involved in the biochemistry of life—specifically protein molecules—demonstrate the ability to configure themselves into structures, and some research seeks to co-opt biochemical machinery to assemble structures designed for entirely human purposes (as described in Section 8.4.3).

8.3 BIOLOGY AS IMPLEMENTER OF MECHANISMS FOR COMPUTING

8.3.1 Evolutionary Computation[55]

8.3.1.1 What Is Evolutionary Computation?

Evolutionary computation is inspired by genetics and evolutionary events.[56] Given a particular problem for which a solution is desired, evolutionary computation requires three components:

[54] H. Abelson, T.F. Knight, G.J. Sussman, et al., "Amorphous Computing," available at http://www.swiss.ai.mit.edu/projects/amorphous/white-paper/amorph-new/amorph-new.html.

[55] The discussion in Section 8.3.1 owes much to Melanie Mitchell, now at Portland State University in Oregon.

[56] Evolutionary computation is a generic name for techniques that are based loosely on evolutionary principles. There are a number of variants, including evolutionary programming, evolution strategies, genetic programming, and genetic algorithms, which have somewhat different emphases but share the generic approach.

Box 8.1
Pattern Formation Using Identical Autonomous Agents

In a 2001 Ph.D. thesis, Nagpal describes a language for instructing a sheet of identically programmed, flexible, and randomly but densely distributed autonomous agents ("cells'') to assemble themselves into a predetermined global shape, using only local interactions. A wide variety of global shapes and patterns can be synthesized (patterns including flat layered shapes, all plane Euclidean constructions, and a variety of tessellation patterns) using only local interactions between identically programmed deformable cells. That is, the global shape results from a coordinated set of local shape changes in individual cells. Despite being programmed identically, each cell deforms in its own individualized manner, depending on the behavior and state of its neighbors. (The governing structural metaphor is that of epithelial cells, which generate a wide variety of structures: skin, capillaries, and many embryonic structures (gut, neural tube) through the coordinated effect of many cells changing their individual shape.)

The global shape is specified as a folding construction on a continuous sheet, using a small set of axioms, simple initial conditions (edges and corners of the sheet), and two types of folds. From an engineering standpoint, the significance of global shape description is that a process that is inherently local can be harnessed to produce a shape of known configuration. This differs significantly from approaches based on cellular automata, in which the local-to-global relationship is not well understood and there is no framework for constructing local rules to obtain any desired pattern (and patterns "emerge" in a non-obvious way from the local interactions).

In this formalism, the specific global shape desired uniquely determines the program executed by all cells. The cellular program is based on several (biologically inspired) primitives for interacting with the cell's local environment. A cell can change the local environment in ways that create the equivalent of chemical gradients, query its local neighborhood and collect information about the state of local companions (e.g., collect neighboring values of a gradient), broadcast messages to all the cells in its local neighborhood, invert its polarity, connect with neighbors in physical contact to establish communication (thus allowing multiple layers of the sheet to act as a single fused layer), and fold itself along a particular orientation by calling the local fold within its program with two arguments: a pair of neighbors and a cell surface.

Each cell has limited resources and reliability. All cells execute the same program and differ only in a small amount of local dynamic state. The cell program does not rely on regular cell placement, global coordinates, or synchronous operation. Robustness against a small amount of random cell death is achieved by depending on large and dense cell populations, using average behavior rather than individual behavior, trading off precision for reliability, and avoiding any centralized control. Further, global coordinates are not required, because cells are able to "discover" positional information. An average cell neighborhood of 15 is sufficient to reliably self-assemble complex shapes and geometric patterns on randomly distributed cells.

SOURCE: R. Nagpal, "Programmable Self-assembly: Constructing Global Shape Using Biologically-inspired Local Interactions and Origami Mathematics," Ph.D. thesis, MIT Department of Electrical Engineering and Computer Science, June 2001.

- A population of candidate solutions to the problem. For example, these candidate solutions may be a sequence of amino acids that can fold into some protein, a computer program, some encoding of the design for something, or some set of rules in a production system.
- A "fitness" metric by which the "goodness" of a candidate solution can be evaluated. For example, if the program was intended to model the output of some designer circuit, the fitness metric might be based on the performance of a candidate program acting on a test case. That is, given the test case, the fitness metric would be the deviation of the output of the program from a known, appropriate answer. Programs that minimized this deviation would be more fit.

- A mechanism (or mechanisms) by which changes to the candidate solutions can be introduced—portions of different candidate solutions are exchanged, for example, or modified in some small random way.[57]

With these components in place, an evolutionary process takes place. The set of new solutions is evaluated for fitness—those with lower fitness scores are thrown out and those with higher scores are retained. This mutation process is iterated many times, and the result at the end is (supposed to be) a solution that is much better than anything in the starting set.

Initially demonstrated on the solving of what might be called "toy" problems, evolutionary techniques have been used in a variety of business applications, including scheduling and production optimization, image processing, engine design, and drug design. Evolutionary computation has also achieved results that are in some sense competitive with human-developed solutions to quite substantive problems. Competitiveness has a number of possible measures, among them results that are comparable to those produced by a succession of human researchers working on a well-defined problem over a period of years, a result that is equivalent to a previously patented or patentable invention, a result that is publishable in its own right (i.e., independent of its origins), or a result that wins or ranks highly in a judged competition involving human contestants.[58]

Evolutionary computation has demonstrated successes according to all of these measures. For example, there are at least 21 instances in which evolutionary techniques have led to artifacts related to previously patented inventions.[59] Eleven of these infringe on previously issued patents, and ten duplicate the functionality of previously patented inventions in a non-infringing way. Also, while some of the relevant patents were issued many years ago (as early as 1917), others were issued as recently as 2001. Some of the inventions created by evolutionary processes include the ladder filter, the crossover filter, a second-derivative controller, a NAND circuit, a PID (proportional, integrative, and derivative) controller, a mixed analog-digital variable capacitor circuit, a voltage-current conversion circuit, and a cubic function generator. They have also created a soccer-playing program that won its first two games in the Robo Cup 1997 competition and another that ranked in the middle of the field of 34 human-written programs in the Robo Cup 1998 competition, four different algorithms for the transmembrane segment identification problem for proteins, and a variety of quantum computing algorithms, and have rediscovered the Campbell ladder topology for low-pass and high-pass filters.

Evolutionary computation also poses intellectual challenges, as described in the next several sections.

8.3.1.2 Suitability of Problems for Evolutionary Computation[60]

Whether or not an evolutionary approach will be successful in solving a given problem is not yet fully understood. Although many components of a full theory of evolutionary algorithms have been worked out, there are critical gaps that remain open questions.

It is known that the relationship between the representation of a problem, genetic operators, and the objective function is the primary determinant of the performance of an evolutionary algorithm. For any optimization problem, there is always a representation or a genetic operator that makes the optima easy to find with an evolutionary algorithm.[61] In addition, evolutionary algorithms are no better or worse

[57] In biology, "crossover" refers to the process in which chromosomal material is exchanged between chromosomes during cell duplication. The exchanged chromosomal material is analogous to portions of the different candidate solutions. "Mutations" are genetic changes induced as the result of random environmental events.

[58] See http://www.genetic-programming.org.

[59] See http://www.genetic-programming.com/humancompetitive.html. More information on these accomplishments can be found in J.R. Koza, M.A. Keane, M.J. Streeter, W. Mydlowec, J. Yu, and G. Lanza, *Genetic Programming IV: Routine Human-Competitive Machine Intelligence, Series in Genetic Programming,* Volume 5, Springer, New York, 2005.

[60] Lee Altenberg of the University of Hawaii was a major contributor to Section 8.3.1.2.

[61] G.E. Liepins and M.D. Vose, "Representational Issues in Genetic Optimization," *Journal of Experimental and Theoretical Artificial Intelligence* 2(2):101-115, 1990.

than any other search algorithm over the space of all problems.[62] Therefore, problem-specific knowledge must be incorporated either implicitly or explicitly in an evolutionary algorithm for it to perform well. Finally, evolutionary algorithms are dynamical systems, and the systems properties necessary to make them good search algorithms are well characterized.[63]

The primary question that remains to tie together the above is the following: How—and when—can knowledge about the problem be translated into representations and genetic operators that produce an evolutionary algorithm with good performance?

In the absence of this critical link in the theory of evolutionary algorithms, the approach taken by designers resorts to the empirical: try it out and see if it works. Evolutionary approaches provide the greatest advantage over other methods in cases where it is not understood how to construct answers from "first principles" (i.e., logico-deductive procedures), but where approximate solutions can be refined by variation and testing. Such problems can be characterized as "difficult inverse problems," where the inverse refers to finding inputs that produce desired outputs of the system in question.

Moreover, evolutionary techniques tend to work best on problems involving relatively large search spaces and large numbers of variables that are not well understood. Evolutionary algorithms have been able to construct and adapt complex neural networks that are intractable analytically or for which derivative-based back-propagation is inapplicable. Genetic programming has produced complex circuits that infringe on patented inventions. By contrast, problems involving small search spaces can usually be searched systematically, and search spaces being well understood generally means that special-purpose heuristics are available. (For example, the Traveling Salesman Problem is reasonably well understood, and there are very good special heuristics for solving that problem.)

For problems in which evolutionary techniques are unable to find global optima, they may nevertheless find very good approximations that are robust to wide-ranging initial conditions. Thus, the solutions generated may be adequate to the task at hand. For this reason, evolutionary techniques may also be better when data are very noisy or in the presence of a varying fitness function: the algorithm may rapidly produce approximate solutions that track the changing environment, just as evolving species can track environmental changes. (An example of a problem calling for a varying fitness function might be a robot that must learn, online, in a dynamic environment, where the task facing the robot changes over time.)

8.3.1.3 Correctness of a Solution

One of the most challenging aspects of evolutionary computation is evaluating the correctness of a solution derived through evolutionary means. Because evolutionary solutions are cumulative, in the sense that they build on previous solutions, the design process does not have an opportunity to develop solutions that are clean and elegantly designed from first principles. Human inspection of a solution so derived is unlikely to yield much insight. Thus, essentially the only way known today to assess the correctness of such a solution is to subject it to extensive testing. Rather than a human being understanding how the solution achieves its goals, the proposed solution convinces a human being that it will do so.

Note, however, that ascertaining the correctness of any large computational artifact (e.g., a complex software system or a VLSI chip) depends to a large degree on testing. Of course, because the thought and decision-making processes of human beings are not available to public inspection, it is only by observing a human being in action that one develops confidence in the designer's ability to perform

[62] D.H. Wolpert and W.G. Macready, "No Free Lunch Theorems for Optimization," *IEEE Transactions on Evolutionary Computation* 1(1):67-82, 1997, available at http://citeseer.ist.psu.edu/wolpert96no.html.

[63] L. Altenberg, "Open Problems in the Spectral Analysis of Evolutionary Dynamics," pp. 73-102 in *Frontiers of Evolutionary Computation*, A. Menon, ed., Genetic Algorithms and Evolutionary Computation Series, Volume 11, Kluwer Academic Publishers, Boston, MA, 2004.

appropriately under certain circumstances. Thus, in the limit of increasing complexity, testing an evolutionary solution may resemble the Turing test. (In the Turing test, an outside observer is asked to distinguish between a human being's answers to a set of questions and a computer's answers. The computer is said to have passed the Turing test if the outside observer is unable to distinguish between the two.)

8.3.1.4 Solution Representation

In biological organisms, the genetic code of DNA is subject to changes (e.g., mutation), and the impact of these changes becomes manifest as the new mutated code is involved in the reproductive process. That is, the particular DNA sequence of an organism can be said to be biology's representation of a "solution" to the problem of adapting the organism to a particular set of evolutionary selective pressures.

From the standpoint of someone solving a problem with techniques from evolutionary computation, the question arises as to the analogue of DNA. More formally, how is a solution to a computational problem to be represented?

In general, the solution to a computational problem is an algorithm. However, an algorithm can be represented in many different ways. Just as data can be represented as lists of numbers or in graphical form, computer programs (which embed algorithms) can be represented as "source code" that is readable by human beings or as "object code"—the raw ones and zeros of binary computation.

If candidate solutions are to be computer programs, one might imagine that their machine language representation is an obvious possible representation. However, changing a machine language program one bit at a time, at random, is highly likely to prevent the (modified) program from running at all (because previously valid op-codes will be turned into invalid ones), and a nonrunning program is useless. The same comments apply to the source code of a program. By randomly changing characters in the source code file, the most likely result is a program that will not compile and therefore cannot be evaluated in any meaningful way. Thus, attempting to evolve a binary program or the source code of a program would likely result in an extraordinarily slow rate of evolution.

A more robust way to conduct this process is to impose the constraint that the program must be executable. Thus, one might insist that the source code of a program be *syntactically* correct but not place any limits whatsoever on its semantics (on what it does). For example, statements in a program can be represented as combinations of functions with various numbers of arguments, and the only requirement for syntactic correctness is that a function have the right number of arguments.[64] Changes to the program can be effected by changing the functions and the specific arguments to the functions. The result, by definition, is a program that is still syntactically correct, still runs, but does not necessarily do what is desirable. A typical initial program is then created by randomly generating a parse tree. A population of such parse trees is then subject to crossovers that exchange different parts of the various parse trees, or mutations that replace one argument or function with a new argument or function.

8.3.1.5 Selection of Primitives

Closely related to the issue of representations is the question of the appropriate semantic primitives (i.e., the smallest meaningful unit that can be changed). For example, in the representation of programs as parse trees, the relevant primitives are functions with arguments, and the efficacy of a genetic algorithm is strongly dependent on the particular set of functions that the evolutionary process can manipulate.

[64] This approach is based on parse trees, a way of representing statements in computer programs. See J.R. Koza, *Genetic Programming: On the Programming of Computers by the Means of Natural Selection*, MIT Press, Cambridge, MA, 1992.

To illustrate, any computable function can in principle be built from the appropriate combination of Boolean operators (AND, OR, and NOT). But these functions operate at too low a level to build the kind of hierarchical structures needed to do anything complicated. It is for this reason that high-level programming languages have emerged that are not based on these operators directly. Such languages allow the creation of many other kinds of structure. For example, a program intended to undertake financial analysis might benefit from an operator or function that would allow finding the average stock price for the previous month. If its task were to evolve a program for financial analysis, such functions might be included in the set of primitives from which an evolutionary process might draw.

One important aspect of the evolutionary approach is the ability to evolve new operators or new functions that can be used subsequently. In some instances, new structures can emerge spontaneously that are more or less stable; more frequently, it is possible to insert rules that will prevent such structures from changing. Alternatively, functions can be defined automatically—the environment provides slots for function and the ability to call on those function (even if they are no-ops), and the subsequent evolutionary process fills in those spaces with functions.[65]

8.3.1.6 More Evolutionary Mechanisms

The model described above is a very crude model of evolution, incorporating only a few bare essential features. However, biologists have characterized other features of evolution. Two of the most important with possible application to computing are coevolution and development; these are discussed below. Other aspects of evolution, such as diploid behavior and sexual selection, do not at this stage provide obvious new approaches to computing.

8.3.1.6.1 Coevolution Coevolution refers to the biological phenomenon in which two or more species interact as they evolve. For example, a host may be susceptible to infection by a parasite. The host evolves some defenses against the parasite, which in turn stimulates the parasite to evolve ways in which to penetrate or circumvent those defenses. In coevolution, other species—which are also evolving—constitute part of the environment in which a given species is embedded.

One application of coevolution to evolutionary programming is to allow the evolution of testing data simultaneously with the solution. Doing so enables the program to account for a wider range of input. In this case, one fitness function is required for the program to evaluate how well it performs against a given set of test data, while a different fitness function is needed for the test data to evaluate how well it breaks the program.[66]

8.3.1.6.2 Development Development refers to the phenomenon in which biological complexity is shaped by growth within the organism (what might be called maturation) and the action of environmental forces on the organism. It is very difficult to create significant complexity using genetic mechanisms alone. Thus, one intellectual thrust in evolutionary computation focuses on the creation of developmental mechanisms that can be evolved to better create their own complexity. For example, evolutionary techniques can be used to evolve neural networks (see Section 8.3.3.2). In designing neural networks, the problems involve various issues related to the topology and configuration of the network. However, a grammar can be used to generate structures of interest. (A grammar is a formal system of rules that can be used to generate far larger structures.) Grammars can evolve as well, with the fitness function being the complexity of the structures it can generate.

[65] J.R. Koza, *Genetic Programming, 11: Automatic Discovery of Reusable Programs*, MIT Press, Cambridge, MA, 1994.

[66] D. Hillis, "Co-evolving Parasites Improve Simulated Evolution as an Optimization Procedure," *Physica D* 42(1-3):228-234, 1990.

Box 8.2
Genetic Programming in Animation

In the world of computer graphics and animation, it can be difficult to build virtual creatures that behave in a realistic manner and simultaneously remain under the user's direct control. For example, directly controlling the positions and angles of moving objects such as limbs can result in detailed behavioral control, but likely at the expense of achieving physically plausible motions. On the other hand, providing a realistic, physics-based environment in which the relevant dynamics are simulated can result in a higher degree of realism, but will likely make it difficult to achieve the desired behavior, especially as the entities involved become more complex.

One way to manage the complexity of control is to optimize the behavior of the creature against some fitness function. Using evolutionary techniques, it is possible to fabricate creatures that behave realistically without understanding the procedures or parameters used to generate them. Different fitness functions can represent different modes of movement (e.g., swimming, walking, jumping, following a source). This approach forces the user to sacrifice some detailed control, but there is also considerable gain in automating the creation of complexity—and the user still influences the outcome by specifying the fitness function.

For purposes of animation, a creature is determined by its physical morphology (e.g., size, shape, number of legs) and the neural system for controlling the relevant muscle forces (the neural system involves sensors that tell the creature about its immediate environment, effectors that cause motion [analogous to muscles], and neurons that retain some memory of its previous states). Both morphology and neural system can be evolved, resulting in a succession of increasingly "fit" creatures that move realistically in a given mode.

In Sims' work, a developmental process was used to generate the creatures and their control systems. The use of such a process allowed similar components, including their local neural circuitry, to be defined once and then replicated, instead of requiring each to be separately specified. Thus, a coded representation—a genotype—of a creature was established that uniquely defined the phenotype of that creature—its morphology and neural system. By evolving the genotype, different phenotypes emerged.

SOURCE: Adapted from K. Sims, "Evolving Virtual Creatures," *Computer Graphics*, Annual Conference Series (SIGGRAPH '94 Proceedings), July 1994, pp. 15-22.

In this case, the goal is to evolve a neural network that has the potential to learn things, rather than evolving the things themselves that are the object of learning. In the case of a robotic brain, it is too difficult to anticipate all of the possibilities that might face the robot, and thus it is impossible to develop a fitness function that fully reflects this diversity. By giving the brain the ability to learn and reason, one can circumvent this difficulty, and as long as one can develop a fitness function for how well the brain has learned over some period, evolutionary techniques can be used to evolve a robotic brain. (Note that the indirect nature of this approach makes it doubly difficult to understand what is going on.)

An example of such work is that of Sims (Box 8.2).

8.3.1.7 Behavior of Evolutionary Processes

Today, those working in evolutionary computation are not able to predict, in general, how long it will take to evolve some desired solution or determine a priori how large an initial population size should be, how rapidly mutations should occur, or how often genetic crossovers should take place. Obviously, all of these parameters have some potential impact on the rate of evolution and how effective a solution might be. Yet how they should be set and their possible relationship to the nature of a given problem are, in general, not known, although some intuitions exist in this area.

For example, variation in a species results from mutations (involving random changes to a genome) and crossovers (involving exchanges of different parts of existing genomes). One hypothesis is that crossovers result in changes that are much more rapid than those driven by mutation. The argument in favor of this is that genomic exchange is in some sense enabling an organism to build on stable substructures. On the other hand, it may be that evolutionary solutions cannot make good use of existing substructures or that crossover is incapable of integrating existing substructures.

If it is true that evolutionary change is more rapid with crossovers than with mutations, this suggests that programs designed to evolve genetic programs may wish to emphasize crossover in their processes for introducing variation.

8.3.2 Robotics 3: Energy and Compliance Management

Biological systems provide an existence proof that self-effected motion is possible. Furthermore, compared to the locomotion made possible by human engineering, biological mechanisms capable of locomotion appear to be energetically efficient, possible in a wide variety of physical environments, and often small in size.

Given these characteristics, it is not unreasonable to ask what lessons biology might hold for the design of engineered systems for locomotion. For example, one reason that biological systems are energetically efficient is that they are not rigid, but rather compliant, and often have mechanisms for energy recovery. That is, these mechanisms store kinetic energy that might otherwise be dissipated, much as a braking electric car can store in batteries the kinetic energy associated with slowing down. A kangaroo employs such a mechanism in its tail, which acts as a spring that compresses as the kangaroo lands from one jump and then assists the kangaroo in pushing off for the next jump. Full has argued that leg locomotion can be described as a point mass attached to a spring and finds that the ratio of relative leg stiffness[67] to body mass is more or less constant across legged animals spanning a wide range of size.[68] In this context, leg musculature functions not just as a source of power but also as an actuator, a springy "strut" that participates in energy absorption, storage, and return.

A second example is that many-legged animals demonstrate an inherent dynamic stability. Contrary to expectations that locomotion would require complex neural control feedback mechanisms, the structure of the leg itself and its inherent multifunctionality provide a key aspect of the control of the system and the combination of stability and forward momentum needed for locomotion. Indeed, analysis of many-legged animals reveals that this inherent stability arises from the production of large lateral and opposing leg forces when the legs are moving. Modeling these forces as a spring between opposing legs reveals that the system is highly stable against perturbations—and the leg assembly is capable of stabilizing itself without any equivalent of neural reflexes at all. Thus, the animal does not need to devote expensive neurological processing to the supervision of locomotive tasks.

Raibert was one of the pioneers of robotics engineering based on physics-inspired control laws—one for height, one for pitch, and one for speed. A fundamental insight was that running animals make use of dynamic stability—a running animal moving forward is out of balance, but legs move forward in rhythm to break its fall. To model this phenomenon, a one-legged "animal" (the "Planar One-legged Hopper") was created. It consisted of a mechanized pogo stick with a three-part control system—one controlling forward running speed, one controlling body attitude, and one controlling hopping height. Stepping motion was not programmed explicitly, but rather emerged under the constraints of balance

[67] Relative leg stiffness is the weight-normalized, size-normalized spring constant of the leg.

[68] R. Blickhan and R.J. Full, "Similarity in Multilegged Locomotion: Bouncing Like a Monopode," *Journal of Comparative Physiology* 173:509-517, 1993; T.M. Kubow and R.J. Full, "The Role of the Mechanical System in Control: A Hypothesis of Self-stabilization in Hexapedal Runners," *Philosophical Transactions of the Royal Society of London B* 354:849-862, 1999; A. Altendorfer et al., "RHex: A Biologically Inspired Hexapod Runner," *Journal of Autonomous Robots* 11:207-213, 2002.

and controlled travel.[69] With this basic unit, a two-legged running animal (the Planar Biped) could be modeled as a body with two pogo sticks working 180° out of phase.[70] A four-legged animal could consist of two two-legged pairs working in opposition (left front and right rear, for example).[71]

Since Raibert's pioneering work, these insights have been applicable to the design of other artificial legged locomotion devices. For example, an autonomous hexapod named "RHex" has a motor associated with each leg, each of which is springy and is able to turn on its central axis. This design enables RHex to have self-correcting reflexes that enable it to respond to obstacles without computational control. Another family of six-legged robots, called the SPRAWL family, is cockroaches. Each leg, driven by a piston, acts as a spring that enables SPRAWL robots to bounce over objects in their path without feedback from the environment. Analysis of the force pattern exerted by the legs closely matches that exerted by a running cockroach.

Other robots are intended to manipulate objects into precise orientations. The traditional way to build such robots is to build them rigidly, with limb motion effected through motors and gear assemblies to increase torque. However, gear assemblies are inherently imprecise, because their very motion requires some degree of play where the gears meet (i.e., some nonzero compliance). In practice, the effect of compliance in the gears introduces a noise function that greatly complicates the prediction of how a limb will move given a certain motor input, and puts limits on the precision with which the final orientation can be known.

One solution to this problem is to use "direct-drive" motors placed at every joint, thus eliminating the gears entirely.[72] Another solution is based on the deliberate introduction of compliance into a gear assembly. This solution is based on the observation that humans can effect precise positioning without precision in their joints. In particular, natural joints are often based on ball-and-socket mechanisms even when they are intended to exhibit 1 degree of freedom. Soft tissue around and in the ball joint introduces compressive compliance in the joint, allowing it to absorb impact and automatically maintain a degree of tightness in the joint.

In the robot context, Pratt et al. inserted a spring mechanism into a limb joint so that the response lags the input.[73] This spring adds a large but known compliance in series into the joint (so-called series elasticity) that is much larger than the unknown compliance of the gears; thus, the gear compliance can safely be ignored in the prediction of final position. Entirely apart from the increased ease of prediction, the introduction of series elasticity enables a local response to any sudden changes in loading—during which time the motors involved can build up torque to handle that load. Other benefits include shock tolerance, lower reflected inertia, more accurate and stable force control, less damage during inadvertent contact, and energy storage.

8.3.3 Neuroscience and Computing

Natural brains demonstrate an alternative to the traditional von Neumann computing architecture (i.e., a fully serial information processor); thus, it is natural to consider possible lessons of neuroscience for computer design. These lessons occur at varying levels of detail.

[69] See http://www.ai.mit.edu/projects/leglab/robots/2D_hopper/2D_hopper.html; see also M.H. Raibert and H.B. Brown, Jr., "Experiments in Balance with a 2D One-legged Hopping Machine," *ASME Journal of Dynamic Systems, Measurement, and Control* 106:75-81, 1984.

[70] See http://www.ai.mit.edu/projects/leglab/robots/2D_biped/2D_biped.html; see also J. Hodgins and M.H. Raibert, "Planar Biped Goes Head Over Heels," Proceedings ASME Winter Annual Meeting, Boston, December 1987.

[71] See http://www.ai.mit.edu/projects/leglab/robots/quadruped/quadruped.html; see also M.H. Raibert, "Four-legged Running with One-legged Algorithms," pp. 311-315 in *Second International Symposium on Robotics Research*, H. Hanafusa and H. Inoue, eds., MIT Press, Cambridge, MA, 1985.

[72] H. Asada and T. Kanade, "Design of a Direct-Drive Mechanical Arm," *ASME Journal of Vibration, Stress, and Reliability in Design* 105(3):312-316, 1983.

[73] G.A. Pratt, M.M. Williamson, P. Dillworth, J. Pratt, K. Ulland, and A. Wright, "Stiffness Isn't Everything," preprints of the Fourth International Symposium on Experimental Robotics, ISER '95, Stanford, CA, June 30-July 2, 1995.

8.3.3.1 Neuroscience and Architecture in Broad Strokes

The most general lesson is that much of human cognition depends on the ability to ignore most of the information made available by the senses.[74] That is, a very high fraction of the raw information that is accessible through sight, sound, and so on does not participate directly in the human's cognitive processes. Human and mammalian cognition is based on an architecture that involves a flexible, but low-capacity, working memory and attentional selection mechanisms that place events and objects into working memory where they become available for cognitive processing.[75]

This approach of selective attention stands in sharp contrast to traditional algorithms that are designed with the goal of seeking optimal solutions and based on the use of as much information about the problem domain as possible. The architecture of biological computation has generally evolved with a different purpose—the adequate management of a complex, changing, and potentially dangerous environment in real time (where "adequate" means "provides for survival").

This architecture is based on a two-track processing arrangement—a very flexible, albeit slow system that implements consciousness, awareness, and cognition but attends to only few things, and a large number of online, fast-acting, sensory-motor systems that bypass attention and awareness (e.g., eye movements, head and hand movements, posture adjustments, and other reflex and reflex-like responses).

Koch et al. have investigated the utility of such a strategy in multiple contexts: (1) a saliency-based visual attention mechanism that selects highly "salient" location in natural images for further processing;[76] (2) a competitive, two-person video game in which an algorithm that focuses on a restricted portion of the playing field outperforms an "optimal" player when a temporal limitation is imposed on the duration of each move;[77] and (3) an algorithm that rapidly solves the NP-complete bin-packing problem under most conditions.[78]

8.3.3.2 Neural Networks

Biology affords an alternative computing model that (1) appears well suited for many ill-posed problems constrained by uncertainty, which is the problem set for which digital machines to date have been reasonably ineffective; and (2) provides an existence proof that slow and noisy circuits can undertake very rapid computations of a certain class. Furthermore, it provides huge numbers of working examples. Although the mechanisms underlying nerve tissue computation are not well understood despite many decades of study, the fact remains that biology has found incredibly good solutions to many engineering problems, and these approaches may well serve to inform practical solutions for engineering problems posed by human beings. Indeed, although biological tissue is not naturally suited for information processing as understood in traditional terms, the fact that biological tissue can do information processing suggests that the underlying architectural principles must be powerful indeed.

Neural networks are among the most successful of biology-inspired computational systems and are modeled on the massively parallel architecture of the brain—and on the brain's inherent ability to learn

[74] C. Koch, "What Can Neurobiology Teach Computer Engineers?" January 31, 2001, unpublished paper, available at http://www7.nationalacademies.org/compbio_wrkshps/Christof_Koch_Position_Paper.doc.

[75] F. Crick and C. Koch, "Consciousness and Neuroscience," *Cerebral Cortex* 8(2):97-107, 1998.

[76] F. Crick and C. Koch, "Consciousness and Neuroscience," *Cerebral Cortex* 8(2):97-107, 1998; L. Itti and C. Koch, "A Saliency-based Search Mechanism for Overt and Covert Shifts of Visual Attention," *Vision Research* 40(10-12):1489-1506, 2000; L. Itti and C. Koch, "Target Detection Using Saliency-based Attention," *Search and Target Acquisition*, RTO Meeting Proceedings 45, NATO, RTO-MP-45, 2000; L. Itti, C. Koch, and E. Niebur, "A Model of Saliency-based Visual Attention for Rapid Scene Analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* 20:1254-1259, 1998.

[77] J.G. Billock, "Attentional Control of Complex Systems," Ph.D. Thesis, 2001, available at http://sunoptics.caltech.edu/~billgr/thesis/thesiscolor.pdf.

[78] J.G. Billock, D. Psaltis, and C. Koch, "The Match Fit Algorithm: A Testbed for the Computational Motivation of Attention," *International Conference on Computational Science* 2: 208-216, 2001.

from experience.[79] A neural network is a network of nodes and links.[80] The nodes, or units, are very simple processors that correspond to neurons—the brain's electrically active cells—and are usually organized in layers, while the links, or connections, are node-to-node data channels that correspond to synapses—the junctions that convey nerve impulses from one neuron to the next. Each node has an activation level that corresponds to a neuron's rate of firing off nerve impulses, while each link has a numeric weight that corresponds to the strength or efficiency of a synapse.

Digital "activation energy" patterns are presented to the network via the "input layer."[81] From the input layer, the activation surges through the various intermediate layers automatically, with the flow being shaped and channeled by the connection strengths in much the same way that the flow of nerve impulses in the brain is shaped by synapses. Once everything has settled down, the answer can be read out from the pattern of activation on a set of designated output nodes in the final layer.

This computation-by-network architecture is where parallelism is relevant:[82] all of the nodes are active at once, and the activation can travel on any number of paths simultaneously. It is also the basis of the system's ability to learn: since the flow of activation (and, thus, the computation) is shaped by the connection weights, it can be *re*shaped by changing the weights according to some form of learning rule. How the connection weights are modified in response to the input patterns is the content of the learning rule. This seems similar in some ways to what happens in the cerebral cortex, where knowledge and experience are encoded as subtle changes in the synaptic strengths. Likewise in a neural network: with very few exceptions, it will always contain some sort of built-in mechanism that can adjust the weights to improve its performance.

These brain-like characteristics give neural networks some decided advantages over traditional algorithms in certain contexts and problem types. Because they *can* learn, for example, the networks can be trained to recognize patterns and compute functions for which no rigorous algorithms are known, simply by being shown examples. ("This is a letter *B*: **B**. So is this: B.") Often, in fact, they can generalize from the training examples well enough to recognize patterns they've never seen before. And their parallel architecture helps them keep on doing so even in the face of noisy or incomplete data or, for that matter, faulty components. The multiple data streams can do a lot to compensate for whatever is missing.

Training a neural network generally involves the use of a large number of individual runs to determine the best solution (i.e., a specific set of connection weights that enables the network to do its job).[83] Most learning rules have a parameter that controls the rate of convergence between the current solution and the global minimum and another that controls the degree to which the network will ignore local minima. Once the network is trained to demonstrate satisfactory performance, it can be presented with other data.[84] With new data, the network no longer invokes the learning rule, and the connection weights remain constant.

[79] Note that neural networks are only one approach to the general problem of machine learning. A second general approach involves what is called statistical learning techniques, so called because they are techniques for the estimation of unknown probabilistic distributions based on data. These techniques have not, as a rule, been derived from the consideration of biological systems.

[80] Useful online tutorials can be found at http://neuralnetworks.ai-depot.com/3-Minutes/ and http://www.colinfahey.com/2003apr20_neuron/2003apr20_neuron.htm.

[81] Some of this discussion is adapted from http://www.cs.wisc.edu/~bolo/shipyard/neural/neural.html.

[82] Note, however, that this does not represent parallelism on the scale of the brain, where the neurons are numbered in the hundreds of billions, if not trillions. The number of units in a neural network is more likely to be measured in the dozens. In practice, moreover, these networks are usually simulated on ordinary, serial computers—although for specific applications they can also be implemented as specialized microchips. (See the online tutorial at http://www.particle.kth.se/~lindsey/HardwareNNWCourse/home.html.) Still, the parallelism is there in principle.

[83] Some of this is adapted from http://www.cs.wisc.edu/~bolo/shipyard/neural/neural.html.

[84] Note that it is possible to "overtrain" a neural network, which means that the network cannot respond properly to anything but the training data. (This might correspond to rote memorization.) Obviously, such a network is not particularly useful.

Neural networks are most useful for problems that are not amenable to computational approaches and are constrained by strict assumptions of normality, linearity, variable independence, and so on.[85] That is, they work well in classifying objects, capturing associations, and discovering regularities within a set of patterns where the volume, number of variables, or diversity of the data is very great; when the relationships between variables are vaguely understood or the relationships are difficult to describe adequately with conventional approaches; or when the problems in question are ill-posed and involve high degrees of uncertainty.[86] In addition, they are well suited for problems that are subject to distortions in the input data.

Neural networks have been applied to a large number of real-world problems of high complexity, including the following.[87]

- *Optical character recognition*. Commercial OCR (optical character recognition) software packages have incorporated neural network technology since the mid-1980s, when it significantly increased their ability to recognize unfamiliar fonts and noisy, degraded documents such as faxes.[88] Today, OCR systems typically use a mix of neural network and rule-based approaches.
- *Finance and marketing*. Neural networks' ability to detect unanticipated patterns has made them a favored tool for analyzing market trends, predicting risky loans, detecting credit card fraud, managing risk, and many other such tasks in the financial sector.[89]
- *Security and law enforcement*. Neural networks' pattern-detection ability has likewise made them a useful tool for fingerprint matching, face identification, and surveillance applications.[90]
- *Robot navigation*. Neural networks' ability to extract relevant features from noisy sensor data can help autonomous robots do a better job of avoiding obstacles.[91]
- *Detection of medical phenomena*. A variety of health-related indices (e.g., a combination of heart rate, levels of various substances in the blood, respiration rate) can be monitored. The onset of a particular medical condition could be associated with a very complex (e.g., nonlinear and interactive) combination of changes on a subset of the variables being monitored. Neural networks have been used to recognize this predictive pattern so that the appropriate treatment can be prescribed.
- *Stock market prediction*. Fluctuation of stock prices and stock indices is another example of a complex, multidimensional, but in some circumstances at least partially deterministic phenomenon. Neural networks are being used by many technical analysts to make predictions about stock prices based on a large number of factors such as past performance of other stocks and various economic indicators.
- *Credit assignment*. A variety of pieces of information are usually known about an applicant for a loan. For instance, the applicant's age, education, occupation, and many other facts may be available. After training a neural network on historical data, neural network analysis can identify the most relevant characteristics and use them to classify applicants as good or bad credit risks.

[85] This material adapted from http://cfei.geomatics.ucalgary.ca/matlab/ann.html.

[86] See http://www.cs.wisc.edu/~bolo/shipyard/neural/neural.html.

[87] See http://www.emsl.pnl.gov:2080/proj/neuron/neural/what.html; see also http://neuralnetworks.ai-depot.com/Applications.html. Examples in the list below for the topics "detection of medical phenomena" through "engine management" are taken from http://www.statsoftinc.com/textbook/stneunet.html#apps.

[88] See http://www.scansoft.com/omnipage/ocr/. At the time, the state of the art in commercial OCR software was the rule-based approach, in which a system broke each character image into simple features and then identified the letters by reasoning about curves, lines, and such. This approach worked well—but only if the fonts were known and the text was very clean.

[89] See http://neuralnetworks.ai-depot.com/Applications.html; see also http://www.nd.com/ and http://www.walkrich.com/value_investing/howdo.htm.

[90] See http://www.neurodynamics.com/.

[91] See http://ai-depot.com/BotNavigation/Obstacle-Introduction.html.

- *Monitoring the condition of machinery*. Neural networks can be instrumental in cutting costs by bringing additional expertise to scheduling the preventive maintenance of machines. A neural network can be trained to distinguish between the sounds a machine makes when it is running normally ("false alarms") versus those it makes when it is on the verge of a problem. After this training period, the expertise of the network can be used to warn a technician of an upcoming breakdown, before it occurs and causes costly unforeseen "downtime."
- *Engine management*. Neural networks have been used to analyze the input of sensors from an engine. The neural network controls the various parameters within which the engine functions, in order to achieve a particular goal, such as minimizing fuel consumption.

8.3.3.3 Neurally Inspired Sensors

One of the first attempts to draw on the principles underlying biological sensors occurred in the mid-1980s, when researchers such as Carver Mead and his coworkers at Caltech made their first attempts to create artificial retinas using VLSI technology,[92] with hoped-for applications that ranged from artificial eyes for the blind to better sensors for robots. A second, more recent example of a neurally inspired sensor is the computational sensor of Brajovic and Kanade.[93] Many approaches toward improving machine vision have been based on better cameras with higher resolution and sensitivity, new sensors such as uncooled infrared cameras, and new recognition algorithms. But standard vision systems typically have high latency (a long time between registration of the image on the vision system's sensors and image recognition), induced by the requirements of transferring large amounts of data from the sensor to the processor and processing those large amounts of data quickly. In addition, latency increases more or less linearly with image size. Standard vision systems can also be very sensitive to small details in the appearance of an object in sensor images. A number of processor-based algorithms have been developed that adjust for such variations, but they are often complex and ad hoc, and hence unreliable.

The computational sensor approach borrows biological architectural principles to use low-latency processing and top-down sensory adaptation as techniques for speeding up vision processes. Computational sensors are (usually) VLSI circuits that include on-chip processing elements tightly coupled with on-chip sensors, exploit unique optical design or geometrical arrangement of elements, and use the physics of the underlying material for computation. The integration of sensor and processor elements on a VLSI chip enables latency to be reduced by a considerable factor and provides opportunities for fast processor-sensor feedback in service of top-down adaptation—and computational sensors have produced an order-of-magnitude improvement in sensing and information processing itself, such as range sensing, sorting, high-dynamic range imaging, and display.

8.3.4 Ant Algorithms

Ant colonies depend on workers that can collectively build nests, find food, and carry out a multitude of other complex tasks while having little or no intelligence of their own. Further, they must do so without the benefit of a leader to organize their efforts. They also continue to do so even in the face of outside disruptions, or the failure and death of individual members, thereby exhibiting a high degree of flexibility and robustness.

[92] M.A. Sivilotti, M.A. Mahowald, and C. Mead, "Real-time Visual Computations Using Analog CMOS Processing Arrays," pp. 295-312 in *Advanced Research in VLSI: Proceedings of the 1987 Stanford Conference*, P. Losleben, ed., MIT Press, Cambridge, MA, 1987.

[93] V. Brajovic, "Computational Sensor for Global Operations in Vision," Ph.D. Thesis, Carnegie Mellon University, Pittsburgh, PA, 1996.

8.3.4.1 Ant Colony Optimization

Entomologists have devoted a great deal of research to figuring out how the social insects achieve these feats.[94] Their answers, in turn, have led computer scientists to devise a variety of "ant algorithms," all of which attempt to capture some of those same qualities of bottom-up self-organization, flexibility, and robustness.[95] Ant algorithms are an example of agent-based models—a broad class of simulations that began to emerge in the early 1990s as researchers tried to model complex adaptive systems on a computer. The idea was to represent different agents with variables that weren't just numbers, as they would be in conventional econometric models, but complex data structures that could respond and adapt to one another—rather like agents in the real world. (In practice, each agent could be modeled as an expert system, a neural network, or any number of other ways.)

The first ant-based optimization—the Ant Colony Optimization algorithm—was created in the early 1990s.[96] The algorithm is based on observations of ant foraging, something that ants do with high efficiency. Imagine that worker ants wandering far from the nest come across a rich food source. Each ant carrying food back to the nest marks her trail by laying pheromone on the ground. When another randomly moving ant encounters this previously marked trail, it will follow it with high probability and reinforce the trail with its own pheromone. This behavior is thus characterized by a positive feedback loop in which the probability with which an ant chooses a given trail increases with the number of ants that previously chose the same trail.

Because the first ant to reach the nest will be the one whose path just happens to be the shortest, there will be a period of time during which the shortest path is the only path to the nest. This fact provides a "seed" around which further pheromone depositions can occur and collectively converge on a path that is one of the shortest possible.

The paradigmatic application of this algorithm is the Traveling Salesman Problem. A salesman is assigned to visit a specified list of cities, going through each of them once and only once before returning to his starting point. In what sequence should he visit them so as to minimize his total distance?

What makes the Traveling Salesman Problem difficult is that there seems to be no guaranteed way of finding the absolute shortest path other than to check every possible sequence, and the number of such sequences grows explosively as the number of cities increases, quickly outstripping the computational ability of any computer imaginable.[97] As a result, practical programmers have had to give up on

[94] See, for example, E.O. Wilson and B. Hölldobler, *The Ants,* Belknap Press of Harvard University Press, Cambridge, MA, 1990.

[95] Overviews can be found in E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*, Oxford University Press, New York, 1999; E. Bonabeau, "Swarm Intelligence," presented at the O'Reilly Emerging Technology Conference, available at http://conferences.oreillynet.com/presentations/et2003/Bonabeau_eric.ppt; and E. Bonabeau and G. Theraulez, "Swarm Smarts," *Scientific American* 282(3):72-79, 2000.

[96] M. Dorigo, "Optimization, Learning, and Natural Algorithms," Ph.D. Dissertation, Politecnico di Milano, Italy, 1992; M. Dorigo, V. Maniezzo, and A. Colorni, "The Ant System: An Autocatalytic Optimizing Process," Technical Report No. 91-016 Revised, Politecnico di Milano, Italy, 1991; M. Dorigo, V. Maniezzo, and A. Colorni, "Positive Feedback as a Search Strategy," Technical Report No. 91-016, Politecnico di Milano, Italy, 1991 (later published as M. Dorigo et al., "The Ant System: Optimization by a Colony of Cooperating Agents," *IEEE Transactions on Systems, Man, and Cybernetics-Part B* 26(1):29-41, 1996, available at ftp://iridia.ulb.ac.be/pub/mdorigo/journals/IJ.10-SMC96.pdf.); M. Dorigo, T. Stützle, and G. Di Caro, eds., *Future Generation Computer Systems* (Special Issues on Ant Algorithms) 16(8), 2000. Dorigo maintains a Web page on ant colony optimization, including an extensive bibliography (with many papers downloadable), plus links to tutorials and software, available at http://iridia.ulb.ac.be/~mdorigo/ACO/about.html.

[97] If there are N cities in the list, then the number of possible routes is on the order of $N!$—that is, $N \times (N-1) \times (N-2). \dots \times 2 \times 1$. (There are N choices of a place to start, $N-1$ choices of a city to visit next, $N-2$ choices to visit after that, and so on.) This is nothing much to worry about for small numbers: 10 cities yield only 10! = 3.628 million paths, which a personal computer could examine fairly quickly, but 20 cities would yield about 2.4×10^{18} paths—a (very fast) computer that examined one path per nanosecond would take more than 77 years to get through all of them; and 30 cities ($30! = 2.65 \times 10^{32}$) would keep that same computer busy for 8 quadrillion years. In computer science, this is a classic example of an NP-complete problem. An NP-complete problem is both NP (i.e., verifiable in nondeterministic polynomial time) and NP-hard (any other NP problem can be translated into this problem). In an NP-complete problem, the number of computations required to solve it grows faster than any power of its size. ("Verifiable in nondeterministic polynomial time" means that a proposed solution to this problem can be verified in polynomial time on a computer that can execute different instructions depending on its input. Polynomial time means a time that is proportional to some power of the problem's size.)

finding the *best* solution to the Traveling Salesman Problem and its relatives, and instead look for algorithms that find an *acceptable* solution in an acceptable amount of time. Many such algorithms have been developed over the years, and the Ant Colony Optimization algorithm has proved to rank among the best—especially after Dorigo and his colleagues introduced several refinements during the 1990s to improve its scaling behavior.[98]

Variations of the algorithm have also been developed for practical applications such as vehicle routing, scheduling, routing of traffic through a data network, or the design of connections between components on a microchip, and the scheduling of special orders in a factory.[99] The technique is particularly useful in such cases because it allows for very rapid *re*routing in the face of unexpected disruptions in the network. Among the successful commercial applications are plant scheduling for the consumer products giant Unilever; truck routing for the Italian oil company Pina Petroli; supply chain optimization and control for the French industrial gas supplier Air Liquide; and network routing for British Telecom, France Telecom, and MCI.[100]

8.3.4.2 Other Ant Algorithms

Ant algorithms are based on two essential principles: (1) self-organization, in which global behavior arises from a myriad of low-level interactions, and (2) stigmergy, in which the individuals interact with one another indirectly using the environment as an intermediary.[101] That is, one individual changes its surroundings (e.g., by laying a pheromone trail), and other individuals then react to those changes at a later time. As researchers have looked to other ant colony behaviors for inspiration, moreover, those same two principles turn up again and again.[102] For example:

- *Sorting behavior*. Certain species of ants apparently have an instinct to keep their surroundings clean; if dead ants are scattered through the nest at random, the workers will immediately begin moving all the corpses into neat little piles (albeit piles in random locations). These ants likewise seem to have an instinct for keeping the brood chambers well organized; if workers are presented with a random jumble of ants-to-be, they will quickly see to it that the eggs and micropupae are in the center, while the larger and more developed pupae and larvae are toward the outside where they have more room. Simulated ants can produce much the same results by following a simple local rule: pick up any item that is isolated—that is, any item that has no others like it in the neighborhood—and drop it whenever many of those items are encountered. Picking things up and then dropping them modifies the environment, while the constant shifting causes the piles and/or broods to self-organize fairly rapidly.

[98] The algorithm and its refinements are discussed at length in Chapter 2 of E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*, Oxford University Press, New York, 1999.

[99] Many of their key papers are available for downloading at M. Dorigo, “Ant Colony Optimization,” 2003, available at http://iridia.ulb.ac.be/~mdorigo/ACO/about.html.

[100] E. Bonabeau, “Swarm Intelligence,” presented at the O’Reilly Emerging Technology Conference, 2003, April 22-25, 2003, Santa Clara, CA, available at http://conferences.oreillynet.com/presentations/et2003/Bonabeau_eric.ppt.

[101] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*, Oxford University Press, New York, 1999.

[102] Among the most notable of these investigators have been entomologist Guy Theraulaz of the French National Center for Scientific Research (CNRS) and telecommunications engineer Eric Bonabeau, formally of France Telecom. Bonabeau, in particular, has been among the most active in the promotion and commercialization of ant algorithms, first as head of European operations for the Santa Fe-based BiosGroup and since 2000 as head of his own company, Icosystem, Inc., of Cambridge, Massachusetts. Details of the various ant behaviors under study, and the algorithms drawn from them, can be found in E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: From Natural to Artificial Systems*, Oxford University Press, New York, 1999; E. Bonabeau and G. Theraulez, “Swarm Smarts,” *Scientific American* 282(3):72-79, 2000; and E. Bonabeau, “Swarm Intelligence,” presented at the O’Reilly Emerging Technology Conference, 2003, available at http://conferences.oreillynet.com/presentations/et2003/Bonabeau_eric.ppt.

- *Division of labor*. In order to gather food, maintain the nest, defend against predators, and so on, a colony has to allocate many different tasks among many different ants simultaneously—again, without the benefit of central planning or individual intelligence. In many cases this is done by a physical caste system, so that workers do certain jobs, soldiers do others, and so on. Yet ants will often allocate tasks even within a single caste. A simple mechanism that reproduces this behavior is to give each individual a response threshold for each task: once the stimuli associated with that task pass the threshold—imagine the smell of accumulating garbage—the individual gets to work. The result is that individuals with higher and higher thresholds keep pitching in until the stimuli are under control, leaving everyone else free to engage in tasks for which *they* have low thresholds.
- *Cooperative transport*. If a single ant encounters a food item that's too big for her to carry alone (e.g., a dead cockroach), she will recruit nest mates via pheromones to help. Now, however, without a leader or brains, they somehow have to start pulling in the same direction. A simple, two-part rule that reproduces the observed behavior is (1) if the object is already moving in the direction you're pulling, keep pulling, and (2) it's not moving at all, or is moving in a different direction, reorient yourself at random and start pulling *that* way. The result is a sequence in which the ants start out pulling their burden from every direction at once, to no effect—until suddenly, when enough ants just happen to line up by accident, a kind of phase transition sets in and the load begins to move.
- *Cooperative construction*. Many species of social insects can build structures of astonishing complexity: witness the vast, hexagonal combs of the honeybee or the multilayered, intricately swirling nests of the paper wasp. And yet again, they manage to do so without the benefit of central planning or individual intelligence. One way to account for such behavior in simulated insects is to equip each individual with a collection of local rules: in situation 1, take action A; in situation 2, take action B; and so on. For a wasp carrying a load of wood pulp, say, such a rule might be, "If you're surrounded by three walls, then deposit the pulp." In general, each insect will modify the environment encountered by the others, and the structure will organize itself in much the same way that the proteins comprising a virus particle assemble themselves inside an infected cell.

Ant algorithms are conceptually similar to the particle swarm optimization algorithm described in Section 8.2.1. However, at least in the case of the Ant Colony Optimization algorithm, it is known that ants really use the algorithm described. For this reason, this algorithm was placed in the category of biologically inspired mechanisms (rather than principles).

8.4 BIOLOGY AS PHYSICAL SUBSTRATE FOR COMPUTING

8.4.1 Biomolecular Computing

The idea of constructing computer components from single molecules or atoms is the logical, if distant, end point of the seemingly inexorable miniaturization of chips and has been foreseen at least since Richard Feynman's lecture "There's Plenty of Room at the Bottom" in 1959.[103] Molecular computing would have significant advantages, most obviously minuscule size of the resulting component, but also a potentially low marginal cost per component and extreme energy efficiency. However, the technology for the precision placing of single atoms or molecules on a large scale is still in its infancy.

However, there is a significant shortcut available: to use biological molecules, including DNA, RNA, and various enzymes, as instruments to perform computational tasks. The sophisticated functions of DNA and related molecules, coupled with the existing technological infrastructure for synthesizing, manipulating, and analyzing them found in molecular biology laboratories, make it feasible to employ them as a universal set of computing components. Also, because the code of DNA is essentially

[103] R.P. Feynman, "There's Plenty of Room at the Bottom," American Physical Society, December 29, 1959; available at http://www.zyvex.com/nanotech/feynman.html.

a digital code, particular strands of DNA can be used to code information, and in particular, joinings and other recombinations of these strands can be used to represent putative solutions to certain computational problems.

This idea is known variously as DNA computation, molecular computation, and biomolecular computation (BMC). The use of DNA as a computational system had been discussed theoretically by T. Head in 1987,[104] but the idea leapt into prominence with Len Adleman's publication in 1994 of a working experiment (Box 8.3) that solved a seven-node instance of the Hamiltonian path problem, an NP-complete problem that is a special case of the Traveling Salesman Problem.[105]

8.4.1.1 Description

Early attention has focused on DNA because its properties are extremely attractive as a basis for a computational system. First, it offers a digital abstraction: the value of a piece of DNA can be precisely *and only* A, G, T, or C. This abstraction is of course quite familiar to the digital abstractions of 0 and 1. Second, the Watson-Crick complementarity of the bases (A with T, G with C) allows matching operations, conceptually similar to "if" clauses in programming. Third, DNA's construction as a string allows a number of useful operations such as insertion, concatenation, deletion, and appending. Next, billions of years of evolution have provided a large set of enzymes and other molecules that perform those operations, some in very specific circumstances. Finally, the last few decades of progress in molecular biology have created a laboratory and instrument infrastructure for the manipulation and analysis of DNA, such as the custom synthesis of sequences of DNA, chips that can detect the presence of individual sequences, and techniques such as polymerase chain reaction (PCR) that can amplify existing sequences. Without such an infrastructure (importantly including the existence of a body of trained laboratory technicians), the use of DNA for computation would be entirely theoretical.

Biomolecular computing provides a number of advantages that make it quite attractive as a potential base for computation. Most obvious are its information density, about 10^{21} bits per gram (billions of times more dense than magnetic tape), and its massive parallelism, 10^{15} or 10^{16} operations per second.[106] Less immediately apparent, but of equal potential importance, is its energy efficiency: it uses approximately 10^{-19} joules per operation, close to the information theoretic limit (compared to 10^{-9} joules per operation for silicon).

One class of biomolecular computing generates witness molecules for all possible solutions to a problem and then uses molecular selection to sift out molecules that represent solutions to the problem at hand. This was the basic architecture developed by Adleman (described in Box 8.3), and with an exponential amount of witness material, this approach can theoretically solve NP-complete problems. Short sequences of DNA (or RNA) are used to represent data, and these are combined to form longer strands, each of which represents a potential solution. Obtaining the particular DNA strand that represents the solution is thus based on laboratory processes that extract the proper DNA strand, and these laboratory processes are based on the existence of an algorithm that can distinguish between correct and incorrect solutions.

A further important step was taken in 2001 by Benenson et al., who developed a programmable finite automaton comprising DNA and DNA-manipulating enzymes that solves certain computational problems autonomously.[107] In particular, the automaton's "hardware" consisted of a restriction nu-

[104] T. Head, "Formal Language Theory and DNA: An Analysis of the Generative Capacity of Specific Recombinant Behaviors," *Bulletin of Mathematical Biology* 49(6):737-759, 1987.

[105] L.M. Adleman, "Molecular Computation of Solutions to Combinatorial Problems," *Science* 266(5187):1021-1024, 1994.

[106] It is only the fact of massive parallelism that makes biological computing at all feasible, because biological switching speeds are diffusion-limited and quite slow.

[107] Y. Benenson, T. Paz-Elizur, R. Adar, E. Keinan, Z. Livneh, and E. Shapiro, "Programmable and Autonomous Computing Machine Made of Biomolecules," *Nature* 414(6862):430-434, 2001.

Box 8.3
Adleman and DNA Computing

Adleman used the tools of molecular biology to solve an instance of the directed Hamiltonian path problem. A small graph was encoded in molecules of DNA, and the "operations" of the computation were performed with standard protocols and enzymes. This experiment demonstrates the feasibility of carrying out computations at the molecular level.

The Hamiltonian path problem is based on finding a special path through an arbitrarily connected set of nodes (i.e., an arbitrary directed graph). (The adjective "directed" means that the connections between nodes are unidirectional, so that a path from A to B does not mean necessarily that another connection from B to A exists.) This path (the Hamiltonian path) is special in the sense that beginning with a specified entering node and ending with a specified exiting node, a continuous path exists that enters and exits every other node once and only once. Hamiltonian paths do not necessarily exist for a given directed graph, and their existence may depend on an appropriate specific choice of entering and exiting nodes.

All known algorithms for determining whether an arbitrary directed graph with designated vertices has a Hamiltonian path exhibit worst-case exponential complexity, which means that there are some directed graphs with a small number of nodes for which this determination would take an impractical amount of computing time.

One method for determining if a Hamiltonian path exists is illustrated in the first column of the table below.

Step	Algorithmic Step	Biological Equivalent
0	Establish directed graph notation as problem representation.	Encode each node and directed node-to-node path as a specific DNA sequence.
1	Generate all possible paths through the graph.	Combine large amounts of these DNA sequences, and with a sufficiently large quantity, the probability that all possible paths will be generated is essentially unity. (In general, these various combinations will be in length several multiples of a single sequence.)
2	Keep only those paths that begin with a specified starting and ending node.	Use polymerase chain reaction (PCR) that amplifies only those molecules encoding paths that begin and end with the specified nodes.
3	If the graph has *n* nodes, then keep only those paths that enter exactly *n* nodes.	Separate only those sequences from step 2 that have the correct length (corresponding to the number of nodes in the graph).
4	Keep only those paths that enter all of the nodes of the graph at least once.	Separate the sequences from step 3 that have a subsequence corresponding to each and every node.
5	If any paths remain, say, "Yes, a Hamiltonian path exists"; otherwise, say "No."	Use PCR amplification on the output of step 4, what remains after step 5 represents the solution to the problem.

SOURCE: Adapted from L.M. Adleman, "Molecular Computation of Solutions to Combinatorial Problems," *Science* 266(5187):1021-1024, 1994.

clease and ligase, while the software and input were encoded by double-stranded DNA. Programming was implemented by choosing appropriate software molecules. The automaton processed the input molecule through a cascade of restriction, hybridization, and ligation cycles, producing a detectable output molecule encoding the automaton's computational result. However, a finite-state automaton is not Turing-complete, and the actual demonstration of a Turing-complete biomolecular machine with a set of primitives sufficient for universal computation has yet to be shown experimentally.[108]

Since Adleman's initial publication, researchers have explored many variants of the basic biological approach. One such variant is the use of RNA, which simplifies the process of removing invalid sequences. In this variant, RNA is used for the solution sequences and DNA is used to represent an element of an invalid solution. Thus, any potential solution that was invalid would be represented by a DNA-RNA hybridized double strand. A single enzyme, ribonuclease H, destroys all DNA-RNA hybridized pairs, leaving only valid solutions. This is significantly simpler than the use of many, potentially noncompatible enzymes necessary to mark and destroy the appropriate DNA-DNA hybrids in the traditional method. (In developing an algorithm based on RNA computing for solving a certain chess problem, Cukras et al.[109] found that although the algorithm was able to recover many more correct solutions than would be expected at random, the persistence of errors continued to present the most significant challenge.)

Other variants of the process seek to automate or simplify the management of stages of the reactions. In the original experiments, the DNA reactions took place in solution in test tubes or other containers, with stages of the process controlled by humans—for example, by introducing new enzymes, changing the temperature (perhaps to break chemical bonds), or mixing DNA solutions. Some of these steps can be automated through the use of laboratory robotics. In some variants, DNA strands are chemically anchored to various types of beads; these beads can be designed with different properties, such as being magnetic or electrically charged, allowing the manipulation of the DNA strands through the application of electromagnetic fields. Another solution is to use microfluidic technologies, which consist of MEMS devices that operate as valves and pumps; a properly designed system of pipettes and microfluidic devices offers significant advantages by automating tasks and reducing the total volume of materials required.[110]

Still another variant is to restrict the chemical operations to a surface, rather than to a three-dimensional volume.[111] In this approach, DNA sequences, perhaps representing all of the solution space of an NP problem, would be chemically attached to a surface. Challenges in this approach include the attachment chemistry, addressing particular strands on the surface, and determining whether chemical attachment interferes with DNA hybridization and enzymatic reactions.

A second class of biomolecular computing begins with an input and a program represented in a molecular form and evolves the program in a number of steps to process the input to produce an output. In this approach, the complexity of the problem does not manifest itself in the number of starting molecules, but rather in the form of the rules provided and the amount of time or number of steps needed to fully evaluate a particular problem and input. For example, in the programmed mutagenesis method, DNA molecules that represent rewrite rules are combined with DNA molecules that encode input data and program. When the combined mixture of these DNA molecules is thermally cycled in the

[108] However, Rothemund has provided a highly detailed description of a Turing-complete DNA computer. See P.W.K. Rothemund, "A DNA and Restriction Enzyme Implementation of Turing Machines," pp. 75-119 in *DNA Based Computers: Proceedings of a DIMACS Workshop,* Vol. 27, R.J. Lipton and E.B. Baum eds., DIMACS Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Princeton, NJ, 1996.

[109] A.R. Cukras, D. Faulhammer, R.J. Lipton, and L.F. Landweber, "Chess Games: A Model for RNA Based Computation," *Biosystems* 52(1-3):35-45, 1999.

[110] A. Gehani and J.H. Reif, "Micro-Flow Bio-Molecular Computation," *BioSystems* 52(1-3):197-216, October 1999.

[111] L.M. Smith, R.M. Corn, A.E. Condon, M.G. Lagally, A.G. Frutos, Q. Liu, and A.J. Thiel, "A Surface-based Approach to DNA Computation," *Journal of Computational Biology* 5(2):255-267, 1998.

presence of DNA polymerase and DNA ligase, the rewrite rules cause new DNA molecules to be produced that represent intermediate states in a computation. These new DNA molecules can be a very general function of the beginning mixture of DNA molecules, and a DNA encoding has been discovered that permits such a system to theoretically implement arbitrary computation.

8.4.1.2 Potential Application Domains

The field of biomolecular computing is still composed of theory and tentative laboratory steps; we are years away from commercial activity. The results of laboratory experiments are proofs of concept; as yet, no biomolecular computer has outperformed an electronic computer.

Biomolecular computing is, in principle, well suited for problems that involve "brute force" solutions, in which candidate solutions can be tested individually to see if they are correct. As noted above, the main application pursued for the first decade of biomolecular computing work is the exhaustive solution of NP-complete problems. While this has been successful for small numbers of nodes (up to 20), the fact that it requires exponential volumes of DNA most likely limits the further development of NP-solving systems (see below for further discussion).

Biomolecular computation also has potential value in the field of cryptography. For example, DNA, with its incredible information density, could serve as an ideal one-time pad, as a tiny sample could provide petabytes of data suitable for use for encryption (as long as it was suitably random). More generally, biomolecules could serve as components of a larger computational system, possibly serving alongside traditional silicon-based semiconductors. For this, and indeed any biomolecular computing system, a challenge is the transformation of information from digital representation into biomolecules and back again. Traditional molecular biological engineering has provided a number of tools for synthesizing DNA sequences and reading them out; however, these tend to be fairly lengthy processes. Recent advances in DNA chips show the potential for more efficient biodigital interfaces. For example, photosensitive chips will synthesize given sequences of DNA based on optical inputs and, similarly, will produce optical signals in the presence of certain sequences. These optical signals are two-dimensional arrays of intensities that can be read by digital image-processing hardware and software. Other approaches for output include the inclusion of fluorescent materials in the DNA molecules or other additives that can be detected with the use of microscopy.

A potential component role for biomolecules is as memory. Whereas biomolecular computation must compete against rapidly improving and increasingly parallel optoelectronic technologies for computation, biomolecular memory is many orders of magnitude superior to conventional magnetic implementations in terms of density. Although DNA memory is unlikely to be used as the rapid-access read-write memory of modern computers, its density makes it useful for "black-box" applications that write a great deal of data, but read only on rare occasions (a fact that would usually tend to increase the acceptable retrieval time).

One such implementation would use DNA as the storage medium of an associative database. A DNA strand would encode the information of a specific record, with sequences on that strand representing attributes of the record and a unique index. Query strings would be composed of the complement of the desired attribute. Although individual lookups would be slow (limited by the speed of DNA chemistry), the total amount of information stored would be enormous and the queries would execute in parallel over the entire database. In contrast, conventional electronic computer implementations of associative memory require linear time with the size of the database.

Such a DNA database might be most useful as a set of tools to manipulate, retrieve, or analyze existing biological or chemical substances. For example, special-purpose DNA computers might search through databases of genetic material. In this model, a large library of genetic material (perhaps representing DNA sequences of various biological lineages, or of criminals) would be stored in its original DNA form, rather than as an electronic digital representation. Biomolecular computers would generate appropriate strands representing a query (matching a sequence found in a new organism, or at a crime scene) and, in massively parallel fashion, identify potential matches. This idea could even be extended to queries of proteins or chemicals, if the appropriate query strand of DNA can be generated.

A separate approach to biomolecular memory uses changes in the sequence of individual strands to represent bits. Certain enzymes known as site-specific recombinases (SSRs) can (among a set of other potential modifications) reverse the sequence of the bases between two marker sequences; repeated application of such an enzyme would flip the sequence back and forth, representing 0 and 1. In this implementation, a single bit requires a long series of bases; research aims at attaining the far more dense use of single bases as bits (in fact, as two bits, since each base can have four values).

8.4.1.3 Challenges

Biomolecular computing faces some significant challenges to adoption beyond the laboratory. The most cited barrier is the exponential doubling of the volume of DNA required to perform exhaustive search of NP-complete problems, such as done by Adleman (Section 8.4.1.1). That is, while the number of different DNA sequences required grows linearly with the number of directed paths in a graph, the volume of those DNA sequences needed to solve a given problem is exponential in the problem's size (in this case, the number of nodes in the graph). Put differently, for the problems to which DNA computing is applicable, a problem that can be solved in exponential time on silicon-based von Neumann computers is replaced by one that can be solved with exponential increases of mass. It is thus an open question today about what kinds of problems can be solved practically using DNA computing. For example, Hartmanis reports that the amount of DNA necessary to replicate Adleman's experiment for a 200-node problem would exceed the weight of the Earth.[112]

While this is a valid concern, standard computers have been widely accepted despite their inability to solve NP-complete problems in a timely fashion. To the best understanding of computer science today, NP-complete problems are fundamentally challenging, and so it ought to be no surprise that even new models of computation struggle with them. Nevertheless some breakthrough may provide subexponential scaling for biomolecular-based exhaustive search.

A second concern involves the time-consuming and expensive laboratory techniques necessary to set up and read out the answer from an experiment—in essence, the input-output problem for biomolecular computing. While DNA reactions themselves offer staggering parallelism (although in fact they take about an hour), the bottleneck may be the time it takes for trained humans to undertake the experiment. Adleman's experiment required about 7 days of laboratory work. And although DNA synthesis itself is cheap, some of the enzymes used in Adleman's experiments cost 10,000 times as much as gold,[113] suggesting that scaling up significantly may not be feasible on economic grounds.

Related to this is the fact that DNA computation is not error-free. Synthesis of sequences can introduce errors; strands of DNA that are close to being complements—but not quite—may still hybridize; point mutations may occur; sheer chance may allow strands of DNA to escape enzymatic destruction; and so forth. Although comparatively high error rates can be acceptable in laboratory environments, this is far more problematic for computation. The problem can be ameliorated partly by the use of techniques familiar to communications protocols, including error-correcting codes and careful design of the code words used in computation, so as to maximize the information distance between any pair. This last example is a good case of computer science and biological cooperation: the distance between a pair of code words composed of a series of bases is a product of both its information content and its biochemical properties. Word design is currently an active area of DNA computation research.

[112] J. Hartmanis, "On the Weight of Computations," *Bulletin of the European Association for Theoretical Computer Science* 55:136-138, 1995.

[113] A.L. Delcher, L. Hood, and R.M. Karp, "Report on the DNA/Biomolecular Computing Workshop (June 6-7, 1996)," National Science Foundation, NSF 97-168, 1998, available at http://www.nsf.gov/pubs/1998/nsf97168/nsf97168.htm.

A related problem is the lack of programmability of current models. Even if experimental verification of Turing-complete biomolecular computing can be achieved, individual runs must still be carefully tuned to a specific instance of a specific problem, much like the hardwiring of the first generation of electronic computers. Worse yet, the sequences of biomolecules synthesized for a particular biomolecular computation are usually consumed or destroyed during the computation. For a replication of the experiment, even with the same dataset, much of the entire process of setup must be repeated. If a different dataset or a different "program" is run, then other steps must be included, such as designing the set of sequences to be used as "words" of the computation and determining the set of enzymes and concentration levels necessary to correctly identify, mark, destroy, and read out the appropriate strands of nucleic acids. The ability to formulate a problem of any generality in terms that map onto a set of chemical processing lab procedures is likely an essential aspect of DNA computing, but it is not at all clear today how such formulations can occur in general.

Finally, the most significant challenge is the high bar that DNA computation will have to surpass to gain wide acceptance. Moore's law is expected to continue unabated for at least a decade, resulting in petaflop machines by 2015. Additionally, biomolecular computation is not the only radical technique in town; quantum computation, various other applications of nanotechnology, analog computing, and other contenders may turn out to offer more favorable performance, programmability, or convenience.

These challenges are quite significant and possibly decisive. Len Adleman himself was pessimistic about the prospect of general computation in a 2002 paper: "Despite our successes, and those of others, in the absence of technical breakthroughs, optimism regarding the creation of a molecular computer capable of competing with electronic computers on classical computational problems in not warranted."[114] Of course, such breakthroughs may yet occur, and this possibility warrants some level of continued research.

8.4.1.4 Future Directions

While it was DNA's resemblance to the tape of a Turing machine that inspired Adleman to investigate the possibility, this model has not yet been pursued experimentally. Nor is it likely that it would have practical computing utility—a Turing machine is extraordinarily slow even executing simple algorithms.

A very different approach would involve single molecules of DNA (or RNA or another biomolecule) acting as the memory of a single process, while enzymes performed the computation by splicing and copying sequences of bases. Although this has been discussed theoretically, it has not yet been shown in an experiment. This model would be best used for massively parallel applications, since the individual operations on DNA are still quite slow compared to electronic components, but it would offer massive improvements of density and energy efficiency over traditional computers.

In a slightly different approach, enzymes that operate on DNA sequences are used as logic gates, such as XOR, AND, or NOT. DNA strands are data, and the enzymes, by reacting to the presence of certain sequences, modify the DNA or generate new strands. Thus, using fairly traditional digital logic design techniques, assemblies of logic gates can be constructed. The resulting circuits will operate in exactly the same manner as traditional silicon electronic-based circuits, but at the energy efficiency and size of molecules.[115]

Even if it turns out that biomolecular computation is a dead end, the research that went into it will not be for naught: the laboratory techniques, enabling technologies, and deeper understanding of

[114] R.S. Braich, C. Johnson, P.W.K. Rothemund, D. Hwang, N. Chelyapov, and L.M. Adleman, "Solution of a 20-Variable 3-SAT Problem on a DNA Computer," *Science* 296(5567):499-502, 2002.

[115] M.N. Stojanovic, T.E. Mitchell, and D. Stefanovic, "Deoxyribozyme-based Logic Gates," *Journal of the American Chemical Society* 124(14):3555-3561, 2002.

biomolecular processes will be valuable. Already, commercial spinoff technologies are available: based on Adleman's research, a company in Japan developed a way to synthesize 10,000 DNA sequences to rapidly search for the presence of genes related to cancer.[116] Also, biologist Laura Landweber's research into biomolecular computation at Princeton has provided insights for her research on DNA and RNA mechanisms in living organisms. For example, her and Lila Kari's analysis of the DNA manipulations that occur in some protozoa is based on techniques of formal languages from computer science, showing that the cellular operations performed by these protozoa are actually Turing-complete. The use of formal computer science theory, in other words, has proven a useful tool for the analysis of natural genetic processes.

8.4.2 Synthetic Biology

As a field of inquiry, the goal of biology—reductionist or otherwise—has been to catalog the diversity of life and to understand how it came about and how it works. These goals emphasize the importance of observation and understanding. Synthetic biology, in contrast, is a new subfield of biology with different intent: based on biological understanding, synthetic biology seeks to modify living systems and create new ones.

Synthetic biology encompasses a wide variety of projects, definitions, and goals and thus is difficult to define precisely. It usually involves the creation of novel biological functions, such as custom metabolic or genetic networks, novel amino acids and proteins, and even entire cells. For example, a synthetic biology project may seek to modify *Escherichia coli* to fluoresce in the presence of TNT, creating in effect a new organism that can be used for human purposes.[117] In one sense, this is a mirror image of natural selection: adding new features to lineages not through mutation and blind adaptation to an environment, but through planned design and forethought. Synthetic biology shares some similarities with recombinant genetic engineering, a common approach that involves transplanting a gene from one organism into the genome of another. However, synthetic biology does not restrict itself to using actual genes found in organisms; it considers the set of all possible genes. In effect, synthetic biology involves writing DNA, not merely reading it.

One basic motivation of this field is that creating artificial cells, or introducing novel biological functions, challenges our understanding of biology and requires significant new insight. In this view, only by reproducing life can we demonstrate that we fully understand it; this is the ultimate acid test for our theories of biology. It is precisely analogous to early synthetic chemistry: only by the successful synthesis of a substance would a theory of its composition be verified.[118]

More broadly, some synthetic biology researchers see created life as an opportunity to explore wider conceptions of life beyond the examples provided by nature. For example, what are the physical limitations of biological systems?[119] Are other self-replicating molecular information systems possible? Are there general principles of biochemical organization? These inquires may help researchers to understand how life began on Earth, as well as the possibility of life in extraterrestrial environments.[120]

Finally, synthetic biology has the potential to contribute significantly to technology, offering in many ways a new industrial revolution. In this view, chemical synthesis, detection, and modification could all be done by creating a microbe with the desired characteristics. This holds the promise of new methods for energy production, environmental cleanup, pharmaceutical synthesis, pathogen detection

[116] *Business Week*, "Len Adleman: Tapping DNA Power for Computers," January 4, 2002.

[117] L.L. Looger, M.W. Dwyer, J.J. Smith, and H.W. Hellinga, "Computational Design of Receptor and Sensor Proteins with Novel Functions," *Nature* 423(6936):185-190, 2003.

[118] S.A. Benner, "Act Natural," *Nature* 421:118, 2003.

[119] D. Endy, quoted in L. Clark, "Writing DNA: First Synthetic Biology Conference Held at MIT," available at http://web.mit.edu/be/news/synth_bio.htm.

[120] J.W. Szostak, D.P. Bartel, and P.L. Luisi, "Synthesizing Life," *Nature* 409(6818):387-390, 2001.

and neutralization, biomaterials synthesis, or any task that can be done by biochemistry. This is essentially a form of nanotechnology, in which the already existing mechanisms of biology are employed to operate on structures at the molecular scale.

However, all of these goals will require a different set of approaches and techniques than traditional biology or any natural science provides. While synthetic biology employs many of the same techniques and tools as systems biology—simulation, computer models of genetic networks, gene sequencing and identification, massively parallel experiments—it is more of an engineering discipline than a purely natural science.

8.4.2.1 An Engineering Approach to Building Living Systems

Although as a viewpoint it is not shared by all synthetic biology researchers, a common desire is to invent an engineering discipline wherein biological systems are both the raw materials and the desired end products. Engineering—particularly, electronics design—is an appropriate discipline to draw on, because no other design field has experience with constructing systems composed of millions or even billions of components. The engineering design approaches of abstraction, modularity, protocols, and standards are necessary to manage the complexity of the biomolecular reality.

One important piece of establishing an engineering discipline of building living systems is to create a library of well-defined, well-understood parts that can serve as components in larger designs. A team led by Tom Knight and Drew Endy at the Massachusetts Institute of Technology (MIT) have created the MIT Registry of Standard Biological Parts, also known as BioBricks, to meet this need.[121] An entry in the registry is a sequence of DNA that will code for a piece of genetic or metabolic mechanism. Each entry has a set of inputs (given concentrations or transcription rates of certain molecules) and a similar set of outputs.

The goal of such a library is to provide a set of components for would-be synthetic biology designers, where the parts are interchangeable, components can be composed into larger assemblies and easily be shared between separate researchers, and work can build on previous success by incorporating existing components. Taken together, these attributes allow the designers to design in ignorance of the underlying biological complexity.

These BioBricks contain DNA sequences at either end that are recognized by specific restriction enzymes (i.e., enzymes that will cut DNA at a target sequence); thus, by adding the appropriate enzymes, a selected DNA section can be spliced. When two or more BioBricks sequences are ligated together, the same restriction sequences will flank the ends of the DNA sequence, allowing the researcher to treat the composite as a single component. BioBricks are in the early stages of research still, and the final product will likely be substantially different in construction.

8.4.2.2 Cellular Logic Gates

Of particular interest to synthetic biologists are modifications to cellular machinery that simulate the operations of classical electronic logic gates, such as AND, NOT, XOR, and so forth. These are valuable for many reasons, including the fact that that their availability in biological systems would mean that researchers could draw on a wide range of existing design experience from electronic circuits. Such logic gates are especially powerful because they increase the ability of designers to build more sophisticated control and reactivity into engineered biological systems. Finally, it is the hope of some researchers that, just as modern electronic computers are composed of many millions of logical gates, a new generation of biological computers could be composed of logic gates embedded in cells.

[121] T. Knight, "Idempotent Vector Design for Standard Assembly of Biobricks," available at http://docs.syntheticbiology.org/biobricks.pdf.

Researchers have begun to construct cellular logic gates in which signals are represented by protein concentrations rather than electrical voltages, with the intent of developing primitives for digital computing on a biological substrate and control of biological metabolic and genetic networks. In other words, the logic gate is an abstraction of an underlying technology (based on silicon or on cellular biology): once the abstraction is available, the designer can more or less forget about the underlying technology.

A biological logic gate uses intracellular chemical mechanisms, such as the genetic regulatory network, metabolic networks, or signaling systems to organize and control biological processes, just as electronic mechanisms are used to control electronic processes.

Any logic gate is fundamentally nonlinear, in the sense that it must be able to produce two levels of output (zero and one), depending on the input(s), in a manner that is highly insensitive to noise (hence, subsequent computations based on the output of that gate are not sensitive to noise at the input). That is, variations in the input levels that are smaller than the difference between 1 and 0 must not be significant to the output of the gate.

Once a logic gate is created, all of the digital logic design principles and tools developed for use in the electronic domain are in principle applicable to the construction of systems involving cellular logic.

A basic construct in digital logic is the inverting gate. Knight et al.[122] describe a cellular inverter consisting of an "output" protein Z and an "input" protein A that serves as a repressor for Z. Thus, when A is present, the cellular inverter does not produce Z, and when A is not present, the inverter does produce Z. One implementation of this inverter is a genetic unit with a binding site for A (an operator), a site on the DNA at which RNA polymerase binds to start transcription of Z (a promoter), and a structural gene that codes for the production of Z.

Protein Z is produced when RNA polymerase binds to the promoter site. However, if A binds to the operator site, it prevents (represses) the binding of RNA polymerase to the promoter site. Thus, if proteins have a finite lifetime, the concentration of Z varies inversely with the concentration of A. To turn this behavior into digital form, it is necessary for the cellular inverter to provide low gain for concentrations of A that are very high and very low, and high gain for intermediate concentrations of A.

Overall gain can be increased by providing multiple copies of the structural gene to be controlled by a single operator binding site. Where high and low concentrations call for low gain, a combination of multiple steps or associations into a single pathway (e.g., the mitogen-activated protein [MAP]-kinase pathway, which consists of many switches that turn on successively) can be used to generate a much sharper nonlinear response for the system as a whole than can be obtained from a single step.

Once this inverter is available, any logic gate can be constructed from combinations of inverters.[123] For example, a NAND gate can be constructed from two inverters that have different input repressors (e.g., A1 and A2) but the same output protein Z, which will be produced unless both A1 and A2 are present. On the other hand, cellular logic and electronic logic differ in that cellular logic circuits are more inherently asynchronous because signal propagation in cellular logic circuits is based on diffusion of proteins, which makes both synchronization and high speed very hard to achieve. In addition, because these diffusion processes are, by definition, not channeled in the same way that electrical signals are confined to wires, a different protein must be used for each unique signal. Therefore, the number of proteins required to implement a circuit is proportional to the complexity of the circuit. Using different proteins means that their physical and chemical properties are different, thus complicating the design and requiring that explicit steps be taken to ensure that the signal ranges for coupled gates are appropriately matched.

[122] T.F. Knight and G.J. Sussman, "Cellular Gate Technology," *Unconventional Models of Computation*, C. Calude, J. Casti, and M.J. Dinneen, eds., Springer, Auckland, New Zealand, 1998.

[123] In general, the availability of an inverter is not sufficient to compute all Boolean functions—an AND or an OR function is also needed. In this particular case, however, the implementing technology permits inverters to be placed side by side to form NOT-AND (NAND) gates.

Cellular circuits capable of logic operations have been demonstrated. For example, Elowitz and Leibler designed and implemented a three-gene network that produced oscillations in protein concentration.[124] The implemented network worked in only a fraction of the cells but did, in fact, oscillate. Gardner et al. built a genetic latch that acted as a toggle between two different stable states of gene expression.[125] They demonstrated that different implementations of the general designs yielded more or less stable switches with differing variances of concentration in the stable states. While both of these applications demonstrate the ability to design a simple behavior into a cell, they also demonstrate the difficulty in implementing these circuits experimentally and meeting design specifications.

In a step toward clinical application of this type of work,[126] Benenson et al. developed a molecular computer that could sense its immediate environment for the presence of several mRNA species of disease-related genes associated with models of lung and prostate cancer and, upon detecting all of these mRNA species, release a short DNA molecule modeled on an anticancer drug.[127] Benenson et al. suggest that this approach might be applied in vivo to biochemical sensing, genetic engineering, and medical diagnosis and treatment.

8.4.2.3 Broader Views of Synthetic Biology

While cellular logic emphasizes the biological network as a substrate for digital computing, synthetic biology can also use analog computing. To support analog computing, the biomolecular networks involved would be sensitive to small changes in concentrations of substances of interest. For example, a microbe altered by synthetic biology research might fluoresce with an intensity proportional to the concentration of a pollutant. Such analog computing is in one sense closer to the actual functionality of existing biomolecular networks (although of course there are many digital elements in such networks as well), but is more alien to the existing engineering approaches borrowed from electronic systems.

For purposes of understanding existing biology, one approach inspired by synthetic biology is to strip down and clean up genomes for maximal clarity and comprehensibility. For example, Drew Endy's group at MIT is cleaning the genome of the T7 bacteriophage, removing all unnecessary sequences, editing it so that genes are contiguous, and so on.[128] Such an organism would be easier to understand than the wild genotype, although such editing would obscure the evolutionary history of the genome.

While synthetic biology stresses the power of hand-designing biological functions, evolution and selection may have their place. Ron Weiss's group at Princeton University has experimented with using artificial selection as a way to achieve desired behavior.[129] This approach can be combined with engineering approaches, using evolution as a final stage to eliminate unstable or faulty designs.

The most extreme goal of synthetic biology is to generate entirely synthetic living cells. In principle, these cells need have no chemical or structural similarity to natural cells. Indeed, achieving an understanding of the range of potential structures that can be considered living cells will represent a profound step forward in biology. This goal is discussed further in Section 9.3.

[124] M.B. Elowitz and S. Leibler, "A Synthetic Oscillatory Network of Transcriptional Regulators," *Nature* 403(6767):335-338, 2000.

[125] T.S. Gardner, C.R. Cantor, and J.J. Collins, "Construction of a Genetic Toggle Switch in *Escherichia coli*," *Nature* 403(6767):339-342, 2000.

[126] Y. Benenson, B. Gil, U. Ben-Dor, R. Adar, and E. Shapiro, "An Autonomous Molecular Computer for Logical Control of Gene Expression," *Nature* 429(6990):423-429, 2004.

[127] In fact, the molecular computer—analogous to a process control computer—is designed to release a suppressor molecule that inhibits action of the drug-like molecule.

[128] W.W. Gibbs, "Synthetic Life," *Scientific American* 290(5):74-81, 2004.

[129] Y. Yokobayashi, C.H. Collins, J.R. Leadbetter, R. Weiss, and F.H. Arnold, "Evolutionary Design of Genetic Circuits and Cell-Cell Communications," *Advances in Complex Systems*, World Scientific, 2003.

8.4.2.4 Applications

While significant from a research view, synthetic biology also has practical applications. A strong driver of this is the rapidly falling cost of custom DNA synthesis. For a few dollars per base pair in 2004, laboratories can synthesize an arbitrary sequence of DNA;[130] these prices are expected to fall by orders of magnitude over the next decade. This not only has enabled research into constructing new genes, but also offers the promise of cost-effective use of synthetic biology for commercial or industrial applications. Once a new lineage is created, of course, organisms can self-replicate in the appropriate environment, implying extremely low marginal cost.

Cells can be abstracted as chemical factories controlled by a host of process control computers. If the programming of these process control computers can be manipulated, or new processes introduced, it is—in principle—possible to co-opt the functional behavior of cells to perform tasks of engineering or industrial interest. Natural biology creates cells that are capable of sensing and actuating functions: cells can generate motion and light, for example, and respond to light or to the presence of chemicals in the environment. Natural cells also produce a variety of enzymes and proteins with a variety of catalytic and structural functions. If logic functions can be realized through cellular engineering, cellular computing offers the promise of a seamlessly integrated approach to process control computing.

Synthetic or modified cells could lead to more rational biosynthesis of a variety of useful organic compounds, including proteins, small molecules, or any substance that is too costly or difficult to synthesize by ordinary bench chemistry. Some of this is already being done by cloning and gene transfection (e.g., in yeast, plants, and many organisms), but synthetic biology would allow finer control, increased accuracy, and the ability to customize such processes in terms of quantity, precise molecular characteristics, and chemical pathways, even when the desired characteristics are not available in nature.

8.4.2.5 Challenges

Synthetic biology brings the techniques and metaphor of electronic design to modify biomolecular networks. However, in many ways, these networks do not behave like electronic networks, and the nature of biological systems provides a number of challenges for synthetic biology researchers in attempting to build reliable and predictable systems.

A key challenge is the stochastic and noisy nature of biological systems, especially at the molecular scale. This noise can lead to random variation in the concentration of molecular species; systems that require a precise concentration will likely work only intermittently. Additionally, as the mechanisms of synthetic biology are embedded in the genome of living creatures, mutation or imperfect replication can alter the inserted gene sequences, possibly disabling them or causing them to operate in unforeseen ways.

Unlike actual electronic systems, the components of biomolecular networks are not connected by physical wires that direct a signal to a precise location; the many molecules that are the inputs and outputs of these processes share a physical space and can commingle throughout the cell. It is therefore difficult to isolate signals and prevent cross-talk, in which signals intended for one recipient are received by another. This physical location sharing also means that it is more difficult to control the timing of the propagation of signals; again, unlike electronics, which typically rely on a clock to precisely synchronize signals, these biomolecular signals are asynchronous and may arrive at varying speeds. Finally, the signals may not arrive, or may arrive in an attenuated fashion.[131]

[130] One firm claims to be able to provide DNA sequences as long as 40,000 base pairs. See http://www.blueheronbio.com/genemaker/synthesis.html. Others suggest that sequences in the 100 base pair range are the longest that can be synthesized today without significant error in most of the resulting strands.

[131] R. Weiss, S. Basu, S. Hooshangi, A. Kalmbach, D. Karig, R. Mehreja, and I. Netravali, "Genetic Circuit Building Blocks for Cellular Computation, Communications, and Signal Processing," *Natural Computing* 2:47-84, 2003.

Aside from the technical challenges of achieving the desired results of synthetic biology projects, there are significant concerns about the misuse or unintended consequences of even successful work. Of major concern is the potential negative effect on the environment or the human population if modified or created organisms became unmanaged, through escape from a laboratory, mutation, or any other vector. This is especially a concern for organisms, such as those intended to detect or treat pollutants, that are designed to work in the open environment. Such a release could occur as a result of an accident, in which case the organism would have been intended to be safe but may enter an environment in which it could pose a threat. More worrisome, an organism could be engineered using the techniques of synthetic biology, but with malicious intent, and then released into the environment. The answer to such concerns must include elements of government regulation, public health policy, public safety, and security. Some researchers have suggested that synthetic biology needs an "Asilomar" conference, by analogy to the conference in 1975 that established the ground rules for genetic engineering.[132]

Some technical approaches to answer these concerns are possible, however. These include "barcoding" engineered organisms, that is, including a defined marker sequence of DNA in their genome (or in every inserted sequence) that uniquely identifies the modification or organism. More ambitiously, modified organisms could be designed to use molecules incompatible with natural metabolic pathways, such as right-handed amino acids or left-handed sugars.[133]

8.4.3 Nanofabrication and DNA Self-Assembly[134]

Nanofabrication draws from many fields, including computer science, biology, materials science, mathematics, chemistry, bioengineering, biochemistry, and biophysics. Nanofabrication seeks to apply modern biotechnological methodologies to produce new materials, analytic devices, self-assembling structures, and computational components from both naturally occurring and artificially synthesized biological molecules such as DNA, RNA, peptide nucleic acids (PNAs), proteins, and enzymes. Examples include the creation of sensors from DNA-binding proteins for the detection of trace amounts of arsenic and lead in ground waters, the development of nonsocial DNA cascade switches that can be used to identify single molecular events, and the fabrication of novel materials with unique optical, electronic, rheological, and selective transport properties.

8.4.3.1 Rationale

Scientists and engineers wish to be able to controllably generate complex two- and three-dimensional structures at scales from 10^{-6} to 10^{-9} meters; the resulting structures could have applications in extremely high-density electronic circuit components, information storage, biomedical devices, or nanoscale machines. Although some techniques exist today for constructing structures at such tiny scales, such as optical lithography or individual atomic placement, in general they have drawbacks of cost, time, or limited feature size.

Biotechnology offers many advantages over such techniques; in particular, the molecular precision and specificity of the enzymatic biochemical pathways employed in biotechnology can often surpass what can be accomplished by other chemical or physical methods. This is especially true in the area of nanoscale self-assembly. Consider the following quote from M.J. Frechet, a chemistry professor at the

[132] D. Ferber, "Synthetic Biology: Microbes Made to Order," *Science* 303(5655):158-161, 2004.

[133] O. Morton, "Life, Reinvented," *Wired* 13.01, 2005.

[134] Section 8.4.3 draws heavily from T.H. LaBean, "Introduction to Self-Assembling DNA Nanostructures for Computation and Nanofabrication," *World Scientific*, CBGI, 2001; E. Winfree, "Algorithmic Self-Assembly of DNA: Theoretical Motivations and 2D Assembly Experiments," *Journal of Biomolecular Structure and Dynamics* 11(2):263-270, 2000; J.H. Reif, T.H. LaBean, and N.C. Seeman, "Challenges and Applications for Self-Assembled DNA Nanostructures," pp. 173-198 in *Proceedings of the Sixth International Workshop on DNA-Based Computers*, A. Condon and G. Rozenberg, eds., DIMACS Series in Discrete Mathematics and Theoretical Computer Science, Springer-Verlag, Berlin, 2001.

University of California, Berkeley, who is a leader in the area of the synthesis and control of molecular architectures on the nanometer scale:[135]

> While most common organic molecules—"small molecules"—have sizes well below one nanometer, macromolecules such as proteins or synthetic polymers have sizes in the nanometer range. Within this size range, it is generally very difficult to control the 3-D structure of the molecules. Nature has learned how to achieve this with proteins and DNA, but most other large synthetic macromolecules have little shape persistence and precise functional group placement is difficult.

It is this fine control of nanoscale architecture exhibited in proteins, membranes, and nucleic acids that researchers hope to harness with these applied biotechnologies, and the goal of research into "self-assembly" is to develop techniques that can create structures at a molecular scale with a minimum of manual intervention.

Self-assembly, also known as bottom-up construction, is a method of fabrication that relies on chemicals forming larger structures without centralized or external control.[136] Because of its ability to run in parallel and at molecular scales, self-assembly is considered to be a potentially important technique for constructing submicron devices such as future electronic circuit components.

Since the role of DNA and related molecules in biology is to generate complicated three-dimensional macromolecules such as proteins, DNA is a natural candidate for a system of self-assembly. Researchers have investigated the potential of using DNA as a medium for self-assembling structures at the nanometer scale. DNA has many characteristics that make it an excellent candidate for creating arbitrary components: its three-dimensional shape is well understood (in contrast to most proteins, which have poorly understood folding behavior); it is a digital, information-encoding molecule, allowing for arbitrary customization of sequence; and it, with a set of easily accessible enzymes, is designed for self-replication. Box 8.4 describes some key enabling technologies for DNA self-assembly.

One important focus of DNA self-assembly research draws on the theory of Wang tiles, a mathematical theory of tiling first laid out in 1961.[137] Wang tiles are polygons with colored edges, and they must be laid out in a pattern such that the edges of any two neighbors are the same color. Later, Berger established three important properties of tiling: the question of whether a given set of tiles could cover an area was undecidable; aperiodic sets of tiles could cover an area; and tiling could simulate a universal Turing machine,[138] and thus was a full computational system.[139]

The core of DNA self-assembly is based on constructing special forms of DNA in which strands cross over between multiple double helices, creating strong two-dimensional structures known as DNA tiles. These tiles can be composed of a variety of combinations of spacing and interconnecting patterns; the most common, called DX and TX tiles, contain two or three double helices (i.e., four or six strands), although other structures are being investigated as well. Ends of the single strands, sequences of unhybridized bases, stick out from the edges of the tile, and are known as "sticky ends" (or "pads") because of their ability to hybridize—stick to—other pads. Pads can be designed to attach to the sticky ends of other tiles. By careful design of the base sequence of these pads, tiles can be designed to connect only with specific other tiles that complement their base sequence.

The congruence between Wang tiles and DNA tiles with sticky ends is straightforward: the sticky ends are designed so that they will bond only to complementary sticky ends on other tiles, just as Wang tiles must be aligned by color of edge. The exciting result of combining Wang tiles with DNA tiles is that DNA tiles have also been shown to be Turing-complete and thus a potential mechanism for computing.

[135] See http://www.cchem.berkeley.edu.

[136] See, for example, G.M. Whitesides et al., "Molecular Self-Assembly and Nanochemistry—A Chemical Strategy for the Synthesis of Nanostructures," *Science* 254(5036):1312-1319, 1991.

[137] H. Wang, "Proving Theorems by Pattern Recognition," *Bell System Technical Journal* 40:1-41, 1961.

[138] A universal Turing machine is an abstract model of computer execution and storage with the ability to perform any computation that any computer can perform.

[139] R. Berger, "The Undecidability of the Domino Problem," *Memoirs of the American Mathematical Society* 66:1-72, 1966.

Box 8.4
Enabling Technologies for DNA Self-replication

DNA Surface Arrays

Current DNA array technologies based on spotting techniques or photolithography extend down to pixel sizes on the order of 1 micron.[1] Examples of these arrays are those produced by Affymetrix and Nanogen.[2] The creation of DNA arrays on the nanometer scale require new types of non-photolithographic fabrication technologies, and a number of methods utilizing scanning probe microscopic techniques and self-assembled systems have been reported.

DNA Microchannels

The separation and analysis of DNA by electrophoresis is one of the driving technologies of the entire genomics area. The miniaturization of these analysis technologies with micron-sized fluidic channels has been vigorously pursued with the end goal of creating "lab on a chip" devices. Examples are the products of Caliper Technologies and Aclara Biosciences.[3] The next generation of these devices will target the manipulation of single DNA molecules through nanometer-sized channels. Attempts to make such channels both lithographically and with carbon nanotubes have been reported.

DNA Attachment and Enzyme Chemistry

Robust attachment of DNA, RNA, and PNA onto surfaces and nanostructures is an absolute necessity for the construction of nanoscale objects—both to planar surfaces and to nanoparticles. The primary strategy is to use modified oligonucleotides (e.g., thiol, amine-containing derivatives) that can be reacted either chemically or enzymatically. The manipulation of DNA sequences by enzymatic activity has the potential to be a very sequence-specific methodology for the fabrication of DNA nanostructures.[4]

DNA-modified Nanoparticles

Nanoscale objects that incorporate DNA molecules have been used successfully to create biosensor materials. In one example, the DNA is attached to a nanometer-sized gold particle, and then the nucleic acid is used to provide biological functionality,while the optical properties of the gold nanoparticles are used to report particle-particle interactions.[5] Semiconductor particles can also be used, and recently the attachment of DNA to dendrimers or polypeptide nanoscale particles has been exploited for both sensing and drug delivery.[6]

DNA Code Design

To successfully self-assemble nucleic acid nanostructures by hybridization, the DNA sequences (often referred to as DNA words) must be "well behaved" (i.e., they must not interact with incorrect sequences). The creation of large sets of well behaved DNA molecules is important not only for DNA materials research by also for large-scale DNA array analysis. An example of the work in this area is the DNA word design by Professor Anne Condon at the University of British Columbia.[7]

DNA and RNA Secondary Structure

The secondary structure of nucleic acid objects beyond simple DNA Watson-Crick duplex formation, whether they are simple single strands of RNA or the complex multiple junctions of Ned Seeman, have to be understood by a combination of experimental methods and computer modeling. The incorporation of nucleic acid structures that include mismatches (e.g., bulges, hairpins) will most likely be an important piece of the self-assembly process of DNA nanoscale objects.[8]

Multistrand DNA Nanostructures and Arrays

The creation of three-dimensional objects with multistrand DNA structures has been pursued for many years by researchers such as Ned Seeman at New York University. Computer scientists such as Erik Winfree at the California Institute of Technology and John Reif at Duke University have been using the assembly of these nanostructures to create mosaics and tile arrays on surfaces. The application of computer science concepts to "program" the self-assembly of materials is the eventual goal. Since single-stranded RNA forms many biologically functional structures, researchers are also pursuing the use of RNA as well as DNA for these self-assembling systems.[9]

[1] A.C. Pease, D. Solas, E.J. Sullivan, M.T. Cronin, C.P. Holmes, and S.P.A. Fodor, "Light-generated Oligonucleotide Arrays for Rapid DNA Sequence Analysis," *Proceedings of the National Academy of Sciences* 91(11):5022-5026, 1994.

[2] See http://www.affymetrix.com and http://www.nanogen.com.

[3] See http://www.caliper.com; and http://www.alcara.com.

[4] A.G. Frutos, A.E. Condon, L.M. Smith, and R.M. Corn, "Enzymatic Ligation Reactions of DNA 'Words' on Surfaces for DNA Computing," *Journal of the American Chemical Society* 120 (40):10277-10282, 1998. Also, Q. Liu, L. Wang. A.G. Frutos, A.E. Condon, R.M. Corn, and L.M. Smith, "DNA Computing on Surfaces," *Nature* 403:175-179, 2000.

[5] C.A. Mirkin, R.L. Letsinger, R.C. Mucic, and J.J. Storhoff, "A DNA-based Method for Rationally Assembling Nanoparticles into Macroscopic Materials," *Nature* 382(6592):607-609, 1996; T.A. Taton, C.A. Mirkin, and R.L. Letsinger, "Scanometric DNA Array Detection with Nanoparticle Probes," *Science* 289(5485):1757-1760, 2000.

[6] F. Zeng and S.C. Zimmerman, "Dendrimers in Supramolecular Chemistry: From Molecular Recognition to Self-Assembly," *Chemical Review* 97(5):1681-1713, 1997; M.S. Shchepinov, K.U. Mir, J.K. Elder, M.D. Frank-Kamenetskii, and E.M. Southern, "Oligonucleotide Dendrimers: Stable Nano-structures," *Nucleic Acids Research* 27(15):3035-3041, 1999.

[7] A. Maranthe, A.E. Condon, and R.M. Corn, "On Combinatorial Word Design," *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* 54:75-90, 2000.

[8] C. Mao, T. LaBean, J.H. Reif, and N.C. Seeman, "Logical Computation Using Algorithmic Self-Assembly of DNA Triple Crossover Molecules," *Nature* 407(6803):493-496, 2000.

[9] E. Winfree, F. Liu, L.A. Wenzler, and N.C. Seeman, "Design and Self-Assembly of Two-Dimensional DNA Crystals," *Nature* 394(6693):539-544, 1998.

Given a set of tiles with the appropriate pads, any arbitrary pattern of tiles can be created. Simple, periodic patterns have been successfully fabricated and formed from a variety of different DNA tiles,[140] and large superstructures involving these systems and containing tens of thousands of tiles have been observed. However, nonperiodic structures are more generally useful (e.g., for circuit layouts), and larger tile sets with more complicated association rules are currently being developed for the assembly of such patterns.

The design of the pads is a critical element of DNA self-assembly. Since the sticky ends are composed of a sequence of bases, the set of different possible sticky ends is very large. However, there are physical constraints that restrict the sequences chosen; pads and their complements should be sufficiently different from other matched pairs, as to avoid unintended hybridization; they should avoid palindromes, and so on.[141] Most importantly, the entire set of pads must be designed so as to produce the desired overall assembly.

The process of DNA self-assembly requires two steps: the first is the creation of the tiles, by mixing input strands of DNA together; then, the tiles are placed in solution and the temperature is lowered slowly until the tiles' pads connect and the overall structure takes form. This process of annealing can take from several seconds to hours.

[140] C. Mao, "The Emergence of Complexity: Lessons from DNA," *PLoS Biology* 2(12):e431, 2004, available at http://www.plosbiology.org/archive/1545-7885/2/12/pdf/10.1371_journal.pbio.0020431-S.pdf.

[141] T.H. LaBean, "Introduction to Self-Assembling DNA Nanostructures for Computation and Nanofabrication," *Computational Biology and Genome Informatics*, J.T.L. Wang et al., eds., World Scientific, Singapore, 2003.

Once the structure is completed, a number of methods can be used to obtain the output if necessary. The first is to image the resulting structure, for example, with an atomic force microscope or transmission electron microscope. In some cases, the structure by itself is visible; in others, tiles can be made distinguishable by reflectivity or the presence of extra atoms such as gold or fluorescents possibly added to a turn of the strand that extends out of the plane. Second, with the use of certain tiles, a "reporter" strand of DNA can be included in such a way that when all the tiles are connected, the single reporter strand winds through all of them. Once the tiling structure completes assembly, that strand can then be isolated and sequenced by PCR or another technique to determine the ordering of the tiles.

8.4.3.2 Applications

DNA self-assembly has a wide range of potential applications, drawing on its ability to create arbitrary, programmable structures. Self-assembled structures can encode data (especially array data such as images); act as a layout foundation for nanoscale structures such as circuits; work as part of a molecular machine; and perform computations.

Since a tiled assembly can be programmed to form in an arbitrary pattern, it is potentially a useful way to store data or designs. In one dimension, this can be accomplished by synthesizing a sequence of DNA bases that encode the data; then, in the self-assembly step, tiles join to the input strand, extending the encoding into the second dimension. This two-dimensional striped assembly can be inspected visually using microscopy, enabling a useful way to read out data. To store two-dimensional data, the input strand is designed with a number of hairpin turns so that the strand weaves across every other line of the assembly; the tiles then attach between adjacent turns of the input strand. The resulting assembly can encode any two-dimensional pattern, and in principle this approach could be extended to three dimensions.

This approach can also be used to create a foundation for nanometer-scale electronic circuits. For this application, the DNA tiles would contain some extra materials, such as tiny gold beads, possibly in a strand fragment that extended above the plain of the tile. After the tiles have formed the desired configuration, chemical deposition would be used to coat the gold beads, increasing their size, until they merge and form a wire. Box 8.5 describes a fantasy regarding a potential application to circuit fabrication.

DNA has been used as a scaffold for the fabrication of nanoscale devices.[142] In crystalline form, DNA has enabled the precise and closely spaced placement of gold nanoparticles (at distances of 10-20 angstroms). Gold nanoparticles might function as a single-electron storage device for one bit, and other nanoparticles might be able to hold information as well (e.g., in the form of electric charge or spin). At one bit per nanoparticle, the information density would be on the order of 10^{13} to 10^{14} bits per square centimeter.

Computation through self-assembly is an attractive alternative to traditional exhaustive search DNA computation. Although traditional DNA computation, such as performed by Adleman, required a linear number of steps with the input size, in algorithmic self-assembly, the computation occurs in a single step. In current experiments with self-assembly, a series of tiles are provided as input, and computation tiles and output tiles form into position around the input. For example, in an experiment that used DNA tiles to calculate cumulative XOR, input tiles represented the Boolean values of four inputs, while output tiles, designed such that a tile representing the value 0 would connect to two identical inputs, and a tile representing the value of 1 would connect to two dissimilar inputs, formed alongside the input tiles. Then, the reporter strand is ligated, extracted, and amplified to read out the answer.[143]

[142] S. Xiao, F. Liu, A.E. Rosen, J.F. Hainfeld, N.C. Seeman, K. Musier-Forsyth, and R.A. Kiehl, "Assembly of Nanoparticle Arrays by DNA Scaffolding," *Journal of Nanoparticle Research* 4:313-317, 2002.

[143] C. Mao, T.H. LaBean, J.H. Reif, and N.C. Seeman, "Logical Computation Using Algorithmic Self-assembly of DNA Triple-crossover Molecules," *Nature* 407:493-496, 2000.

Box 8.5
A Fantasy of Circuit Fabrication

Consider:

> . . . a fantasy of nanoscale circuit fabrication in a future technology. Imagine a family of primitive molecular-electronic components, such as conductors, diodes, and switches, is available from generic parts suppliers. Perhaps we have bottles of these common components in the freezer. . . . Suppose we have a circuit to implement. The first stage of the construction begins with the circuit and builds a layout incorporating the sizes of the components and the ways they might interact. Next, the layout is analyzed to determine how to construct a scaffold. Each branch is compiled into a collagen strut that links only to its selected targets. The struts are labeled so that they bind only to the appropriate electrical component molecules. For each strut, the DNA sequence to make that kind of strut is assembled, and a protocol is produced to insert the DNA into an appropriate cell. These various custom parts are then synthesized by the transformed cells.
>
> Finally, we create an appropriate mixture of these custom scaffold parts and generic electrical parts. Specially programmed worker cells are added to the mixture to implement the circuit edifice we want. The worker cells have complex programs, developed through amorphous computing technology. The programs control how the workers perform their particular task of assembling the appropriate components in the appropriate patterns. With a bit of sugar (to pay for their labor), the workers construct copies of our circuit we then collect, test, and package for use.

SOURCE: H. Abelson, R. Weiss, D. Allen, D. Coore, C. Hanson, G. Homsy, T.F. Knight, Jr., et al., "Amorphous Computing," *Communications of the ACM* 43(5):74-82, 2000.

This approach has two main drawbacks: the speed of individual assemblies, and the error rate. First, the DNA reactions can take minutes or hours, and so any individual computation by self-assembly will likely be substantially slower than using a traditional computer. The potential for self-assembly is that, like exhaustive DNA computation, it can occur in parallel, with a parallelism factor as high as 10^{18}. In the XOR experiment, researchers observed an error rate of 2 to 5 percent. Certainly, this rate may be lowered as experience is gained in designing laboratory procedures and assembly methods; however, the error rate is likely to remain higher than that for electronic computers. For certain classes of problems, an ultraparallel though unreliable approach may be an effective way to compute a solution.

8.4.3.3 Prospects

So far, DNA self-assembly has been demonstrated successfully in the laboratory, constructing relatively simple patterns (e.g., alternating bands, or the encoding of a binary string) that are visible through microscopy. It has also been used successfully for simple computations such as counting, XOR, and addition.

Moving forward, laboratory techniques must improve in sophistication to handle the more complex assemblies and reactions that will accompany large-scale computations or designs. Along with progress in the lab, further theoretical developments are possible in developing algorithms for constructing arbitrary aperiodic patterns.

Although so far DNA self-assembly has used only naturally occurring variants of DNA, a possible improvement is to employ alternative chemistries, such as peptide nucleic acid, an artificial form of DNA in which the backbone has peptide links in place of the phosphate that occurs in natural DNA.

Also, a wide variety of potential geometries exists for crossover tiles. There have been experiments with a so-called 4 × 4 tile, where the sticky ends extend at right angles.

DNA also has the property that its length scale can bridge the gap between molecular systems and microelectronics components. If the issues of surface attachment chemistry, secondary structure, and self-assembly can be worked out, hybrid DNA-silicon nanostructures may be feasible, and a DNA-controlled field effect transistor is one possible choice for a first structure to fabricate. Some other specific near-term objectives for research in DNA self-assembly include the creation of highly regular DNA nanoparticles and the creation of programmable DNA self-assembling systems. For the cell regulatory systems and enzymatic pathways, some specific near-term objectives include the creation of sets of coupled protein-DNA interactions or genes, the simulation and emulation of kinase phosphor-relay systems, and the creation of networks of interconnecting nanostructures with unique enzyme communication paths.

To be adopted successfully as an industrial technology, however, DNA self-assembly faces challenges similar to solution-based exhaustive search DNA computing: a high error rate, the need to run new laboratory procedures for each computation, and the increasing capability of non-DNA technologies to operate at nanoscales. For example, while it is likely true that current lithography technology has limits, various improvements already demonstrated in laboratories such as extreme ultraviolet lithography, halo implants, and laser-assisted direct imprint techniques can achieve feature sizes of 10 nm, comparable to a single DNA tile. Some other targets might be the ability to fabricate biopolymers such as oligonucleotides and polypeptides as long as 10,000 bases for the creation of molecular control systems and the creation of biochemical and hybrid biomolecular-inorganic systems that can be self-assembled into larger nanoscale objects in a programmable fashion.

8.4.3.4 Hybrid Systems

A hybrid system is one that is assembled from both biological and nonbiological parts. Hybrid systems have many applications, including biosensors, measurement devices, mechanisms, and prosthetic devices.

Biological sensors, or biosensors, probe the environment for specific molecules or targets through chemical, biochemical, or biological assays. Such devices consist of a biological detection element attuned to the target and a transduction mechanism to translate a detection event into a quantifiable electronic or optical signal for analysis. For example, antennae from a living silkworm moth have been used as an olfactory sensor connected to a robot.[144] Such antennae are much more sensitive than artificial gas sensors, in this case to moth pheromones. A mobile robot, so equipped, has been shown to be able to follow a pheromone plume much as a male silkworm moth does. When a silkworm moth's antennae are stimulated by the presence of pheromones, the moth's nervous system activities alternate between active and inactive states in a pattern consistent with the activity pattern of neck motor neurons that guide the moth's direction of motion. In the robot, the silkworm moth's antennae are connected to an electrical interface, and a signal generated by the right (left) antenna results in a "turn right" ("turn left") command. This suggests that such signals may play an important role in controlling the pheromone-oriented zigzag walking of a silkworm moth.

[144] Y. Kuwana et al., "Synthesis of the Pheromone-oriented Behaviour of Silkworm Moths by a Mobile Robot with Moth Antennae as Pheromone Sensors," *Biosensors and Bioelectronics* 14:195-202, 1999.

9

Illustrative Problem Domains at the Interface of Computing and Biology

9.1 WHY PROBLEM-FOCUSED RESEARCH?

Problems offered by nature do not respect disciplinary boundaries. That is, nature does not package a problem as a "biology" problem, a "computing" problem, or a "physics" problem. Many disciplines may have helpful insights to offer or useful techniques to apply to a given problem, and to the extent that problem-focused research can bring together practitioners of different disciplines to work on shared problems, this can only be a good thing.

This chapter describes problem domains in which the expenditure of serious intellectual effort can reasonably be expected to generate (or to require!) significant new knowledge in biology and/or computing. Biological insight could take different forms—the ability to make new predictions, the understanding of some biological mechanism, the construction of a new biological organism. The same is true for computing—insight might take the form of a new biologically inspired approach to some computing problem, different hardware, or novel architecture. It is important to note that these domains contain very difficult problems—and it is unrealistic to expect major progress in a short time.

Challenge problems can often be found in interesting problem domains. A "challenge problem" is a scientific challenge focused on a particular intellectual goal or application (Box 9.1). Such problems have a long history of stimulating important research efforts, and a list of "grand challenges" in computational biology originating with David Searls, senior vice president of Worldwide Bioinformatics for GlaxoSmithKline, includes protein structure prediction, homology search, multiple alignment and phylogeny construction, genomic sequence analysis, and gene finding.[1] Appendix B provides a sampling of grand challenge problems found in other reports and from other life scientists.

The remainder of this chapter illustrates problem domains that display the intertwined themes of understanding biological complexity and enabling a novel generation of computing and information science. It incorporates many of the dimensions of the basic knowledge sought by each field and discusses some of the technical and biological hurdles that must be overcome to make progress. However, no claim whatsoever is made that these problems exhaust the possible interesting or legitimate domains at the BioComp interface.

[1] D.B. Searls, "Grand Challenges in Computational Biology," *Computational Methods in Molecular Biology*, S.L. Salzberg, D. Searls, and S. Kasif, eds., Elsevier Science, 1999.

Box 9.1
On Challenge Problems

Challenge problems have a history of stimulating scientific progress. For example:

- The U.S. High Performance Computing and Communications Program focused on problems in applied fluid dynamics, meso- to macroscale environmental modeling, ecosystem simulations, biomedical imaging and biomechanics, molecular biology, molecular design and process optimization, and cognition.[1] These problem domains were selected because they drove applications needs for very high-performance computing.
- A second example is the Text REtrieval Conference (TREC), sponsored by the National Institute of Standards and Technology, in cooperation with the National Security Agency and the Defense Advanced Research Projects Agency. The purpose of this conference is to "support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. . . . The TREC workshop series has the following goals: to encourage research in information retrieval based on large test collections; to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas; to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems."[2] TREC operates by presenting a problem in text retrieval clearly and opening it up to all takers. It makes available to the community at large all basic tools, and its structure and organization have attracted a large number of research sites.
- Still another approach to challenge problems is to offer prizes for the accomplishment of certain well-specified tasks. For example, in aeronautics, the Kremer Prize was established in 1959 for the first human-powered flight over a specific course; this prize was awarded to Paul MacReady for the flight of the *Gossamer Condor* in 1977. The Kremer Prize is widely regarded as having stimulated a good deal of innovative research in human-powered flight. A similar approach was taken in cryptanalysis, in which nominal prizes were offered for the first parties to successfully decrypt certain coded messages. These prizes served to motivate the cryptanalytic community by providing considerable notoriety for the winners. On the other hand, pressures to be the first to achieve a certain result often strongly inhibit cooperation, because sharing one's own work may eliminate the competitive advantage that one has over others.

[1] See http://www.ccic.gov/pubs/blue96/index.html.
[2] See http://trec.nist.gov/overview.html.

9.2 CELLULAR AND ORGANISMAL MODELING[2]

A living cell is a remarkable package of biological molecules engaged in an elaborate and robust choreography of biological functions. Currently, however, we have very incomplete knowledge about all of the components that make up cells and how these components interact to perform those functions. Understanding how cells work is one of biology's grand challenges. If it were possible to understand more completely how at least some of the machinery of cells works, it might be possible to anticipate the onset and effects of disease and create therapies to ameliorate those effects. If it were possible to influence precisely the metabolic operations of cells, they might be usable as highly controllable factories for the production of a variety of useful organic compounds.

However, cell biology is awash in data on cellular components and their interactions. Although such data are necessary starting points for an understanding of cellular behavior that is sufficient for prediction, control, and redesign, making sense out of the data is difficult. For example, diagrams tracing all of the interactions, activities, locations, and expression times of the proteins, metabolites, and nucleic acids involved have become so dense with lines and annotations that reasoning about their functions has become almost impossible.

[2] Section 9.2 is based largely on A.P. Arkin, "Synthetic Cell Biology," *Current Opinion in Biotechnology* 12(6):638-644, 2001.

As noted in Section 5.4.2, cellular simulation efforts have for the most part addressed selected aspects of cellular functionality. The grand challenge of cellular modeling and simulation is a high-fidelity model of a cell that captures the interactions between the many different aspects of functionality, where "high fidelity" means the ability to make reasonably accurate and detailed predictions about all interesting cellular behavior under the various environmental circumstances encountered in its life cycle. Of course, a model perforce is an abstraction that omits certain aspects of the phenomenon it is representing. But the key term in this description is "interesting" behavior—behavior that is interesting to researchers. In this context, the model is intended to integrate—as a real cell would—different aspects of its functionality. Although the grand challenge may well be unachievable, almost by definition, the goal of increasing degrees of integration of what is known and understood about various aspects of cellular function remains something for which researchers strive.

The development of a high-fidelity simulation of a cell—even the simplest cell—is an enormous intellectual challenge. Indeed, even computational models that are very well developed, such as models of neural and cardiac electrophysiology, often fail miserably when they are exercised beyond the data that have been used to construct them. Yet if a truly high-fidelity simulation could be developed, the ability to predict cellular response across a wide range of environmental conditions *using a single model* would imply an understanding of cellular function far beyond what is available today, or even in the immediate future, and would be a tangible and crowning achievement in science. And, of course, the scientific journey to such an achievement would have many intermediate payoffs, in terms of tools and insights relevant to various aspects of cellular function. From a practical standpoint, such a simulation would be an invaluable aid to medicine and would provide a testbed for biological scientists and engineers to explore techniques of cellular control that might be exploited for human purposes.

An intermediate step toward the high-fidelity simulation of a real cell would be a model of a simple hypothetical cell endowed with specific properties of real cells. This model would necessarily include representations of several key elements (Box 9.2). The hundreds of molecules and hundreds of thousands of interactions required do not appear computationally daunting, until it is realized that the time scale of molecular interaction is on the order of femtoseconds, and interesting time scales of cellular response may well be hours or days.

The challenges fall into three general categories:

- *Mechanistic understanding*. High-fidelity simulations will require a much more profound physical understanding of basic biological entities at multiple levels of detail than is available today. (For example, it is not known how RNA polymerase actually moves along a DNA strand or what rates of binding or unbinding occur in vivo.) An understanding of how these entities interact inside the cell is equally critical. Mechanistic understanding would be greatly facilitated by the development of new mathematical formalisms that would enable the logical parsing of large networks into small modules

Box 9.2
Elements of a Hypothetical Cell

- An outside and inside separated by some coat or membrane (e.g., lipid)
- One or more internal compartments inside the cell
- Genes and an internal code for regulation of function
- An energy supply to keep the cell "alive" or "working"
- Reproductive capability
- At least hundreds of biologically significant molecules, with potentially hundreds of thousands of interactions between them
- Responsiveness to environmental conditions that affect the internal operation and behavior of the cell (e.g., changes in temperature, acidity, salinity)

whose behavior can be analyzed. Such modules would be building blocks that researchers could use to build functionality, understand controllable aspects, and identify points of failure.

- *Data acquisition*. Simulation models are data-intensive, and today there are relatively few systems with enough quality data to create highly detailed models of cellular function. It will be important to develop ways of measuring many more aspects of internal cellular state, and in particular, new techniques for measuring rates of processes and biochemical reactions in situ in living cells will be necessary. Besides additional reporter molecules, selective fluorescent dyes, and so on, a particular need is to develop good ways of tracking cellular state at different points in time, so that cellular dynamics can be better understood. Large volumes of data on reaction rates will also be necessary to model many cellular processes.
- *Integrative tools*. Because cellular function is so complex, researchers have used a variety of data collection techniques. Data from multiple sources—microarrays, protein mass spectroscopy, capillary and high-pressure chromatographies, high-end fluorescence microscopy, and so on—will have to be integrated—and are indeed required—if validated, high-fidelity cellular models are to be built. Moreover, because existing models and simulations relevant to a given cell span multiple levels of organizational hierarchy (temporal, spatial, etc.), tools are necessary to facilitate their integration. With such tools at the researcher's disposal, it will be possible to develop complex models rapidly, assembling molecular components into modules, linking modules, computing dynamic interactions, and comparing predictions to data.

Finally, despite the power of cellular modeling and simulation to advance understanding, models should not be regarded as an end product in and of themselves. Because all models are unfaithful to the phenomena they represent in some way, models should be regarded as tools to gain insight and to be used in continual refinement of our understanding, rather than as accurate representations of real systems, and model predictions should be taken as promising hypotheses that will require experimental validation if they are to be accepted as reliable.

The discussion above suggests that many researchers will have to collaborate in the search for an integrated understanding. Such coordinated marshaling of researchers and resources toward a shared goal is a common model for industry, but this multi-investigator approach is new for the academic environment. Large government-funded projects such as the Alliance for Cellular Signaling (discussed in Chapter 4) or private organizations like the Institute for Systems Biology[3] are the new great experiments in bringing a cooperative approach to academic biology.

Still more ambitious—probably by an order of magnitude or more—is the notion of simulating the behavior of a multicelled organism. For example, Harel proposes to develop a model of the *Caenorhabditis elegans* nematode, an organism that is well characterized with respect to its anatomy and genetics.[4] Harel describes the challenge as one of constructing "a full, true-to-all-known-facts, 4-dimensional, fully animated model of the development and behavior of this worm. . . , which is easily extendable as new biological facts are discovered."

In Harel's view, the feasibility of such a model is based on the notion that the complexity of biological systems stems from their high reactivity (i.e., they are highly concurrent and time-intensive, exhibit hybrid behavior that is predominantly discrete in nature but with continuous aspects as well, and consist of many interacting, often distributed, components). The structure of a reactive system may itself be dynamic, with its components being repeatedly created and destroyed during the system's life span. Harel notes:

[3] See http://www.systemsbiology.org/home.html.

[4] D. Harel, "A Grand Challenge for Computing: Towards Full Reactive Modeling of a Multi-Cellular Animal," *European Association for Theoretical Computer Science (EATCS) Bulletin*, 2003, available at http://www.wisdom.weizmann.ac.il/~dharel/papers/GrandChallenge.doc.

> [B]iological systems exhibit the characteristics of reactive systems remarkably, and on many levels; from the molecular, via the cellular, and all the way up to organs, full organisms, and even entire populations. It doesn't take much to observe within such systems the heavy concurrency, the event-driven discrete nature of the behavior, the chain-reactions and cause-effect phenomena, the time-dependent patterns, etc.

Harel concludes that biological systems can be modeled as reactive systems, using languages and tools developed by computer science for the construction of man-made reactive systems (briefly discussed in Section 5.3.4 and at greater length in the reference in footnote 4 of this chapter).

If the Harel effort is successful, a model of *C. elegans* would result that is fully executable, flexible, interactive, comprehensive, and comprehensible. By realistically simulating the worm's development and behavior, it would help researchers to uncover gaps, correct errors, suggest new experiments, predict unobserved phenomena, and answer questions that cannot be addressed by standard laboratory techniques alone. In addition, it would enable users to switch rapidly between levels of detail (from the entire macroscopic behavior of the worm to the cellular and perhaps molecular levels). Most importantly, the model would be extensible, allowing biologists to enter new data themselves as they are discovered and to test various hypotheses about aspects of behavior that are not yet known.

9.3 A SYNTHETIC CELL WITH PHYSICAL FORM

The most ambitious goal of synthetic biology (Section 8.4.2) is the biochemical instantiation of a real—if synthetic—cell with the capability to grow and reproduce. Such an achievement would necessarily be accompanied by new insights into the molecular dynamics of cells, the origins of life on Earth, and the limits of biological life. In practical terms, such cells could be engineered to perform specific functions, and thus could serve as a platform for innovative industrial and biomedical applications.

Cellular modification has a long history ranging from the development of plasmids carrying biosynthetic genes, or serving as "engineering blanks" for production of new materials, to the creation of small genetic circuits for the control of gene expression. However, the synthetic cells being imagined today would differ from the original cell much more substantially than those that have resulted from modifications to date. In principle, these cells need have no chemical or structural similarity to natural cells. Since they will be designed, not evolved, they may contain functions or structures unachievable through natural selection.

Synthetic cells are a potentially powerful therapeutic tool that may be able to deliver drugs to damaged tissue to seek and destroy foreign cells (in infections), destroy malignant cells (in cancer), remove obstructions (in cardiovascular disease), rebuild or correct defects (e.g., reattach severed nerves), or replace parts of tissue that was injured—and doing so without affecting nonproblematic tissues, while reducing the side effects of current conventional treatments.

The applications of synthetic cells undertaking cell-level process control computing are not limited to those of medicine and chemical sensing. There are also potential applications to the nanofabrication of new and useful materials and structures. Indeed, natural biology exhibits propulsive rotors and limbs at the microscale, and synthetic cells may be an enabling technology for nanofabrication—the building of structures at the microscopic level. There may be other techniques to accomplish this, but synthetic cells offer a promise of high efficiency through massively parallel reproduction. The gene regulatory networks incorporated into synthetic cells allow for the simultaneous creation of multiple oligonucleotide sequences in a programmable fashion. Conversely, self-assembled DNA nanostructures can potentially be used as control structures that interact with intracellular components and molecules. Such control could enable the engineering construction of complex extracellular structures and precise control of fabrication at the subnanometer level, which might in turn lead to the construction of complex molecular-scale electronic structures (Section 8.4.3.2) and the creation of new biological materials, much as natural biological materials result from natural biological processes.

Constructing these structures will require the ability to fabricate individual devices and the ability to assemble these devices into a working system, since it is likely to be very difficult to assemble a system directly from scratch. One approach to an assembly facility is to use a mostly passive scaffold, consisting of selectively self-assembling molecules that can be used to support the fabrication of molecular devices that are appropriately interconnected. Indeed, DNA molecules and their attendant enzymes are capable of self-assembly. By exploiting that capability, it has been possible to create a number of designed nanostructures, such as tiles and latticed sheets. Although the characteristics of these biomaterials need further exploration, postulated uses of them include as scaffolds (for example, for the crystallization of macromolecules); as photonic materials with novel properties; as designable zeolite-like materials for use as catalysts or as molecular sieves; and as platforms for the assembly of molecular electronic components or biochips.[5] Uses of DNA as a molecular "Lego" kit with which to design nanomachines, such as molecular tweezers and motors on runways, are also under investigation.

The relevance of synthetic cell engineering to nanofabrication is driven by the convergence of developments in several areas, including the miniaturization of biosensors and biochips into the nanometer-scale regime, the fabrication of nanoscale objects that can be placed in intracellular locations for monitoring and modifying cell function, the replacement of silicon devices with nanoscale, molecular-based computational systems, and the application of biopolymers in the formation of novel nanostructured materials with unique optical and selective transport properties. The highly predictable hybridization chemistry of DNA, the ability to completely control the length and content of oligonucleotides, and the wealth of enzymes available for modification of DNA make nucleic acids an attractive candidate for all of these applications.

Furthermore, by designing and implementing synthetic cells, a much better understanding will be gained of how real cells work, how they are regulated, and what limitations are inherent in their machinery. Here, the discovery process is iterative, in that our understanding and observations of living cells serve as "truthing" mechanisms to inform and validate or refute the experimental constructs of synthetic cells. In turn, the mechanisms underlying synthetic cells are likely to be more easily understood than comparable ones in natural cells. Using this combined information, the behavior of biological processes in living cells can slowly be unraveled. For such reasons, the process of creating synthetic cells will spin off benefits to biology and science, just as the Human Genome Project led to dramatic improvements in the technology of molecular biology.

To proceed with the creation of synthetic cells, three separate but interrelated problems must be addressed:

- The theoretical and quantitative problem of formulating, understanding, and perhaps even optimizing the design of a synthetic cell;
- The biological problem of applying lessons learned from real cells to such designs and using synthetic cells to inform our understanding of more complicated natural cells; and
- The engineering and chemistry problem of assembling the parts into a physical system (or to design self-assembling pieces).

One approach to building such a cell de novo is to start with a set of parts and assemble them into a functional biomolecular machine. Conceiving a cell de novo means that cellular components and their assembly are predetermined, and that the cell engineer has a quantifiable understanding of events and outcomes that can be used to predict the behavior of the components and their interactions at least probabilistically. A key aspect of de novo construction is that a de novo cellular design is not constrained by evolutionary history and hence is much more transparent than cells found in nature. Be-

[5] E. Winfree, F. Liu, L.A. Wenzler, and N.C. Seeman, "Design and Self-Assembly of Two-Dimensional DNA Crystals," *Nature* 394(6693):539-544, 1998.

cause an engineered cell would be designed by human beings, the functions of its various elements would be much better known. This fact implies that it would be easier to identify critical control points in the system and to understand the rules by which the system operates.

A second approach is to modify an existing living cell to give it new behaviors or to remove unwanted behaviors; classical metabolic engineering and natural product synthesis would be relevant to this approach. One starting point would be to use the membrane of an existing cell, but modification of these lipid bilayers to incorporate chemically inducible channels, integrated inorganic structures for sensing and catalysis, and other biopolymer structures for the identification and modification of biological substrates will provide a greater degree of freedom in the manipulation of the chemical state of the synthetic cell.

A third approach is to abandon DNA-based cells. Szostak et al.[6] argue that the "stripping-down" of a present-day bacterium to its minimum essential components still leaves hundreds of genes and thousands of different proteins and other molecules. They suggest that synthetic cells could use RNA as the repository of "genetic" information and as enzymes that catalyze metabolism. In their view, the most important requirements of a synthetic cell from a scientific standpoint are that it replicates autonomously and that it is subject to evolutionary forces. In this context, autonomous replication means continued growth and division that depends only on the input of small molecules and energy, not on the products of preexisting living systems such as protein enzymes. Evolution in this context means that the structure is capable of producing different phenotypes that are subject to forces of natural selection, although being subject to evolutionary forces has definite disadvantages from an engineering perspective seeking practical application of synthetic cells.

The elements of a synthetic cell are likely to mirror those of simulations (see Box 9.2), except of course that they will take physical representation. Inputs to the synthetic cell would take the form of environmental sensitivities of various kinds that direct cellular function. (Another perspective on "artificial cells" similar to this report's notion of synthetic cells is offered by Pohorille.[7] In general, synthetic cells share much with artificial cells, and the dividing line between them is both blurry and somewhat arbitrary. The modal use of the term "artificial cell" appears to refer to an entity with a liposome membrane, whose physical dimensions are comparable to those of natural cells, that serves a function such as enzyme delivery, drug delivery for cell therapy, and red blood cell substitutes.[8]) However, if synthetic cells are to be useful or controllable, it will be necessary to insert control points that can supply external instructions or "reprogram" the cell for specialized tasks (e.g., a virus that injects DNA into the cell to insert new pieces of code or instructions).

Researchers are interested in expanding the size and complexity of pathways for synthetic cells that will do more interesting things. But there is little low-hanging fruit in this area, and today's computational and mathematical ability to predict cellular behavior quantitatively is inadequate to do so, let alone to select for the desired behavior. To bring about the development of synthetic cells from concept to practical reality, numerous difficulties and obstacles must be overcome. Following is a list of major challenges that have to be addressed:

- *A framework for cellular simulation that can specify and model cellular function at different levels of abstraction* (as described in Section 9.2). Simulations will enable researchers to test their proposed designs, minimizing (though not eliminating) the need for in vivo construction and experimentation. Note that the availability of such a framework implies that the data used to support it are also available to assist in the engineering development of synthetic cells.

[6] J.W. Szostak, D.P. Bartel, and P.L. Luisi, "Synthesizing Life," *Nature* 409(6818):387-390, 2001.

[7] A. Pohorille, "Artificial Cells: Prospects for Biotechnology," *Trends in Biotechnology* 20(3):123-128, 2002.

[8] See, for example, T.M.S. Chang, "Artificial Cell Biotechnology for Medical Applications," *Blood Purification* 18:91-96, 2000, available at http://www.medicine.mcgill.ca/artcell/isbp.pdf.

Box 9.3
Tool Suites

One tool suite is a simulator and verifier for genetic digital circuits, called BioSPICE. The input to BioSPICE is the specification of a network of gene expression systems (including the relevant protein products) and a small layout of cells on some medium. The simulator computes the time-domain behavior of concentration of intracellular proteins and intercellular message-passing chemicals. (For more information, see http://www.biospice.org.)

A second tool would be a "plasmid compiler" that takes a logic diagram and constructs plasmids to implement the required logic in a way compatible with the metabolism of the target organism. Both the simulator and the compiler must incorporate a database of biochemical mechanisms, their reaction kinetics, their diffusion rates, and their interactions with other biological mechanisms.

- *Stability and robustness in the face of varying environmental conditions and noise.* For example, it is well known that nature provides a variety of redundant pathways for biological function, so that (for example) the incapacitation of one gene is often not unduly disruptive to the cell.
- *Improvement in the libraries of DNA-binding proteins and their matching repressor patterns.* These are at present inadequate, and good data about their kinetic constants are unavailable (hence signal transfer characteristics cannot be predicted). Any specific combination of proteins might well interact outside the genetic regulatory mechanisms involved, thus creating potentially undesirable side effects.
- *Control point design and insertion.*
- *Data measurement and acquisition.* To facilitate the monitoring of a synthetic cell's behavior, it is desirable to incorporate into the structure of the cell itself methods for measuring internal state parameters. Such measurements would be used to parameterize the functionality of cellular elements and compare performance to specifications.
- *Deeper understanding of biomolecular design rules.* Engineering of proteins for the modification of biointeractions will be required in all aspects of cell design, because it is relevant to membrane-based receptors, protein effectors, and transcriptional cofactors. Today, metabolic engineers are frequently frustrated in attempts to reengineer metabolic pathways for new functions because, at this point, the "design principles" of natural cells are largely unknown. To design, fabricate, and prototype cellular modules, it must be possible to engineer proteins that will bind to DNA and regulate gene expression. Current examples of DNA binding proteins are zinc fingers, response regulators, and homeodomains. The goal is to create flexible protein systems that can be modified to vary binding location and strength and, ultimately, to insert these modules into living cells to change their function.
- *A "device-packing" design framework that allows the rapid design and synthesis of new networks inside cells.* This framework would facilitate designs that allow the reuse of parts and the rapid modification of said parts for creating various "modules" (switches, ramps, filters, oscillators, etc.). The understanding available today regarding how cells reproduce and metabolize is not sufficient to enable the insertion of new mechanisms that interact with these functions in predictable and reliable ways.
- *Tool suites to support the design, analysis, and construction of biologic circuits.* Such suites are as yet unavailable (but see Box 9.3).

9.4 NEURAL INFORMATION PROCESSING AND NEURAL PROSTHETICS

Brain research is a grand challenge area for the coming decades. In essence, the goal of neuroscience research is to understand how the interplay of structural dynamics, biochemical processes, and electrical signals in nervous tissue gives rise to higher-order functions such as normal or abnormal thoughts, actions, memories, and behaviors. Experimental advances of the past decades have given the brain researcher an increasingly powerful arsenal of tools to obtain data—from the level of molecules to nervous systems—and to compare differences between individuals.

Today, neuroscientists have begun the arduous process of adapting and assembling neuroscience data at all scales of resolution and across disciplines into electronically accessible, distributed databases. These information repositories will complement the vast structural and sequence databases created to catalog, organize, and analyze gene sequences and protein products. Such databases have proven enormously useful in bioinformatics research; whether equal rewards will accrue from similar efforts for tissue-level data, whole-brain imaging, physiological data, and so forth remains to be seen, but based on the successes of the molecular informatics activities and the challenge questions of the neuroscientist, big payoffs can be anticipated.

At the very least, multiscale informatics efforts for brain research will provide organizing frameworks and computational tools to manage neuroscience data, from the lab notebook to published data. An ideal and expected outcome will be the provisioning for new opportunities to integrate large amounts of biological data into unified theories of function and aid in the discovery process.

To provide some perspective on the problem, consider that animal brains are the information-processing systems of nature. A honeybee's brain contains roughly 100 million synapses; a contemporary computer contains roughly 100 million transistors. Given a history of inputs, both systems choose from among a set of possible outputs. Yet although it is understood how a digital computer adds and subtracts numbers and stores error-free data, it is not understood how a honeybee learns to find nectar-rich flowers or to communicate with other honeybees.

We do not expect a honeybee to perform numerical computations; likewise, we do not expect a digital computer to learn autonomously, at least not today. However, an interesting question is the extent to which the structure of an information-processing system and the information representations that it uses predispose the system to certain types of computation. Put another way, in what ways and under what circumstances, if any, are neuronal circuits and neural information-processing systems inherently superior to von Neumann architectures and Shannon information representations for adaptation and learning? Given the desirability of computers that can learn and adapt, an ability to answer this question might provide some guidance in the engineering of such systems.

Some things are known about neural information processing:

- Animal brains find good solutions to real-time problems in image and speech processing, motor control, and learning. To perform these tasks, nervous systems must represent, store, and process information. However, it is highly unlikely that neural information is represented in digital form.
- It is likely that neurons are the nervous system's primary computing elements. A typical neuron is markedly unlike a typical logic gate; it possesses on average 10,000 synaptic inputs and a similar number of outputs.
- The stored memory of a neural information-processing system is contained in the pattern and strength of the analog synapses that connect it to other neurons. Nervous systems use vast numbers of synapses to effect their computations: in neocortical tissue, the synapse density is roughly 3×10^8 synapses per cubic millimeter.[9] Specific memories are also known not to be localized to particular neurons or sets of neurons in the brain.[10]

[9] R. Douglas, "Rules of Thumb for Neuronal Circuits in the Neocortex," *Notes for the Neuromorphic VLSI Workshop*, Telluride, CO, 1994.

[10] The essential reason is that specific memories are generally richly and densely connected to other memories, and hence can be reconstructed through that web of connections.

- The disparity between the information processing that can be done by digital computers and that done by nervous systems is likely to be a consequence of the different way in which nerve tissue represents and processes information, although this representation is not understood.
- At the device level, nervous tissue operates on physical principles that are similar to those that underlie semiconductor electronics.[11] Thus, differences between neural and silicon computation must be the result of differences in computational architecture and representation. It is thus the higher-level organization underlying neural computation that is of interest and relevance. Note also that for the purposes of understanding neural signaling or computation, a neuron-by-neuron simulation of nervous tissue per se cannot be expected to reveal very much about the principles of organization, though it may be necessary for the development of useful artifacts (e.g., neural prostheses).

Some of the principles underlying neural computation are understood. For example, neurobiology uses continuous adaptation rather than absolute precision in responding to analog inputs. The dynamic range of the human visual system is roughly 10 decades in input light intensity—about 32 bits. But biology doesn't process visual signals with 32-bit precision; rather, it uses a 7- or 8-bit instantaneous dynamic range and adapts the visual pathway's operating point based on the background light intensity. Although this approach is similar to the automatic gain control used in electronic amplifiers, biology takes the paradigm much farther: adaptation pervades every level of the visual system, rather than being concentrated just at the front end.[12]

There are essentially two complementary approaches toward gaining a greater understanding of neural information processing. One approach is to reproduce physiological phenomena to increase our understanding of the nervous system.[13] A second approach is based on using a manageable subset of neural properties to investigate emergent behavior in networks of neuron-like elements.[14] Those favoring the first approach believe that these details are crucial to understanding the collective behavior of the network and are developing probes that are increasingly able to include the relevant physiology. Those favoring the second approach make the implicit assumption that reproducing many neurophysiological details is secondary to understanding the collective behavior of nervous tissue, even while acknowledging that only a detailed physiological investigation can reveal definitively whether the details are in fact relevant.

What can be accomplished by building silicon circuits modeled after biology? First, once the neuronal primitives are known, it will be possible to map them onto silicon. Once it is understood how biological systems compute with these primitives, biologically based silicon computing will be possible. Second, we can investigate how physical and technological limits, such as wire density and signal delays and noise, constrain neuronal computation. Third, we can learn about alternative models of computation. Biology demonstrates nondigital computing machines that are incredibly space- and energy-efficient and that find adequate solutions to ill-posed problems naturally.

[11] In both integrated circuits and nervous tissue, information is manipulated principally on the basis of charge conservation. In the former, electrons are in thermal equilibrium with their surroundings and their energies are Boltzmann distributed. In the latter, ions are in thermal equilibrium with their surroundings and their energies also are Boltzmann distributed. In semiconductor electronics, energy barriers are used to contain the electronic charge, by using the work function difference between silicon and silicon dioxide or the energy barrier in a *pn* junction. In nervous tissue, energy barriers are also erected to contain the ionic charge, by using lipid membranes in an aqueous solution. In both systems, when the height of the energy barrier is modulated, the resulting current flow is an exponential function of the applied voltage, thus allowing devices that exhibit signal gain. Transistors use populations of electrons to change their channel conductance, in much the same way that neurons use populations of ionic channels to change their membrane conductance.

[12] Adaptation helps to explain why some biological neural systems never settle down—they can be built so that when faced with unchanging inputs, the inputs are adapted away. This phenomenon helps to explain many visual aftereffects. A stabilized image on the retina disappears after a minute or so, and the whole visual field appears gray.

[13] M.A. Mahowald and R.J. Douglas, "A Silicon Neuron," *Nature* 354(6354):515-518, 1991.

[14] J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Reading, MA, 1991.

The challenges of neural information processing fall into two primary categories: the semantics of neural signaling and the development of neural prostheses. Signaling is the first challenge. It is known that the spike trains of neurons carry information in some way—neurons that cannot "fire" are essentially dead.[15] Also, the physical phenomena that constitute "firing" are known—electrical spikes of varying amplitude and timing. However, the connections among these patterns of signaling in multiple neurons to memories of specific events, motor control of muscles, sensory perception, or mental computation are entirely unknown. How do neurons integrate data from large numbers of multimodal sensors? How do they deal with data overload? How do they decide a behavioral response from multiple alternatives under severe and ill-posed constraints?

Today's neural instrumentation (e.g., positron emission tomography [PET] scans, functional magnetic resonance imaging [fMRI]) can identify areas of the brain that are active under various circumstances, but since the spatial resolution of these probes is wholly inadequate to resolve individual neuronal activity,[16] such instrumentation can provide only the roughest guidance about where researchers need to look for more information about neuronal signaling, rather than anything specific about that information itself. The primary challenge in this domain is the development of a formalism for neuronal signaling (most likely a time-dependent one that takes kinetics into account), much like the Boolean algebra that provides a computational formalism based on binary logic levels in the digital domain.

A step toward a complete molecular model of neurotransmission for an entire cell is provided by MCell, briefly mentioned in Chapter 5. MCell is a simulation program that can model single synapses and groups of synapses. To date, it been used to understand one aspect of biological signal transduction, namely the microphysiology of synaptic transmission. MCell simulations provide insights into the behavior and variability of real systems comprising finite numbers of molecules interacting in spatially complex environments. MCell incorporates high-resolution physical structure into models of ligand diffusion and signaling, and thus can take into account the large complexity and diversity of neural tissue at the subcellular level. It models the diffusion of individual ligand molecules used in neural signaling using a Brownian dynamics random walk algorithm, and bulk solution rate constants are converted into Monte Carlo probabilities so that the diffusing ligands can undergo stochastic chemical interactions with individual binding sites, such as receptor proteins, enzymes, and transporters.[17]

The second challenge is that of neural prosthetics. A neural prosthesis is a device that interfaces directly with neurons, receiving and transmitting signals that affect the function and activity of those neurons, and that behaves in predictable and useful ways. Perhaps the "simplest" neural prosthesis is an artificial implant that can seamlessly replace nonfunctioning nerve tissue.

Today, some measure of cognitive control of artificial limbs can be achieved through bionic brain-machine or peripheral-machine interfaces. William Craelius et al.[18] have designed a prosthetic hand that offers amputees control of finger flexion using natural motor pathways, enabling them to undertake slow typing and piano playing. The prosthetic hand is based on the use of natural tendon movements in the forearm to actuate virtual finger movement. A volitional tendon movement within the residual limb causes a slight displacement of air in foam sensors attached to the skin in that location, and the resulting pressure differential is used to control a multifinger hand.

[15] It is also known that not all neural signaling is carried by spikes. A phenomenon known as graded synaptic transmission also carries neural information and is based on a release of neurotransmitter at synaptic junctions whose volume is voltage dependent and continuous. Graded synaptic transmission appears to be much more common in invertebrates and sometimes exists alongside spike-mediated signaling (as in the case of lobsters). The bandwidth of this analog channel is as much as five times the highest rates measured in spiking neurons (see, for example, R.R. de Ruyter van Steveninck and S.B. Laughlin, "The Rate of Information Transfer at Graded-Potential Synapses," *Nature* 379:642-645, 1996), but the analog channel is likely to suffer a much higher susceptibility to noise than do spike-mediated communications.

[16] The spatial resolution of neural instrumentation is on the order of 1 to 10 mm. See D. Purves et al., *Neuroscience*, Sinauer Associates Inc., Sunderland, MA, 1997. Given about 3×10^8 synapses per cubic millimeter, not much localization is possible.

[17] See http://www.mcell.cnl.salk.edu/.

[18] W. Craelius, R.L. Abboudi, and N.A. Newby, "Control of a Multi-finger Prosthetic Hand," *ICORR '99: International Conference on Rehabilitation Robotics*, Stanford, CA, 1999.

A second example of a neural prosthesis is a retinal prosthesis intended to provide functionality when the retina of the eye is nonfunctional. In one variant, a light-sensitive microchip is implanted into the back of the eye. Light striking the microchip (which has thousands of individual sensors) generates electrical signals that travel through the optic nerve to the brain and are interpreted as an image.[19] In another variant, the retina is bypassed entirely through the use of a camera mounted on a pair of eyeglasses to capture and transmit a light image via a radio signal to a chip implanted near the ganglion cells, which send nerve impulses to the brain.[20] In a third variant, an implanted microfluidic chip that controls the flow of neurotransmitters translates digital images into neurochemical signals that provide meaningful visual information to the brain. The microfluidic chip has a two-dimensional array of small controllable pores, corresponding to pixels in an image. An image is created by the selective drip of neurotransmitters onto specific bipolar cells, which are the cells that carry retinal information to the brain.[21]

A third example of work in this area is that of Musallam et al., who have demonstrated the feasibility of a neural interface that enables a monkey to control the movement of a cursor on a computer screen by thinking about a goal the monkey would like to achieve and assigning a value to that goal.[22] The interesting twist to this work is the reliance of signals from parts of the brain related to higher-order ("cognitive") brain functions for movement planning for the control of a prosthetic device. (Previous studies have relied on lower-level signals from the motor cortex.[23])

The advantage of using higher-level cognitive signals is that they capture information about the monkey's goal (moving the cursor) and preferences (the destination on the screen the monkey wants). Musallam et al. point out that once the signals associated with the subject's goals are decoded, a smart external device can perform the lower-level computations necessary to achieve the goals. For example, a smart robotic arm would be able to understand what the intended goal of an arm movement is and then compute—on its own—the trajectory needed to move the arm to that position. Furthermore, the abstract nature of a cognitive command would allow it to be used for the control and operation of a number of different devices. If higher-level signals associated with speech or emotion could be decoded, it would become possible to record thoughts from speech areas (reducing the need for the use of cumbersome letter boards and time-consuming spelling programs) or to provide online indications of a patient's emotional state.

A fourth example is provided by Theodore Berger of the University of Southern California, who is attempting to develop an artificial hippocampus—a silicon implant that will behave neuronally in a manner identical to the brain tissue that it replaces.[24] The hippocampus is the part of the brain responsible for encoding experiences so that they can be stored as long-term memories elsewhere in the brain; without the hippocampus, a person is unable to store new memories but can recall ones stored prior to its loss. Because the manner in which the hippocampus stores information is unknown, Berger's approach is based on designing a chip that can provide the identical input-output response. The input-

[19] N.S. Peachey and A.Y. Chow, "Subretinal Implantation of Semiconductor-based Photodiodes: Progress and Challenges," *Journal of Rehabilitation Research and Development* 36(4):371-376, 1999.

[20] W. Liu, E. McGucken, M. Clements, S.C. DeMarco, K. Vichienchom, C. Hughes, et al., "Multiple-Unit Artificial Retina Chipset System to Benefit the Visually Impaired," to be published in *IEEE Transactions on Rehabilitation Engineering*. Available at http://www.icat.ncsu.edu/projects/retina/files/MARC_system_paper.pdf.

[21] B. Vastag, "Future Eye Implants Focus on Neurotransmitters," *Journal of the American Medical Association* 288(15):1833-1834, 2002.

[22] S. Musallam, B.D. Corneil, B. Greger, H. Scherberger, and R.A. Andersen, "Cognitive Control Signals for Neural Prosthetics," *Science* 305(5681):258-262, 2004. A Caltech press release of July 8, 2004, available at http://pr.caltech.edu/media/Press_Releases/PR12553.html, describes this work in more popular terms.

[23] J. Wessberg, C.R. Stambaugh, J.D. Kralik, P.D. Beck, M. Laubach, J.K. Chapin, J. Kim, S.J. Biggs, M.A. Srinivasan, and M.A.L. Nicolelis, "Real-Time Prediction of Hand Trajectory by Ensembles of Cortical Neurons in Primates," *Nature* 408(6810):361-365, 2000. Similar work on rats is described in J.K. Chapin, K.A. Moxon, R.S. Markowitz, and M.A.L. Nicolelis, "Real-Time Control of a Robot Arm Using Simultaneously Recorded Neurons in the Motor Cortex," *Nature Neuroscience* 2(7):664-670, 1999.

[24] R. Merritt, "Nerves of Silicon: Neural Chips Eyed for Brain Repair," *EE Times*, March 17, 2003 (10:37 a.m. EST), available at http://www.eetimes.com/story/OEG20030317S0013.

output response of a hippocampal slice was determined by stimulating it with a random-signal generator, and a mathematical model was developed to account for its response to these different stimuli. This model is then the basis for the chip circuitry.

By December 2003, Berger and his colleagues had completed the first test of using a microchip model to replace a portion of the hippocampal circuitry contained in a specific hippocampal brain slice. In that slice is the major intrinsic circuitry of the hippocampus that consists of three major cell fields, designated A, B, and C. Field A projects to and excites field B, which projects to and excites field C. Berger et al. developed a predictive mathematical model of the signal transformations that field B performs on the input signals that come from field A, and that field B then projects onto field C, and implemented the model in a field-programmable gate array (FPGA) for field B. When field B was surgically removed and the FPGA model of B was substituted, the result was that the output from area C of the hippocampal slice remained unchanged in all meaningful respects. Next steps beyond this work (e.g., developing circuitry that is less sensitive to the details of slice preparation, understanding the hardware in terms of meaningful abstractions) remain to be realized.

One result of such work may be the creation of building blocks that can be used to calculate universal mathematical functions and ultimately be the basis of families of devices for neural pattern matching. Such building blocks may also serve as a point of departure for understanding neural functions at a higher level of abstraction than is possible today.

An analogy might be drawn to finding a mathematical representation of a particular dataset. The approach of mapping an exhaustive input-output response is similar to a curve-fitting process that generates a function capable of reproducing the dataset perfectly. Knowledge of such a function does not necessarily entail any understanding of the casual mechanisms underlying that dataset; thus, a function resulting from a curve-fitting process is highly unlikely to be able to account for new data. Still, developing such a function may be the first step toward such understanding.

As suggested above, building a successful neural prosthetic implies some understanding of the semantics of neural information processing: how the relevant nerve tissue stores and replicates and processes information. However, it also requires a well-understood interface between a biological organism (e.g., a person) and the engineered device.

One of the primary challenges in the area of neural interface design is the physical connection of neurons to a chip—the right neurons must make connection with the right electrodes. The body's natural response to an electrode implanted in living tissue is to wall it off with glial cells that prevent neuron and electrode from making contact. One approach to solving this problem is to coat the electrode with a substance that does not trigger the glial reaction. Another is to rely on the neural tissue to reconfigure itself. Based on the knowledge that auditory nerves can reconfigure themselves to accommodate the signals emitted by cochlear implants, it may be possible to send out a signal that attracts the right nerves to the right contacts.

Prosthetic devices that restore or augment human physical abilities are increasingly sophisticated, and follow-on work will focus on enabling control of more complex actions by robotic arms and other devices. On the other hand, although some early work on prostheses that help to replace cognitive abilities has been successful, prostheses that improve cognitive abilities, by enhancing perception (superhuman sense) and decision-making (superhuman computation or knowledge) capabilities, must at present be regarded as being on the distant horizon.

9.5 EVOLUTIONARY BIOLOGY[25]

Although the basic principles of evolution (natural selection and mutation) are understood in the large, both population genetics and phylogenetics have been radically transformed by the recent avail-

[25] Section 9.5 is adapted largely from the Web page of John Huelsenbeck, University of California, San Diego, http://biology.ucsd.edu/faculty/huelsenbeck.html.

ability of large quantities of molecular data. For example, in population genetics (the study of mutations in populations), more molecular variability was found in the 1960s than had been expected, and this finding stimulated Kimura's neutral theory of molecular evolution.[26] Phylogenetics (the study of the evolutionary history of life) makes use of a variety of different kinds of data, of which DNA sequences are the most important, as well as whole-genome, metabolic, morphological, geographical, and geological data.[27]

Evolutionary biology is founded on the concept that organisms share a common origin and have diverged through time. The details and timing of these divergences—that is, the estimation or reconstruction of an evolutionary history—are important for both intellectual and practical reasons, and phylogenies are central to virtually all comparisons among species. From a practical standpoint, phylogenetics has helped to trace routes of infectious disease transmission (e.g., dental transmission of AIDS/HIV) and to identify new pathogens such as the New Mexico hantavirus. Moret (footnote 27) notes that phylogenetic analysis is useful in elucidating functional relationships within living cells, making functional predictions from sequence data banks of gene families, predicting ligands, developing vaccines, antimicrobials, and herbicides, and inferring secondary structure of RNAs. A clear picture of how life evolved from its humble origins to its present diversity would answer the age-old question, Where do we come from?

There are many interesting phylogenetic problems. For example, consider the problem of estimating large phylogenies, which is a central challenge in evolutionary biology. Given three species, there are only three possible trees that could represent their phylogenetic history: (A,(B,C)); (B,(A,C)); and (C,(A,B)). (The notation (A,(B,C)) means that B and C share a common ancestor, who itself shares a different common ancestor with A. Thus, even if one picks a tree at random, there is a one in three chance that the tree chosen will be correct. But the number of possible trees grows very rapidly with the number of species involved. For a "small" phylogenetic problem involving 10 species, there are 34,459,425 possible trees. For a problem involving 22 species, the number of trees exceeds 10^{23}. Today, most phylogenetic problems involve more than 80 species and some data sets contain more than 500 species. (For 500 species, there are approximately 1.0085×10^{1280} possible trees, only one of which can be correct.) Of course, the grandest of all challenges in this area is the construction of the entire phylogeny of all organisms on the planet—the complete "Tree of Life" involving some 10^7 to 10^8 species.

Given the existence of such large state spaces, it is clear that exhaustive search for the single correct phylogenetic tree is not a feasible strategy, regardless of how fast computers become in the foreseeable future. Researchers have developed a number of methods for coping with the size of these problems, but many of these methods have serious deficiencies. For example, the optimality criteria used by these methods often have dubious statistical justifications. In addition, many of these methods are simply stepwise addition algorithms and make no effort to explore the space of trees. Methods with the best statistical justification, such as maximum likelihood and Bayesian inference, are also the most difficult to implement for large problems.

Thus, the algorithmics of evolutionary biology are a fertile area for research. Moret (footnote 27) notes that reconstruction of the Tree of Life will require either the scaling-up of existing reconstruction methods or the development of entirely new ones. He notes that sequence-based reconstruction methodologies are available that are likely to scale effectively from 15,000 to 100,000 taxa, but that these methodologies are not likely to scale to millions of taxa. Moret also points out that the use of gene-order data (i.e., lists of genes in the order in which they occur along one or more chromosomes) can circumvent many of the difficulties associated with using sequence data. On the other hand, there are relatively

[26] M. Kimura, "Evolutionary Rate at the Molecular Level," *Nature* 217(129):624-626, 1968; Motoo Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, MA, 1983.

[27] B.M.E. Moret, "Computational Challenges from the Tree of Life," *Proceedings of the 7th Workshop on Algorithm Engineering and Experiments*, ALENEX '05, Vancouver, SIAM Press, Philadelphia, PA, 2005. This paper presents a number of computational challenges in evolutionary biology, of which only a few are mentioned in the subsequent discussion in this section.

few whole-genome data today, few models for the evolution of gene content and gene order, and a far greater complexity of the mathematics for gene orders compared to that for DNA sequences.

A related problem is that of comparing one or more features across species. The comparative method has provided much of the evidence for natural selection and is probably the most widely used statistical method in evolutionary biology. But comparative analyses must account for phylogenetic history, since the similarity in features common to multiple species that originate in a common evolutionary history can inappropriately and seriously bias the analyses. A number of methods have been developed to accommodate phylogenies in comparative analyses, but most of these methods assume that the phylogeny is known without error. However, this is patently unrealistic, because almost all phylogenies have a large degree of uncertainty. An important question is therefore to understand how comparative analyses can be performed that accommodate phylogenetic history without depending on any single phylogeny being correct.

Still another interesting problem concerns the genetics of adaptation—the genomic changes that occur when an organism adapts to a new set of selection pressures in a new environment. Because the process of adaptive change is difficult to study directly, there are many important and unanswered questions regarding the genetics of adaptation. For example, how many mutations are involved in a given adaptive change? Does this figure change when different organisms or different environments are involved? What is the distribution of fitness effects implied by these genetic changes during a bout of adaptation? How and to what extent are adaptations constrained by phylogenetic history? To what extent are specific genetic changes inevitable given a change of selection pressures?

9.6 COMPUTATIONAL ECOLOGY[28]

The long-term scientific goal of computational ecology is the development of methods to predict the response of ecosystems to changes in their physical, biological, and chemical components. Computational ecology seeks to combine realistic models of ecological systems with the often large datasets available to aid in analyzing these systems, utilizing techniques of modern computational science to manage the data, visualize model behavior, and statistically examine the complex dynamics that arise.[29] Questions raised immediately by computational ecology have a direct bearing on issues of important policy significance today—potential losses of biodiversity, achievement of sustainable futures, and impact of global change on local communities.[30]

The scientific questions to be addressed by computational ecology have both theoretical and applied significance. These questions include the following:[31]

- How are communities organized in space and time?
- What factors maintain or reduce biodiversity?
- What are the implications for ecosystem function?
- How should biodiversity be measured?
- How is ecological robustness maintained?

Consider, for example, ecological robustness. In ecological communities, many of the salient features remain unchanged, despite the fact that the identities of the relevant actors are continually in flux.

[28] Much of the discussion in this section is based on J. Helly, T. Case, F. Davis, S. Levin, and W. Michener, eds., *The State of Computational Ecology*, National Center for Ecological Analysis and Synthesis, Santa Barbara, CA, 1995, available at http://www.sdsc.edu/compeco_workshop/report/report.html.

[29] J. Helly et al., eds., *The State of Computational Ecology*, National Center for Ecological Analysis and Synthesis, Santa Barbara, CA, 1995, available at http://www.sdsc.edu/compeco_workshop/report/report.html.

[30] J. Lubchenco et al., "The Sustainable Biosphere Initiative: An Ecological Research Agenda," *Ecology* 72(2):371-412, 1991.

[31] Much of this list is taken from Helly et al., *The State of Computational Ecology*, 1995.

Species richness, species abundance relations, and biogeochemical cycles exhibit remarkable regularity, despite changes at lower levels of organization. In marine systems, the Redfield ratios,[32] which characterize the mean stoichiometry of plankton and of the water column, summarize the great constancy seen in the concentration ratios of carbon, nitrogen, and phosphorus relative to each other, although absolute levels vary considerably across the oceans. Similarly, Sheldon et al.[33] observed that the size spectrum, from the smallest particles to large fish, follows a power law with a characteristic exponent, valid across a range of trophic levels.

Ecosystems and the biosphere are complex adaptive systems,[34] in which macroscopic patterns emerge from interactions at lower levels of organization and feed back to influence dynamics on those scales. Although macroscopic investigations, such as those of Carlson and Doyle,[35] can shed considerable light on designed or managed systems, or on organ systems that have been the direct products of evolution, they provide at best a benchmark for comparisons for complex adaptive systems in which selection acts well below the level of the whole system.

The robustness of complex adaptive systems is dependent upon the same suite of characteristics that govern the robustness of any system—heterogeneity and diversity, redundancy and degeneracy, modularity, and the tightness of feedback loops. Heterogeneity, for example, provides the adaptive capacity that allows a system to persist in a changing environment; indeed, the robustness of the macroscopic features of such systems may arise despite, in fact even because of, the lack of robustness of their components. Yet these systems are neither designed nor selected for their macroscopic features. How different then are such systems from those in which the level of selection is the whole system? Should robustness be expected to emerge from the bottom up, and how does this self-organized robustness differ from what would be optimal for the robustness of systems as a whole?

Given that selection is most effective at much lower levels of organization, it is unclear what sustains ecological robustness at the macroscopic level. A key problem is to understand the properties of such self-organized, complex adaptive systems—to develop theories that facilitate scaling from individuals to whole systems and relating structure to function in order to identify signals warning of collapse. What are the consequences of the erosion of biodiversity, the homogenization of systems, and the breakdown of ecological barriers? How, indeed, will such changes affect the spread of disturbances, from forest fires to novel infectious diseases? Addressing these questions will require iterative integration of computational approaches with explorations into large-scale stochastic and distributed dynamical systems, with the goal of developing more parsimonious descriptors of essential aspects.

General theory concerning the robustness of complex systems focuses on a few key features: heterogeneity and diversity, redundancy and degeneracy, modularity, and the tightness of feedback loops.[36] Robustness is a design objective for most engineering applications, and investigations such as those of Carlson and Doyle have demonstrated how one might select on complex systems as a whole to achieve tolerance to particular classes of perturbations. One general principle that emerges from such studies is that there are trade-offs between robustness on diverse scales. Systems in general may be characterized as "robust, yet fragile." That is, their robustness to one class of perturbations, or on one scale, may

[32] A.C. Redfield, "On the Proportions of Organic Derivatives in Sea Water and Their Relation to the Composition of Plankton," pp. 176-192 in *James Johnstone Memorial Volume*, R.J. Daniel, ed., University Press of Liverpool, Liverpool, UK, 1934.

[33] R.W. Sheldon and T.R. Parsons, "A Continuous Size Spectrum for Particulate Matter in the Sea," *Journal of the Fisheries Research Board of Canada* 24:909-915, 1967; R.W. Sheldon, A. Prakash, and W.H. Sutcliffe, Jr., "The Size Distribution of Particles in the Ocean," *Limnological Oceanography* 17:327-340, 1972.

[34] S.A. Levin, *Fragile Dominion: Complexity and the Commons*, Perseus Books, Reading, MA, 1999; S.A. Levin, "Complex Adaptive System: Exploring the Known, the Unknown and the Unknowable," *Bulletin of the American Mathematical Society* 40:3-19, 2003.

[35] J.M. Carlson and J. Doyle, "Highly Optimized Tolerance: Robustness and Design in Complex Systems," *Physical Review Letters* 84(11):2529-2532, 2000.

[36] S.A. Levin, *Fragile Dominion: Complexity and the Commons*, Perseus Books, Reading, MA, 1999; S.A. Levin, "Complex Adaptive Systems; Exploring the Known, the Unknown and the Unknowable," *Bulletin of the American Mathematical Society* 40:3-19, 2003.

necessarily lead to fragility to other classes of perturbations, or on other scales. Understanding such trade-offs is one dimension of considerable intellectual challenge and problem richness.

These general points are instantiated in many different problem areas. Two illustrative areas—each important in its own right—include the dynamics of infectious diseases and the dynamics of marine microbial systems. In the first case, increased computational resources have fostered the development of models that relate individual behaviors to the spread of novel diseases, including smallpox and new strains and subtypes of influenza. Such models have been given added stimulus by concerns about the introduction and spread of infectious agents as weapons of bioterror, but the potential for new pandemics of influenza and other infectious diseases is probably a greater motivation for their development.

Marine microbial systems represent a vast and important storehouse of biodiversity, about which much too little is known. Recent efforts, stimulated by the success of genomics, have directed attention to characterizing the massive genetic diversity found in these systems. The computational challenges are substantial, even to catalog the vast array of data being collected. Yet just as sequencing efforts in genomics have highlighted the importance of knowing what the catalog of genetic detail reveals about how systems function in their ecological environments, the mass of accumulating information about marine microbial diversity spurs efforts at understanding how those marine ecosystems are organized and what maintains the robustness of features such as microbial diversity.

To address the scientific questions described above, researchers need techniques for dealing with systems across scales of space, time, and organizational complexity. Ultimately, an essential enabling tool will be a statistical mechanics of heterogeneous and nonindependent entities, in which the components of a system of interest are continually changing through processes of mutation and other forms of change.[37] Such a system differs dramatically from systems that have traditionally been analyzed through the machinery of traditional statistical mechanics (e.g., systems composed of identical, independently moving particles), and analytical methods for dealing with heterogeneous, nonindependent entities are generally very sophisticated. In general, such methods rely on the ability to capture the heterogeneity of the distribution (e.g., of traits) in terms of a small number of moments or other descriptors or rely on "equation-free" approaches[38] that finesse the need for explicit closures. In the absence of such an analytical characterization, computation is generally the only alternative to gaining insights about ensemble behavior, although computation may often provide analytical insights (and vice versa).

Today, computational ecology makes use of continuum and individual descriptions. Continuum modeling focuses on the impact on local ecological communities of large-scale (global) influences such as climate and fluxes of key elements such as carbon and nitrogen. These models are typically characterized by parameterized partial differential equations that represent appropriately averaged continuum quantities of ecological significance (e.g., density of a species). A central intellectual challenge of the top-down approach is reconciling the hundred-kilometer resolution of models that predict global climate change and elemental fluxes with the meter and centimeter scales of interest in natural and managed ecosystems.

The ab initio formulation of realistic continuum models is difficult, because the details of the underlying populations and entities matter a great deal. For example, naïve assumptions of independence, random motion, zero mixing time, or infinite propagation speed, which are often used in the ab initio formulation of continuum models, simply do not hold at the underlying individual level.[39] Accordingly, great care must be taken to derive a continuum description from knowledge of the individual elements in play.

[37] S. Levin, *Mathematics and Biology: The Interface,* Lawrence Berkeley Laboratory Pub-701, Berkeley, CA, 1992, available at http://www.bio.vu.nl/nvtb/Interface.html.

[38] C. Theodoropolous, Y. Quan, and I.G. Kevrekidis, "Coarse Stability and Bifurcation Analysis Using Time-Steppers: A Reaction-Diffusion Example," *Proceedings of the National Academy of Sciences* 97(18):9840-9843, 2000.

[39] S.A. Levin, "Complex Adaptive Systems: Exploring the Known, the Unknown and the Unknowable," *Bulletin (New Series) of the American Mathematical Society* 40(1):3-19, 2002.

Individual-based modeling seeks to extrapolate from the level of effects on individual plants and animals to changes in community-level patterns, which are necessarily characterized by longer time scales and broader space scales than those of individuals. Individual-based models, an ecological form of agent-based models, are rule-based approaches that can track the growth, movement, and reproduction of many thousands of individuals across the landscape[40] and, in looking at the global consequences of local interactions of individuals, are particularly well suited to address questions that relate to spatial heterogeneities (e.g., ecological sanctuaries).

In individual-based models, the inherent parallelism of ecological systems—that organisms interact concurrently across space—is manifest.[41] (By contrast, the parallelism in many computational models of other biological systems such as genomes and proteins is primarily a speedup mechanism for computation-intensive problems.) Individual-based models have been used to represent populations of predators, trees, and endangered species, and they are very useful in understanding the detailed response of the population of interest to alternative environmental circumstances.

In general, individual-based models are powerful tools for investigating systems that are analytically intractable, and they provide opportunities for the consideration of various scenarios and for exploring ecosystem management protocols that would not otherwise be possible. Nevertheless, such simulations often contain too many degrees of freedom to allow robust prediction. Thus, efforts to develop macroscopic representations that reduce dimensionality and that suppress irrelevant detail are essential—a point that reinforces the desirability of developing an appropriate statistical mechanics as described above.

Individual-based modeling is generally computation-intensive, for two reasons. The first is that a multitude of individuals must be represented, the behavior of each must be computed, and the entire ecosystem being modeled must be time-stepped at appropriately fine intervals. The second is that realism demands a certain amount of stochasticity; thus, an ensemble of simulations must be run in order to understand how changes in environmental and other parameters affect predicted outcomes. Grid implementations, taking advantage of the inherent parallelism of ecosystems, are one recent effort to advance individual-based modeling. The development of algorithms implementing parallelization for individual-based ecological models has enabled a number of simulations, including simulations for fish populations in the Everglades[42] and for more general models aimed ultimately at resource management.[43]

Data issues in computational ecology are also critical. Information technology has been a key enabler for a great deal of ecological data. For example, high-resolution multispectral images captured by satellites provide a wealth of information about ecosystems, resulting in maps that can depict how ecologically significant quantities can vary across large areas. While such images cannot yield significant information on the behavior of individuals, modern telemetry can be used to follow the movements of many individual organisms, a method applied routinely for certain endangered and threatened species.

At the same time, much remains to be done. Ground-based sensors take data only in their immediate locality. Thus, the spatial resolution provided by such sensors is a direct function of their areal density. Therefore, the advent of inexpensive networked sensors, described in Chapter 7, is potentially the harbinger of a new explosion of ecological data. For example, a survey of thirty papers chosen randomly from the journal *Ecology* illustrates that most ecological sampling is conducted with measurements being taken in small areas or at low frequency (often including one-time sampling).[44] Wireless

[40] See D.L. DeAngelis and L.J. Gross, eds., *Individual-Based Models and Approaches in Ecology*, Routledge, Chapman and Hall, New York, 1992.

[41] J. Haefner, "Parallel Computers and Individual-Based Models: An Overview," pp. 126-164 in D. DeAngelis and L. Gross, eds., *Individual-Based Models and Approaches in Ecology*, Chapman and Hall, New York, 1992.

[42] D. Wang, M.W. Berry, E.A. Carr, and L.J. Gross, "A Parallel Landscape Model for Fish as Part of a Multi-Scale Ecological System," available at http://www.tiem.utk.edu/gem/papers/dalipaper.pdf.

[43] D. Wang, E.A. Carr, M.R. Palmer, M.W. Berry, and L.J. Gross, "A Grid Service Module for Natural-Resource Managers," *IEEE Internet Computing* 9(1):35-41, 2005, available at http://www.tiem.utk.edu/gem/papers/gridservice.pdf.

[44] J. Porter et al., "Wireless Sensor Networks for Ecology," *Biosciences*, 2005, in press.

sensor networks can fill a gap in our current capabilities by enabling researchers to sample at finer spatial scales or faster rates not currently possible. It is this range of space-time (widely distributed spatial sensing with high temporal frequency) that will be critical to address the grand challenges of the environmental sciences (biogeochemical cycles, biological diversity and ecosystem functioning, climate variability, hydrologic forecasting, infectious disease and the environment, institutions and resource use, land-use dynamics, reinventing the use of materials) proposed by the National Research Council.[45] Similarly, an explosion of data and of information will arise from sensors carried by individual animals. The extent of information potentially provided by continuous monitoring of position and physiological data, compared to tags and radio collars, is obvious.

Note also an important synergy between modeling and the use of sensor networks. The effective use of sensor networks relies on modeling and analytical work to guide the placement of sensors. In turn, sensor data provide data to models that allow for prediction and interpretation of models, to understand the underlying processes. In this sense, models are the basis for an adaptive sampling scheme for sensor use.

Another data issue is progress in capturing specimen data in electronic form. Over the years, hundreds of millions of specimens have been recorded in museum records. While the information in extant collections could provide numerous opportunities for modeling and increased understanding, very few records are in electronic form and even fewer have been geocoded. Museum records carry a wealth of image and text data, and digitizing these records in a meaningful and useful way remains a serious challenge, in terms of both appropriate technical methods and the practical effort and resources required.

9.7 GENOME-ENABLED INDIVIDUALIZED MEDICINE

By many accounts, knowledge of the sequence of the human genome has enormous potential for changing the practice of medicine and the delivery of health care services. As more is understood about human biology, it is increasingly feasible for medicine to be predictive—to have advance knowledge of how a person's health status will respond (positively or negatively) to various exposures to different foods and environmental events, and to prevent disease and sustain lifelong health and well-being. Both these goals depend on a personalized medicine that begins with deep knowledge of the implications of the genetic makeup of any given individual, as well as his or her health and medical life history. Indeed, one of the most important implications of knowledge of the genome is the possibility that medical treatment and interventions might be more customized to the genetic profile of individuals or groups in ways that maximize the likelihood of successful outcomes.[46]

One necessary precondition for genome-based individualized medicine is technology for the inexpensive acquisition of sequence information—perhaps a few hundred dollars for an individual's complete genome, for example.[47] On the other hand, from a cost-effectiveness standpoint, it is better to stratify individuals into subcategories that are relevant to various treatment or intervention regimes by looking at a limited number of genetic markers, rather than to acquire the complete genetic sequence of all individuals involved. This vision has led major pharmaceutical companies to proclaim that genomic

[45] National Research Council, *Grand Challenges in Environmental Sciences*, National Academy Press, Washington, DC, 2001.

[46] One of the most ambitious efforts to exploit the potential of genome-enabled individualized medicine is being undertaken by Mexico, whose population is composed of more than 65 native Indian groups and Spaniards. Because the overall genetic makeup of this population is associated with a characteristic set of disease susceptibilities, Mexico has undertaken this initiative to reduce the social and financial burden of health problems, since new strategies for prevention, early diagnosis, and more effective treatment are essential to meet the mid- and long-term health care goals in Mexico. See Gerardo Jimenez-Sanchez, "Developing a Platform for Genomic Medicine in Mexico," *Science* 300:295-296, 2003.

[47] Note that this is 10^5 times less expensive than the sequencing of the first genome. Whether the least expensive approach turns out to be sequencing individual genomes from scratch, or sequencing only those portions specific to individuals and integrating those portions into the genome of the generic human, remains to be seen.

medicine and related technologies will allow physicians to provide the right drug to the right patient at the right time. Thus, the term "individualized medicine" should be regarded as one that ranges from single individuals (likely in the farther-term future) to genetically differentiated subpopulations (more likely to happen in the near term).

The fundamental challenge is to correlate genetic variation to susceptibility for specific diseases, specific drug reactions, and specific responses to environmental insult. But even with these correlations in hand, it is a very long way from examination of individual drug-gene interactions to individualized medicine—what might be called translational medicine—that affects the well-being of the citizenry at large. Traversing this distance will require considerable advances on multiple fronts: in the laboratory, on the computer, and in how scientists conceptualize the relationships between all of the individual components involved.

9.7.1 Disease Susceptibility[48]

It has been known for many years that many medical conditions have a genetic basis. Indeed, for many illnesses, the strongest predictor of risk is an individual's family history. The association of specific genomic differences with the likelihood of disease will provide physicians and patients with more specific and more certain information. Such knowledge will allow individuals to takes steps that reduce the likelihood and/or severity of such disease in the future. These steps might include greater medical surveillance or screening, environmental changes, diet, exercise, or preventive drug therapy (e.g., more frequent colonoscopies starting earlier in life for individuals with genetic profiles that imply a high degree of risk for colon cancer).

It is useful to distinguish between genetic signatures that are highly penetrant and those that are highly prevalent. A highly penetrant genetic signature associated with a disease is one whose presence implies a high likelihood that the disease will develop in an individual with that signature: examples provided by Guttmacher and Collins (Footnote 48) include mutations in the BRCA1 and BRCA2 genes that increase the risk of breast and ovarian cancer, in the HNPCC gene set that increases the risk of hereditary nonpolyposis colorectal cancer, and in the gene for synuclein that causes Parkinson's disease. A highly prevalent genetic signature is one that occurs frequently in the population, but its presence may or may not be associated with a large increase in the likelihood that a disease will develop in an individual with that signature: as examples, Guttmacher and Collins (Footnote 48) include a mutation in the factor V Leiden gene that increases the risk of thrombosis, in the APC (adenomatosis polyposis coli) gene that increases the risk of colorectal cancer, and in the apolipoprotein gene that increases the risk of Alzheimer's disease.

From the standpoint of the individual, identification of a *highly penetrant* genetic signature associated with disease will have important clinical ramifications. However, from a public health standpoint, it is the identification of *highly prevalent* genetic signatures associated with disease that is most significant.

The best-understood genetic disorders leading to disease are those associated with the inheritance of a single gene. Such disease conditions have been cataloged in the Online Mendelian Inheritance in Man (OMIM) catalog.[49] Examples of single-gene conditions cited by Guttmacher and Collins include hereditary hemochromatosis, cystic fibrosis, alpha$_1$-antitrypsin deficiency, and neurofibromatosis. These

[48] The discussion in this section on monogenic and highly penetrant signatures is based on excerpts from A.E. Guttmacher and F.S. Collins, "Genomic Medicine—A Primer," *New England Journal of Medicine* 347(19):1512-1520, 2002. The discussion in this section on polygenic and highly prevalent signatures is based on excerpts from P.D. Pharoah, A. Antoniou, M. Bobrow, R.L. Zimmern, D.F. Easton, and B.A. Ponder, "Polygenic Susceptibility to Breast Cancer and Implications for Prevention," *Nature Genetics* 31(1):33-36, 2002. A highly positive and optimistic view of the impact of the genome on medicine can be found in F.S. Collins and V.A. McKusick, "Implications of the Human Genome Project for Medical Science," *Journal of the American Medical Association* 285(5):540-544, 2001. A somewhat contrary view can be found in N.A. Holtzman and T.M. Marteau, "Will Genetics Revolutionize Medicine?" *New England Journal of Medicine* 343(2):141-144, 2000.

[49] See http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.

disorders are highly penetrant, but occur relatively rarely in the population (with approximate incidences of one in several hundred or less).

On the other hand, multifactor genetic causality for disease is almost certainly much more common than monogenic causality. In principle, knowledge of the range of genetic variations across many loci in the population will allow researchers to estimate risks arising from the combined effect of such variations.

Using breast cancer as a case study, Pharoah et al. (footnote 48) compared the potential for prediction of risk based on common genetic variations with the predictions that could be made using known and established risk factors. They concluded that a typical polygenic approach for analysis would suggest that the half of the population at highest risk would account for 88 percent of all affected individuals, if all of the susceptibility genes could be identified. However, using currently known factors for breast cancer to stratify the population, they estimated that the half of the population at highest risk would account for only 62 percent of all cases. Pharoah et al. thus suggest that genetic profiles may provide significant improvement in the ability to differentiate at-risk individuals from individuals not at risk.

Nevertheless, for a variety of reasons, identifying the relevant genetic signatures over multiple genes that account for disease susceptibility will pose significant intellectual challenges. Probably the most important point is that the contribution of any given gene involved is likely to be weak; hence detecting its clinical significance may be problematic. Nongenomic effects, such as posttranslational modifications, may also be relevant. Zimmern[50] notes that even monogenic conditions can result in variable expressivity and incomplete penetrance, and that similar disease phenotypes may result from genetic heterogeneity, whether in the form of allelic heterogeneity (different mutations at the same locus) or locus heterogeneity (where mutations occur at different loci). Different mutations of the same gene may also give rise to separate clinical effects. Environmental factors may be difficult to disentangle from genetic ones. As a consequence of such issues, definitive conclusions about the relationship of a given polygenic genotype to a specific disease condition may well be difficult to draw.

An extension of the genomic approach to disease susceptibility applies to understanding the impact of an individual's genomic composition on that individual's response to various environmental insults to the body, such as those caused by exposure to chemicals (e.g., from drinking water or air pollution) or electromagnetic fields (e.g., from cell phones or ambient radiation). Furthermore, in dealing with certain environmental insults, stochasticity is likely to play an important role. For example, in considering the effects of radiation on the genome, macroscopic parameters that characterize radiation such as duration and intensity are insufficient to determine its effect, simply because what part of a genome is affected is mostly a matter of chance. Thus, a given dose of a certain kind of radiation will not affect individuals in equal measure and, more to the point, could not be expected to affect even an ensemble of identical twins similarly.

Overall, there is wide variability in individual responses to environmental influences. While existing diseases, differences in gender, or differences in nutritional status affect such variability, genetic influences are also important. Genes that affect the human response to environmental exposure (called environmentally responsive genes by the Environmental Genome Project [EGP] of the National Institute of Environmental Health Sciences [NIEHS] tend to fall into several categories.[51] That is, they affect the cell cycle, DNA repair, cell division, cell signaling, cell structure, gene expression, apoptosis, and metabolism. The initial phases of the EGP are focused on identifying single nucleotide polymorphisms (SNPs) associated with 554 genes identified by the scientific community as environmentally responsive. Identification of the SNPs associated with environmentally responsive genes would make it possible to conduct epidemiological studies that classify subjects by SNPs, thus increasing the utility of these

[50] R.L. Zimmern, "The Human Genome Project: A False Dawn?" *British Medical Journal* 319(7220):1282, 1999.

[51] See http://www.niehs.nih.gov/envgenom/egp.htm.

studies in detecting genetic contributions to the likelihood of various diseases with at least partial environmental causation.

The challenges of polygenic data analysis are formidable. An example of methodological research in this area is that of Nelson et al.,[52] who developed the combinatorial partitioning method (CPM) for examining multiple genes, each containing multiple variable loci, to identify partitions of multilocus genotypes that predict interindividual variation in quantitative trait levels. The CPM offers a strategy for exploring the high-dimensional genotype state space so as to predict the quantitative trait variation in the population at large that does not require the conditioning of the analysis on a prespecified genetic model, such as a model that assumes that interacting loci can each be identified through their independent, marginal contribution to trait variability. On the other hand, a brute-force approach to this correlation problem explodes combinatorially. Therefore, it is likely that finding significant correlations will depend on the ability to prune the search space before specific combinations are tested—and the ability to prune will depend on the availability of insight into biological mechanisms.

9.7.2 Drug Response and Pharmacogenomics[53]

As with disease susceptibility, it has been known for many years that different individuals respond differently to the same drug at the same dosages and that the relevant differences in individuals are at least partly genetic in origin. However, characterization of the first human gene containing DNA sequence variations that influence drug metabolism did not take place until the late 1980s.[54] Today, pharmacogenomics—the impact of an individual's genomic composition on his or her response to various drugs—is an active area of investigation that many believe holds significant promise for changing the practice of medicine by enabling individual-based prescriptions for compound and dosage.[55] An individual's genetic profile may well suggest which of several drugs is most appropriate for a given disease condition. Because genetics influence drug metabolism, an individual's weight will no longer be the determining factor in setting the optimal dosage for that individual.

Similarly, many drugs are known to be effective in treating specific disease conditions. However, because of their side effects in certain subpopulations, they are not available for general use. Detailed "omic" knowledge about individuals may help to identify the set of people who might benefit from certain drugs without incurring undesirable side effects, although some degree of empirical testing will be needed if such individuals can be identified.[56] In addition, some individuals may be more sensitive than others to specific drugs, requiring differential dosages for optimal effect.

As in the case of disease susceptibility, the best-understood genetic polymorphisms that affect drug responses in individuals are those that involve single genes. As an example, Evans and Relling note that

[52] M.R. Nelson, S.L.R. Kardia, R.E. Ferrell, and C.F. Sing, "A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative Trait Variation," *Genome Research* 11(3):458-470, 2001.

[53] Much of the discussion in Section 9.7.2 is based on excerpts from W.E. Evans and M.V. Relling, "Moving Towards Individualized Medicine with Pharmacogenomics," *Nature* 429(6990):464-468, 2004.

[54] F.J. Gonzalez, R.C. Dkoda, S. Kimura, M. Umeno, U.M. Zanger, D.W. Nebert, H.V. Gelboin, et al., "Characterization of the Common Genetic Defect in Humans Deficient in Debrisoquine Metabolism," *Nature* 331(6155):442-446, 1988. Cited in Evans and Relling, 2004.

[55] If the promise of pharmacogenomics is realized, a number of important collateral benefits follow as well. Drug compounds that have previously been rejected by regulatory authorities because of their side effects on some part of the general population at large may become available to those individuals genomically identified as not being subject to those side effects. Thus, these individuals would have options for treatment that would not otherwise exist. Furthermore, clinical trials for drug testing could be much more targeted to appropriate subpopulations with a higher likelihood of ultimate success, thus reducing expenses associated with failed trials. Also, in the longer term, pharmacogenomics may enable the customized creation of more powerful medicines based on the specific proteins and enzymes associated with genes and diseases.

[56] Another application often discussed in this context is the notion of drugs customized to specific individuals based on "omic" data. However, the business model of pharmaceutical companies today is based on large markets for their products. Until it becomes possible to synthesize and manufacture different drug compounds economically in small quantity, custom-synthesized drugs for small groups of individuals will not be feasible.

individuals that are deficient in thiopurine *S*-methyltransferase (TPMT) can be treated with much lower doses of the thiopurine drugs mercaptopurine and azathiopurine used as immunosuppressants and to treat neoplasias. There is a clinical diagnostic test available for the genomic detection of the TPMT deficiency, but routine use of TPMT genotyping to make treatment decisions is limited. A second example also discussed by Evans and Relling is that polymorphisms in a gene known as CYP2D6 have a strong effect on individuals' responses to the antihypertensive drug debrisoquine and in the metabolism of the oxytocic drug sparteine.

A second example is found in the area of certain drugs for the treatment of cardiovascular disease. Numerous examples of differences among individuals have been seen as potential candidate pharmacodynamic loci (e.g., those for angiotensinogen, angiotensin-converting enzyme, and the angiotensin II receptor). Polymorphisms at these loci predict responses to specific treatments such as the inhibition of angiotensin-converting enzyme. Here, researchers hope to establish and utilize antihypertensive drugs that are matched to the genetic variations among individuals, and thus to optimize blood pressure control and reduce side effects.[57]

A number of monogenic polymorphisms have been found, encoding drug-metabolizing enzymes, drug transporters, and drug targets, as well as disease-modifying genes, that have been linked to drug effects in humans. However, these are the "low-hanging fruit" of pharmacogenetics, and for most drug effects and treatment outcomes, monogenic polymorphisms with clearly recognizable drug-response phenotypes do not characterize the situation. For example, as in the case of disease susceptibility, nongenomic effects (e.g., posttranslational modifications) on protein function may be relevant. Or, multiple genes may act together in networks to create a single drug-response phenotype.

As Evans and Relling note, genome-wide approaches, such as gene expression arrays, genome-wide scans, or proteomic assays, can contribute to the identification of as-yet-unrecognized candidate genes that may have an influence on a drug response phenotype. For example, it may be possible to detect genes whose expression differentiates drug responders from nonresponders (or those for whom certain drugs are toxic from those for whom they are not), genomic regions with a paucity of heterozygosity in responders compared with nonresponders, or proteins whose abundance differentiates drug responders from nonresponders.

In expression-array and proteomic approaches, the level of the signal may directly reflect functional variation—a distinct advantage from an experimental point of view. Yet there can be many other reasons for differences in signal level, such as the choice of tissue from which the samples are drawn (which may not be the tissue of interest where toxicity or response is concerned) or changes in function not reflected by levels of mRNA or protein. Thus, when such studies suggest that a given gene or gene product is relevant to drug response, Evans and Redding point out that large-scale molecular epidemiological association studies (in vivo or in vitro with human tissues), biochemical functional studies, and studies on preclinical animal models of candidate gene polymorphisms become necessary to further establish the link between genetic polymorphism and drug response.

A second challenge in pharmacogenomics relates to integrating pharmacogenomics with the everyday practice of medicine. Although there are cultural and historical sources of resistance to such integration, it is also true that definitive clinical pharmacogenomic studies have not been conducted that demonstrate unambiguously the benefits of integration on clinical outcomes. Indeed, there are many difficulties in conducting such studies, including the multigenic nature of most drug effects and the difficulty in controlling for nongenetic confounders such as diet or exercise. Until such difficulties are overcome, it is unlikely that a significant change will occur in clinical practice.

One of the most important databases for the study of pharmacogenomics is a database known as the Stanford PharmGKB, described in Box 3.4. Supported by the National Institute of General Medical

[57] P. Cadman and D. O'Connor, "Pharmacogenomics of Hypertension," *Current Opinion in Nephrology and Hypertension* 12(1):61-70, 2003.

Science (NIGMS), PharmGKB is a publicly available, Internet-accessible database for pharmacogenetics and pharmacogenomics. Its overall aim is to aid researchers in understanding how genetic variation among individuals contributes to differences in reactions to drugs.[58] The database integrates pharmacodynamics (drug actions), pharmacokinetics (drug metabolism), toxicity, sequence and other molecular data, pathway information, and patient data.

9.7.3 Nutritional Genomics

Traditional nutrition research has had among its goals the establishment of overarching dietary recommendations for everyone—in principle, for the world's entire population. Today, and more so in the future, the implications of individual genetic makeup for optimal diet have changed that perspective. To understand and exploit the interplay of diet and genetics, nutritional genomics is a relatively new specialization within the life sciences with two separate but related foci. One focus relates an individual's genetic makeup to dietary regimes that are more or less healthy for him or her. For example, it is well known that some individuals are more likely to suffer from high blood pressure if they consume salt in relatively large quantities, while others are not. Poch et al.[59] found a possible genetic basis on which to differentiate salt-sensitive individuals and salt-insensitive ones. If it is possible to develop genetic tests for salt sensitivity, salt-sensitive individuals could be advised specifically to limit their salt intake, and salt-insensitive individuals could continue to indulge at will their taste for salty snacks.

The traditional focus of nutrition research is not in any way rendered irrelevant by nutritional genomics. Still, beyond general good advice and informed common sense, in the most ambitious scenarios, recommended dietary profiles could be customized for individuals based on their specific genomic composition. Ordovas and Corella write:[60]

> Nutritional genomics has tremendous potential to change the future of dietary guidelines and personal recommendations. Nutrigenetics will provide the basis for personalized dietary recommendations based on the individual's genetic makeup. This approach has been used for decades for certain monogenic diseases; however, the challenge is to implement a similar concept for common multifactorial disorders and to develop tools to detect genetic predisposition and to prevent common disorders decades before their manifestation. . . . [P]reliminary evidence strongly suggests that the concept should work and that we will be able to harness the information contained in our genomes to achieve successful aging using behavioral changes; nutrition will be the cornerstone of this endeavor.

A second focus of nutritional genomics is on exploiting the potential for modifying foodstuffs to be more healthy, and so dietary advice and discipline might be supplanted *in part* by such modifications. For example, it may be possible to redesign the lipid composition of oil seed crops using genetic modification techniques (through either selective breeding or genetic engineering). However, whether this is desirable depends on how consumption of a different mix of lipids affects human health. Watkins et al.[61] argue for an understanding of the overall metabolomic expression of lipid metabolism to ensure that a particular metabolite composition truly improves overall health, so that a change in lipid composition that is deemed healthy when viewed as lowering the risk of one disease does not simultaneously increase the risk of developing another.

[58] T.E. Klein and R.B. Altman, "PharmGKB: The Pharmacogenetics and Pharmacogenomics Knowledge Base," *Pharmacogenomics Journal* 4(1):1, February 2004.

[59] E. Poch, D. Gonzalez, V. Giner, E. Bragulat, A. Coca, and A. de La Sierra, "Molecular Basis of Salt Sensitivity in Human Hypertension: Evaluation of Renin-Angiotensin-Aldosterone System Gene Polymorphisms," *Hypertension* 38(5):1204-1209, 2001.

[60] J.M. Ordovas and D. Corella, "Nutritional Genomics," *Annual Review of Genomics and Human Genetics* 5:71-118, 2004.

[61] S.M. Watkins, B.D. Hammock, J.W. Newman, and J.B. German, "Individual Metabolism Should Guide Agriculture Toward Foods for Improved Health and Nutrition," *American Journal of Clinical Nutrition* 74(3):283-286, 2001.

More generally, Watkins et al. point out that the goal of nutritional improvement of agriculture—to produce changes in crops and foods that provide health benefits to all—is difficult to achieve because modifications of existing foodstuffs are likely to advantage some people while disadvantaging others. Watkins et al. cite the example of recent attempts to increase the carotenoid content of the food supply—a move that was thought to have protective value against certain cancers, especially lung cancer. In the midst of this effort, it was found that high intakes of ß-carotene as a supplement actually *increased* the incidence of lung cancer in smokers—and the move was abandoned.

The intellectual underpinning of this effort is thus metabolomics, the quantitative characterization of the set of metabolites—generally small, nonprotein molecules—involved in the metabolism or a cell, tissue, or organism over its lifetime. In the context of nutritional genomics, metabolomic studies attempt to characterize the levels, activities, regulation, and interactions of all metabolites in an individual and determine how this characterization changes in response to various foods that are consumed. Genomics is important because genetic makeup is an important influence on the specific nature of the metabolomic changes that result as a function of food consumption.

9.8 A DIGITAL HUMAN ON WHICH A SURGEON CAN OPERATE VIRTUALLY

A surgical act on a human being is by definition an invasive process, one that inflicts many insults on the body. Prior to the advent of medical imaging techniques, surgeons relied on their general knowledge of anatomy to know where and what to cut. Today's imaging technologies provide the surgeon with some idea of what to expect when he or she opens the patient.

At the same time, a surgeon in the operating room has no opportunity to practice the operation on this particular patient. Experience with other patients with similar conditions helps immeasurably, of course, but it is still not uncommon even in routine surgical operations to find some unexpected problem or complication that the surgeon must manage. Fortunately, most such problems are minor and handled easily. Surgeons-in-training operate first on cadavers and move to live patients only after much practice and under close supervision.

Consider then the advantages that a surgeon might have if he or she were to be able to practice a difficult operation before doing it on a live patient. That is, a surgeon (or surgeon-in-training) would practice or train on a digital model of a human patient that incorporates static and dynamic physical properties of the body in an operating room environment (e.g., under anesthesia, in real gravity) when it is subject to surgical instruments.

In this environment, the surgeon would likely wear glasses that projected an appropriate image to his or her retina and use implements that represented real instruments (e.g., a scalpel). Kinetic parameters of the instrument (e.g., speed, velocity, orientation) would be monitored and registered onto the image that the surgeon sees. When "touched" by the instrument, the image would respond appropriately with a change in shape and connectivity (e.g., when a scalpel touches a piece of skin, it might separate into two parts and a cut would appear). Blood would emerge at realistic rates, and tissue under the skin would appear.

Even in this very simple example, many challenges can be seen. To name just a few:

- *Realistic modeling of body subsystems.* From the perspective of a surgeon's scalpel, the body is simply a heterogeneous and spatially organized mass of tissue. Of course, this mass of tissue is functionally a collection of subsystems (e.g., organs, muscle tissue, bone) that have different properties. These subsystems must be separated so that the physiological responses of surgery are appropriately propagated through them when surgery occurs.
- *Integration of person-specific information with a generic model of a human being.* Because of the labor involved in constructing a digital model of a human being, it makes sense to consider an approach in which a model of a generic human being is developed and then adjusted according to person-specific information of any given patient.

- *Spatial registration and alignment of instruments, the surgeon's hands, and the digital body being operated on.* The surgeon must see an instrument move to the position in the body to which the surgeon has moved it. When a cutting motion is made, the appropriate tissue should split in the appropriate place and amount.
- *The different feel and texture of tissue depending on whether the instrument is a scalpel or a finger.* A digital human for surgical use must provide appropriate force feedback ("haptic capability") to the surgeon so that, for example, cutting into soft tissue feels different than cutting into bone.
- *Incorporation of gravity in the model.* Many organs consist of soft tissue that is deformed easily under pressure from instruments and touch. As importantly, tissues are subject to gravitational forces that will change their shape as their orientation is changed (the breast of a woman lying on her back has an entirely different shape than when she is lying on her side).

Some first steps have been taken in many of these areas. For example, a project at the Ohio Supercomputer Center (OSC) in 1996 sought to develop a virtual reality-based simulation of regional anesthesia that employed haptic techniques to simulate the resistance felt when an injection is given in a certain area (Box 9.4).

A second example is work in computational anatomy, one application of which has sought to characterize the structure of human brains in a formal manner. Structure is interesting to neuroscientists because of a presumed link between physical brain structure and neurological function. Through mathematical transformations that can deform one structure into another, it is possible to develop metrics that can characterize how structurally different two brains are. These metrics can then be correlated with understanding of the neurological functions of which each brain is capable (Box 9.5). Such metrics can also be used to identify normal versus diseased states that are reflected anatomically.

Box 9.4
A Virtual Reality Simulation of Regional Anesthesia

A collaborative effort between researchers at the Ohio State University Hospitals, Immersion Corporation, and the Ohio Supercomputer Center has led to the creation of a virtual reality simulator that enables anesthesiologists-in-training to practice in a realistic environment the injection of a local anesthetic into the epidural space of the spinal column. The system includes a workstation capable of stereo display, a real-time spatial volume renderer, a voice-activated interface, and most importantly, a one-dimensional haptic probe capable of simulating the resistive forces of penetrated tissues.

Although this procedure appears simple, it is in fact a delicate manual operation that requires the placement of a catheter into a small epidural space using only haptic cues (i.e., cues based on tactile sensations of pressure) to guide the needle. By feeling the resistive forces of the needle passing through various tissues, the anesthesiologist must maneuver the tip of the needle into the correct space without perforating or damaging the spinal cord in the process.

The system is designed to enable the trainee to practice the procedure on a variety of datasets representative of what he or she might experience with real patients. That is, the pressure profile as a function of needle penetration would vary from patient to patient. By training in this environment, the trainee can gain proficiency in the use of this technique in a non-harmful manner.

SOURCE: L. Hiemenz, J.S. McDonald, D. Stredney, and D. Sessanna, "A Physiologically Valid Simulator for Training Residents to Perform an Epidural Block," *Proceedings of the 15th Southern Biomedical Engineering Conference*, March 29-31, 1996, Dayton, OH. See also http://www.osc.edu/research/Biomed/past_projects/anesthesia/index.shtml.

Box 9.5
Computational Anatomy

Computational anatomy seeks to make more precise the commonsense notion that samples of a given organ from a particular species are both all the same and all different. They are the same in the sense that all human brains, for example, exhibit similar anatomical characteristics and can be associated with the canonical brain of *Homo sapiens*, rather than the canonical brain of a dog. They are all different in the sense that each individual has a slightly different brain, whose precise anatomical characteristics differ somewhat from those of other individuals.

Computational anatomy is based on a mathematical formalism that allows one structure (e.g., a brain) to be deformed reversibly into another. (Reversibility is important because irreversible processes destroy information about the original structure.) In particular, the starting structure is considered to be a deformable template. The template anatomy is morphed into the target structure via transformations applied to subvolumes, contours, and surfaces. These computationally intensive transformations are governed by generalizations of the Euler equations of fluid mechanics and are required only to preserve topological relationships (i.e., to transform smoothly from one to the other).

Key to computational anatomy is the ability to calculate a measure of difference between similar structures. That is, a distance parameter should represent in a formalized manner the extent to which two structures differ—and a distance of zero should indicate that they are identical. In the approach to computational anatomy pioneered by Grenander and Miller,[1] the distance parameter is the square root of the energy required to transform the first structure onto the metric of the second with the assumption that normal transformations follow the least-energy path.

One instance in which computational anatomy has been used is in understanding the growth of brains as juveniles mature into adults. Thompson et al.[2] have applied these deformation techniques to the youngest brains, with results that accord well with what was seen in older subjects. In particular, they are able to predict the most rapid growth in the isthmus, which carries fibers to areas of the cerebral cortex that support language function. A second application has sought to compare monkey brains to human brains.

[1] U. Grenander and M.I. Miller, "Computational Anatomy: An Emerging Discipline," *Quarterly Journal of Applied Mathematics* 56:617-694, 1998.

[2] P.M. Thompson, J.N. Giedd, R.P. Woods, D. Macdonald, A.C. Evans, and A.W. Toga, "Growth Patterns in the Developing Brain Detected by Using Continuum Mechanical Tensor Maps," *Nature* 404:190-193, March 9, 2000; doi:10.1038/35004593.

SOURCE: Much of this material is adapted from "Computational Anatomy: An Emerging Discipline," *EnVision* 18(3), 2002, available at http://www.npaci.edu/envision/v18.3/anatomy.html#establishing.

9.9 COMPUTATIONAL THEORIES OF SELF-ASSEMBLY AND SELF-MODIFICATION[62]

Self-assembly is any process in which a set of components joins together to form a larger, more complex structure without centralized or manual control. For example, it includes biologically significant processes ranging from the joining of amino acids to form a protein and embryonic development to nonbiological chemical processes such as crystallization. More recently, the term has become widely used as researchers attempt to create artificial self-assembling systems as a way to fabricate structures efficiently at nanometer scale.

One kind of structure—that can be described as a simple repeating pattern in which molecules form into a regular structure or lattice—is the basis for creating artifacts such as crystals or batteries that can be extended to potentially macroscopic scale; this process is known as periodic self-assembly. However, for applications such as electronic circuits, which cannot be described as a simple repeating pattern, a

[62] Section 9.9 is based largely on material from L. Adleman, Q. Cheng, A. Goel, M.-D. Huang, D. Kempe, P. Moisset de Espanés, P. Wilhelm, and K. Rothemund, "Combinatorial Optimization Problems in Self-Assembly," STOC '02, available at http://www.usc.edu/dept/molecular-science/optimize_self_assembly.pdf.

more expressive form of self-assembly is required. Ideally, a designer could select a set of components and a set of rules by which they connect, and the system would form itself into the desired final shape.

This kind of self-assembly, called nonperiodic or programmable self-assembly, would allow the creation of arbitrary arrangements of components. Nonperiodic self-assembly would be useful for the efficient execution of tasks such as electronic circuit design, material synthesis, micro- and nanomachine construction, and many other technological feats. For the purposes of artificial self-assembly technology, the pinnacle result of a theory would be to be able to select or design an appropriate set of components and assembling rules to produce an arbitrary desired result.

Self-assembly, both as a biological process and as a potential technology, is poorly understood. A range of significant (and possibly insuperable) engineering and technological challenges stands in the way of effectively programming matter to form itself into arbitrary arrangements. A less prominent but no less important challenge is the lack of a theoretical foundation for self-assembly.

A theory of self-assembly would serve to guide researchers to determine which structures are achievable, select appropriate sets of components and assembling rules to produce desired results, and estimate the likely time and environmental conditions necessary to do so. Such a theory will almost certainly be based heavily on the theory of computation and will more likely be a large collection of theoretical results and proofs about the behavior of self-assembling systems, rather than a single unified theory such as gravity.

The grandest form of such a theory would encompass and perhaps unify a number of disparate concepts from biology, computer science, mathematics, and chemistry—such as thermodynamics, catalysis and replication, computational complexity, and tiling theory[63] and would require increases in our understanding of molecular shape, the interplay between enthalpy and entropy, and the nature of noncovalent binding forces.[64] A central caveat is that self-assembly occurs with a huge variety of mechanisms, and there is no a priori reason to believe that one theory can encompass all or most of self-assembly and also have enough detail to be helpful to researchers. In more limited contexts, however, useful theories may be easier to achieve, and more limited theories could serve in guiding researchers to determine which structures are achievable or stable, to identify and classify failure modes and malformation, or to understand the time and environmental conditions in which various self-assemblies can occur. Furthermore, theories in these limited contexts may or may not have anything to do with how real biological systems are designed.

For example, progress so far on a theory of self-assembly has drawn heavily from the theory of tilings and patterns,[65] a broad field of mathematics that ties together geometry, topology, combinatorics, and elements of group theory such as transitivity. A tiling is a way for a set of shapes to cover a plane, such as M.C. Escher's famous tesselation patterns. Self-assembly researchers have focused on nonperiodic tilings, those in which no regular pattern of tiles can occur. Most important among aperiodic patterns are Wang tiles, a set of tiles for which the act of tiling a plane was shown to be equivalent to the operation of a universal Turing machine.[66] (Because of the grounding in the theory of Wang tiles in particular, the components of self-assembled systems are often referred to as "tiles" and collections of tiles and rules for attaching them as "tiling systems.")

With a fundamental link between nonperiodic tilings and computation being established, it becomes possible to consider the possibility of programming matter to form desired shapes, just as Turing machines can be programmed to perform certain computations. Additionally, based on this relationship, computationally inspired descriptions might be sufficiently powerful to describe biological self-assembly processes.

Today, one of the most important approaches to a theory of self-assembly focuses on this abstract model of tiles, which are considered to behave in an idealized, stochastic way. Tiles of different types are present in the environment in various concentrations, and the probability of a tile of a given type

[63] L.M. Adleman, "Toward a Mathematical Theory of Self-Assembly," USC Tech Report, 2000, available at http://www.usc.edu/dept/molecular-science/papers/fp-000125-sa-tech-report-note.pdf.

[64] G. Whitesides, "Self-Assembly and Nanotechnology," *Fourth Foresight Conference on Molecular Nanotechnology*, 1995.

[65] B. Grunbaum and G.C. Shephard, *Tilings and Patterns*, W. H. Freeman and Co., New York, 1987.

[66] H. Wang, "Notes on a Class of Tiling Problems," *Fundamenta Mathematicae* 82:295-305, 1975.

attempting to connect to an established shape is proportional to its share of the total concentration of tiles. Then, the "glues" of touching sides of the adjacent tiles have a possibility of attaching. Simulating such self-assembly is actually relatively simple. Given a set of tiles and glues, a simulation can predict with arbitrary accuracy the end result. However, this is complicated by the fact that a given tiling system might not have a unique end result; situations could arise in which two different tiles join the assembly in the same location. While this may seem an undesirable situation, such ambiguous systems may be necessary to perform universal computation.[67]

The more challenging question is the converse of simulation: Given a desired result, how do we get there? Research into the theory of self-assembly has focused on two more specific framings of this question. First, what is the minimum number of tile types necessary to create a desired shape (the "Minimum Tile Set Problem") and, given a specific tiling system, what concentrations produce the end result the fastest (the "Tile Concentrations Problem")? The former has been shown to be NP-complete, but has polynomial solutions given certain restrictions on shape and temperature.

The current state of the art in the theory of self-assembly abstracts away much of the details of chemistry. First, the theory considers only the assembly of two-dimensional patterns. For artificial DNA tiles, designed to be flat, rigid, and square, this may be a reasonable approximation. For a more general theory that includes the self-assembly in three dimensions of proteins or other nonrigid and highly irregularly shaped macromolecules, it is less clear that such a theory is sufficient. Extending the current theory to irregular shapes in three dimensions is a key element of this challenge problem.

The history of the motivation of research into the theory of self-assembly provides a lesson for research at the BioComp interface. Originally, researchers pursued the link between self-assembly and computation because they envisioned self-assembled systems constructed from DNA as potential competitors to electronic digital computing hardware, that is, using biochemistry in the service of computation. However, as it became less obvious that this research would produce a competitive technology, interest has shifted to using the computational theory of self-assembly to increase the sophistication of the types of molecular constructs being created. In other words, today's goal is to use computational theory in the service of chemistry. This ebb and flow of both source theory and application between computation and biochemistry is a hallmark of a successful model of research at the interface.

Another area related to theories of self-assembly is what might be called adaptive programming. Today, most programs are static; although variables change their values, the structure of the code does not. Because computer hardware does not fundamentally differentiate between "code" and "data" (at the machine level, both are represented by 1's and 0's), there is no reason in principle that code cannot modify itself in the course of execution. Self-modifying code can be very useful in certain contexts, but its actual execution path can be difficult to predict and, thus, the results that might be obtained from program execution are uncertain.

However, biological organisms are known to learn and adapt to their environments—that is, they self-modify under certain circumstances. Such self-modification occurs at the genomic level, where the DNA responsible for the creation of cellular proteins contains both genetic coding and regions that regulate the extent to which, and the circumstances under which, genes are activated. It also occurs at the neural level, where cognitive changes (e.g., a memory or a physical skill) are reflected in reorganized neural patterns. Thus, a deep understanding of how biology organizes self-modification in using DNA or in a neural brain may lead to insights about how one might approach human problems that call for self-modifying computer programs.

9.10 A THEORY OF BIOLOGICAL INFORMATION AND COMPLEXITY

Much of this report is premised on the notion of biology as an information science and has argued that information technology is essential for acquiring, managing, and analyzing the many types of

[67] P.W. Rothemund, "Using Lateral Capillary Forces to Compute by Self-Assembly," *Proceedings of the National Academy of Sciences* 97(3):984-989, 2000.

biological data. This fact—that an understanding of biological systems depends on so many different kinds of biological data, operating at so many different scales, and in such volume—suggests the possibility that biological information and/or biological complexity might be notions with some formal quantitative meaning.

How much information does a given biological system have? How should biological complexity be conceptualized? Can we quantify or measure the amount of information or the degree of complexity resident in, say, a cell, or perhaps even more challengingly, in an organelle, an ecosystem, or a species? In what sense is an organism more complex than a cell or an ecosystem more complex than an individual organism? Establishing an intellectually rigorous methodology through which such information could be measured, capturing not only the raw scale of information needed to describe the constituent elements of a system but also its complexity, could be a powerful tool for answering questions about the nature of evolution, for quantifying the effects of aging and disease, and for evaluating the health of ecologies or other complex systems.

Developing such a theory of biological information and complexity will be extraordinarily challenging, however. First, complexity and information exist at a vast range of orders of magnitude in size and time, as well as in the vast range of organisms on Earth, and it is not at all clear that a single measure or approach could be appropriate for all scales or creatures. Second, progress toward such a theory has been made in fields traditionally separate from biology, including physics and computer science. Transferring knowledge and collaboration between biology and these fields is difficult at the best of times, and doubly challenging when the research is at an early stage. Finally, such a theory may prove to be the basis of a new organizing principle for biology, which may require a significant reorientation for practicing biologists and biological theory.

Some building blocks for such a theory may already be available. These include information theory, formulated by Claude Shannon in the mid-20th century for analyzing the performance of noisy communication channels; an extension of information theory, developed over the last few decades by theoretical physicists, that defines information in thermodynamic terms of energy and entropy; the body of computational complexity theory, starting from Turing's model of computation and extending it to include classes of complexity based on the relative difficulty of families of algorithms; and complexity theory (once called "chaos theory"), an interdisciplinary effort by physicists, mathematicians, and biologists to describe how apparently complex behavior can arise from the interaction of large numbers of very simple components.

Measuring or even defining the complexity of a biological system—indeed, of any complex, dynamic system—has proven to be a difficult problem. Traditional measures of complexity that have been developed to analyze and describe the products of human technological engineering are difficult to apply or inappropriate for describing biological systems. For example, although both biological systems and engineered systems often have degrees of redundancy (i.e., multiple instances of the same "component" that serve the same function for purposes of reliability), biological systems also show many other systems-level design behaviors that are rarely if ever found in engineered systems. Indeed, many such behaviors would be considered poor design. For example, "degeneracy" in biological systems refers to the property of having different systems produce the same activity. Similarly, in most biological systems, many different components contribute to global properties, a design that if included in a human-engineered system would make it very difficult to understand.

Other attempts at measuring biological complexity include enumerating various macroscopic properties of an organism, such as the number of distinct parts, number of distinct cell types, number of biological functions performed, and so forth. In practice this can be difficult (what is considered a "distinct" part?) or inconclusive (is an organism with more cell types necessarily more complex?).

More conveniently, the entire DNA sequence of an organism's genome can be analyzed. Since DNA plays a major role in determining the structure and functions of an organism, one approach is to consider the information content of the DNA string. Of course, biological knowledge is nowhere close to actually being able to infer the totality of an organism merely from a DNA sequence, but the argument is that sequence complexity will be highly correlated with organismal complexity. (Some advantages of dealing with strings of letters as an abstraction are discussed in Section 4.4.1.)

Because information theory treats all bits as alike and of equal significance, a purely information-theoretic view would suggest that a gene of a thousand base pairs that encode a crucial protein required for the development of a human characteristic has the same information content (about 2,000 bits) as a random sequence of the same length with no biological function. This view strains plausibility or, rather, would have limited applicability to biology. Thus, the example suggests that something more is needed.

Generally, the types of complexity measures applied to DNA sequences are defined by their relationship to the process of computation. For example, a string might be considered to be a program, an input to a program, or the output of a program, and the resulting complexity measure might include the size of the Turing machine that produced it, its running time, or the number of states. Each measure captures a different sense of complexity of the DNA string and will consider different strings to be relatively more or less complex.

One such approach is the notion of Kolmogorov (or more formally, Kolmogorov-Chaitin-Solomonoff) complexity. Kolmogorov complexity is a measure of the extent to which it is possible to eliminate redundancies from a bit string without loss of information. Specifically, a program is written to generate the bit string in question. For a truly random string, the program is at least as long as the string itself. But if there are information redundancies in the string, the string can be compressed, with the compressed representation being the program needed to reproduce it. A string with high Kolmogorov complexity is one in which the difference in length between the string and its program is small; a string with low Kolmogorov complexity is one that contains many redundancies and thus for which the generating program is shorter than the string.

However, for the purpose of analyzing overall complexity, a purely random string will have a maximal Kolmogorov score, which is not what seems appropriate intuitively for estimating biological complexity. In general, a desired attribute of measures of biological complexity is the so-called one-hump criterion. A measure that incorporated this criterion would indicate a very low complexity for both very ordered sequences (e.g., a purely repeating sequence) and very random sequences and the highest complexity for sequences in the middle of a notional continuum, neither periodic nor random.[68] Feldman and Crutchfield further suggest that biological complexity must also be defined in a setting that gives a clear interpretation to what structures are quantified.[69]

Other measures that have been proposed include thermodynamic depth, which relates a system's entropy to the number of possible histories that produced its current state; logical depth, which considers the minimal running time of a program that produced a given sequence; statistical complexity measures, which indicate the correlation among different elements of an entity's components and the

[68] A related phenomenon, highly investigated but poorly understood, is the ubiquity of so-called 1/*f* spectra in many interesting phenomena, including biological systems. The term "1/*f* spectra" refers to a type of signal whose power distribution as a function of frequency obeys an inverse power law in which the exponent is a small number. A 1/*f* signal is not random noise (random noise would result in an exponent of zero; i.e., the power spectrum of a random noise source is flat). On the other hand, there is some stochastic component to 1/*f* spectra as well as some correlation between signals at different nonadjacent times (i.e., 1/*f* noise exhibits some degree of long-range correlation). Similar statistical analyses have been applied to spatial structures, such as DNA, although power and frequency are replaced by frequency of base-pair occurrence and spatial interval, respectively (see, for example, A.M. Selvam, "Quantumlike Chaos in the Frequency Distributions of the Bases A, C, G, T in Drosophila DNA," *APEIRON* 9(4):103-148, 2002; W. Li, T.G. Marr, and K. Kaneko, "Understanding Long-range Correlations in DNA Sequences," *Physica D* 75(1-3):392-416, 1994 [erratum:82, 217,1995]). 1/*f* spectra have been found in the temporal fluctuations of many biological processes, including ion channel kinetics, auditory nerve firings, lung inflation, fetal breathing, human cognition, walking, blood pressure, and heart rate. (See J.M. Hausdorff and C.K. Peng, "Multiscaled Randomness: A Possible Source of 1/*f* Noise in Biology," *Physical Review E* 54(2):2154-2157, 1996, and references therein. Hausdorff and Peng suggest that if the time scales of the inputs affecting a biological system are "structured" and there are a large number of inputs, it is very likely that the output will exhibit 1/*f* spectra, even if individual input amplitudes and time scales are loosely correlated.)

[69] D.P. Feldman and J.P. Crutchfield, "Measures of Statistical Complexity: Why?," *Physics Letters A* 238:244-252, 1997.

degree of structure or pattern in that entity; and physical complexity, which interprets the shared Kolmogorov complexity of an ensemble of sequences as information stored in the genome about the environment. This last makes the interesting point that one cannot know anything about the meaning of a DNA sequence without considering the environment in which the corresponding organism is expected to live.

All of these capture some aspect of the way in which complexity might arise over time through an undirected evolutionary process and be stored in the genome of a species. However, in their physics-inspired search for minimal descriptions, they may be missing the fact that evolution does not produce optimal or minimal descriptions. That is, because biological organisms are the result of their evolutionary histories, they contain many remnants that are likely to be irrelevant to their current environmental niches, yet contribute to their complexity. Put differently, any given biological organism is almost certainly not optimized to perform the functions of which it is capable.

Another difficulty with many of these measures' application to biology is that, regardless of their theoretical soundness, they will almost certainly be hard to determine empirically. More prosaically, they often involve a fair amount of mathematics or theoretical computational reasoning (e.g., to what level of the Chomskian hierarchy of formal languages does this sequence belong?) completely outside the experience of the majority of biologists. Regardless, this is an area of active research, and further integration with actual biological investigation is likely to produce further progress in identifying accurate and useful measures of complexity.

10

Culture and Research Infrastructure

Earlier chapters of this report have focused on what might be achieved experimentally and on the scientific and technical hurdles that must be overcome at the interface of biology and computing. This chapter focuses on the infrastructural underpinnings needed to support research at this interface. Note that because the influence of computing on biology has been much more significant than the influence of biology on computing, the discussion in this chapter is focused mostly on the former.

10.1 SETTING THE CONTEXT

In 1991, Walter Gilbert sketched a vision of 21st century biology (described in Chapter 1) and noted the changes in intellectual orientation and culture that would be needed to realize that vision. He wrote:

> To use [the coming] flood of [biological] knowledge [i.e., sequence information], which will pour across the computer networks of the world, biologists not only must become computer-literate, but also change their approach to the problem of understanding life. . . . The next tenfold increase in the amount of information in the databases will divide the world into haves and have-nots, unless each of us connects to that information and learns how to sift through it for the parts we need. This is not more difficult than knowing how to access the scientific literature as it is at present, for even that skill involves more than a traditional reading of the printed page, but today involves a search by computer. . . . We must hook our individual computers into the worldwide network that gives us access to daily changes in the database and also makes immediate our communications with each other. The programs that display and analyze the material for us must be improved—and we [italics added] must learn how to use them more effectively.[1]

In short, Gilbert pointed out the need for institutional change (in the sense of individual life scientists learning to cooperate with each other) and for biologists to learn how to use the new tools of information technology.

Because the BioComp interface encompasses a variety of intellectual paradigms and disparate institutions, Section 10.2 describes the organizational and institutional infrastructure supporting work at this interface, illustrating a variety of programs and training approaches. Section 10.3 addresses some

[1] W. Gilbert, "Toward a Paradigm Shift in Biology," *Nature* 349(6305):99, 1991.

of the barriers that affect research at the BioComp interface. Chapter 11 is devoted to proposing possible ways of helping to reduce the negative impact of these barriers.

10.2 ORGANIZATIONS AND INSTITUTIONS

Efforts to pursue research at the BioComp interface, as well as the parallel goal of attracting and training a sufficient workforce, are supported by a number of institutions and organizations in the public and private sectors. A prime mover is the U.S. government, both by pursuing research in its own laboratories, and by providing funding to other, largely academic, organizations. However, the government is only a part of a larger web of collaborating (and competing) academic departments, private research institutions, corporations, and charitable foundations.

10.2.1 The Nature of the Community

The members of an established scientific community can usually be identified by a variety of commonalities—fields in which their degrees were received, journals in which they publish, and so on.[2] The fact that important work at the BioComp interface has been undertaken by individuals who do not necessarily share such commonalities indicates that the field in question has not jelled into a single community, but in fact is composed of many subcommunities. Members of this community may come from any of a number of specialized fields, including (but not restricted to) biology, computer science, engineering, chemistry, mathematics, and physics. (Indeed, as the various epistemological and ontological discussions of previous chapters suggest, even philosophers and historians of science may have a useful role to play.)

Because the intellectual contours of work at the intersection have not been well established, the definition of the community must be broad and is necessarily somewhat vague. Any definition must encompass a multitude of cultures and types, leaving room for approaches that are not yet known. Furthermore, the field is sufficiently new that people may enter it at many different stages of their careers.

For perspective, it is useful to consider some possible historical parallels with the establishment of biochemistry, biophysics, and bioengineering as autonomous disciplines. In each case, the phenomena associated with life have been sufficiently complex and interesting to warrant the bringing to bear of specialized expertise and intellectual styles originating in chemistry, physics, and engineering. Nonbiologists, including chemists, physicists, and engineers, have made progress on some biologically significant problems precisely because their approaches to problems differed from those of biologists and thus have advanced biological understanding because they were not limited by what biologists felt could not be understood. On the other hand, chemists, physicists, and engineers have also pursued many false or unproductive lines of inquiry because they have not appreciated the complexity that characterizes many biological phenomena or because they addressed problems that biologists already regarded as solved. Eventually, biochemistry, biophysics, and bioengineering became established in their own right as education and cultural inculcation from both parent disciplines came to be required.

It is also to be expected that the increasing integration of computing and information into biology will raise difficult questions about the nature of biological research and science. If an algorithm to examine the phylogenetic tree of life is too slow to run on existing hardware, clearly a new algorithm must be developed. Does developing such an algorithm constitute biological research? Indeed, modern biology is sufficiently complex that many of the most important biological problems are not easily tamed by existing mathematical theory, computational models, or computing technologies. Ultimately, success in understanding biological phenomena will depend on the development and application of new tools throughout the research process.

[2] T.S. Kuhn, *The Structure of Scientific Revolutions,* Third Edition, University of Chicago Press, Chicago, IL, 1996.

10.2.2 Education and Training

Education, either formal or informal, is essential for practitioners of one discipline to learn about another, and there are many different venues in which training for the BioComp interface may occur. (Contrast this to a standard program in physics, for example, in which a very typical career path involves an undergraduate major in physics, graduate education in physics culminating in a doctorate, and a postdoctoral appointment in physics.)

Reflecting this diversity, it is difficult to generalize about approaches toward academic training at the BioComp interface, since different departments and institutions approach it with varied strategies. One main difference in approaches is whether the initiative for creating an educational program and the oversight and administration of the program come from the computer science department or the biology department. Other differences include whether it is a stand-alone program or department, or a concentration or interdisciplinary program that requires a student or researcher to have a "home" department as well, and whether the program was established primarily as a research program for postdoctoral fellows and professors (and is slowly trickling down to undergraduate and graduate education), or as an undergraduate curriculum that is slowly building its way up to a research program. Those differences in origin result in varying emphases on what constitutes core subject matter, whether interdisciplinary work is encouraged and how it is handled, and how research is supported and evaluated.

What is clear is that this is an active area of development and investment, and many major colleges and universities have a formal educational program of some sort at the BioComp interface (generally in bioinformatics or computational biology) or are in the process of developing one. Of course, there is not yet widespread agreement on what the curriculum for this new course of study should be[3] or indeed if there should be a single, standard, curriculum.

10.2.2.1 General Considerations

As a general rule, serious work at the BioComp interface requires knowledge of both biology and computing. For example, many models and simulations of biological phenomena are constrained by lack of quantitative data. The paucity of measurements of in vivo rates or parameters associated with dynamics means that it is difficult to understand systems from a dynamic, rather than a static, point of view. For example, to further the use of biological modeling and simulation, kinetics should be an important part of early biological courses, including biochemistry and molecular biology, to instill an appreciation in experimental biologists that kinetics is important. The requisite background in quantitative methods is likely to include some nontrivial exposure to continuous mathematics, nonlinear dynamics, linear algebra, probability and statistics, as well as computer programming and algorithm design.

From the engineering side, few nonbiologists get any exposure to biological laboratory research or develop an understanding of the collection and analysis of biological data. This also leads to unrealistic expectations of what can be done practically, how repeatable (or unrepeatable) a set of experiments can be, and how difficult it can be to understand the system in detail. Computer scientists also require exposure to probability, statistics, laboratory technique, and experimental design in order to understand the biologist's empirical methodology. More fundamentally, nonbiologists working at the BioComp interface must have an understanding of the basic principles relevant to the biological problem domains of interest, such as physiology, phylogeny, or proteomics. (A broad perspective on biology, including some exposure to evolution, ecosystems, and metabolism, is certainly desirable, but is likely not absolutely necessary.)

Finally, it must be noted that many students choose to study biology because it is a science whose study has traditionally not involved mathematics to any significant extent. Similarly, W. Daniel Hillis

[3] R. Altman, "A Curriculum for Bioinformatics: The Time Is Ripe," *Bioinformatics* 14(7):549-550, 1998.

has noted that "biologists are biologists because they love living things. A computation is not alive."[4] Indeed, this has been true for several generations of students, so that many of these same students are now incumbent instructors of biology. Managing this particular problem will pose many challenges.

10.2.2.2 Undergraduate Programs

The primary rationale for undergraduate programs at the BioComp interface is that the undergraduate years of university education in the sciences carry the greatest burden in teaching a student the professional language of a science and the intellectual paradigms underlying the practice of that science. The term "paradigm" is used here in the original sense first expounded by Kuhn, which includes the following:[5]

- Symbolic generalizations, which the community uses without question,
- Beliefs in particular models, which help to determine what will be accepted as an explanation or a solution,
- Values concerning prediction (e.g., predictions must be accurate, quantitative) and theories (e.g., theories must be simple, self-consistent, plausible, compatible with other theories in current use), and
- Exemplars, which are the concrete problem solutions that students encounter from the start of their scientific education.

The description in Section 10.3.1 suggests that the disciplinary paradigm of biology is significantly different from that of computer science. Because the de novo learning of one paradigm is easier than subsequently learning a second paradigm that may (apparently) be contradictory or incommensurate with one that has already been internalized, the argument for undergraduate exposure is based on the premise that simultaneous exposure to the paradigms of two disciplines will be more effective than sequential exposure (as would be the case for someone receiving an undergraduate degree in one field and then pursuing graduate work in another).

Undergraduate programs in most scientific courses of study are generally designed to prepare students for future academic work in the field. Thus, the goal of undergraduate curricula at the BioComp interface is to expose students to a wide range of biological knowledge and issues and to the intellectual tools and constructs of computing such as programming, statistics, algorithm design, and databases. Today, most such programs focus on bioinformatics or computational biology, and in the most typical cases, the integration of biology and computing occurs later rather than earlier in these programs (e.g., as senior-year capstone courses).

Individual programs vary enormously in the number of computer science classes required. For example, the George Washington University Department of Computer Science offers a concentration in bioinformatics leading to a B.S. degree; the curriculum includes 17 computer science courses and 4 biology courses, plus a single course on bioinformatics. The University of California, Los Angeles (UCLA) program in cybernetics offers a concentration in bioinformatics, in contrast, in which the students can take as few as seven computer science courses, including four programming classes and two biology-themed classes. In other cases, a university may have an explicit undergraduate major in bioinformatics associated with a bioinformatics department. Such programs are traditionally structured in the sense of having a set of specific courses required for matriculation.

In addition to concentrations at the interface, a number of other approaches have been used to prepare undergraduates:

- An explicitly interdisciplinary B.S. science program can expose students to the interrelationships of the basic sciences. Sometimes these are co-taught as single units: students in their first year may take

[4] W.D. Hillis, "Why Physicists Like Models, and Biologists Should," *Current Biology* 3(2):79-81, 1993.

[5] T.S. Kuhn, *The Structure of Scientific Revolutions*, Third Edition, University of Chicago Press, Chicago, 1996.

mathematics, physics, chemistry, and biology as a block, taught by a team of dedicated professors. Modules in which ideas from one discipline are used to solve problems in another are developed and used as case studies for motivating the connections between the topics. Other coordinated science programs intersperse traditional courses in the disciplines with co-taught interdisciplinary courses (Examples: Applications of Physical Ideas in Biological Systems; Dimensional Analysis in the Sciences; Mathematical Biology).

- A broad and unrestricted science program can allow students to count basic courses in any department toward their degree or to design and propose their personal degree program. Such a system gives graduates an edge in the ability to transcend boundaries between disciplines. A system of co-advising to help students balance needs with interests would be vital to ensure that such open programs function well.
- Courses in quantitative science with explicit ties to biology may be more motivating to biology students. Some anecdotal evidence indicates that biology students can do better in math and physics when the examples are drawn from biology; at the University of Washington, the average biology student's grade in calculus rose from C to B+ when "calculus for biologists" was introduced.[6] (Note that such an approach requires that the instructor have the knowledge to use plausible biological examples—a point suggesting that simply handing off modules of instruction will not be successful.)
- Summer programs for undergraduates offer undergraduates an opportunity to get involved in actual research projects while being exposed to workshops and tutorials in a range of issues at the BioComp interface. Many such programs are funded by a National Science Foundation or National Institutes of Health program.[7]

When none of these options are available, a student can still create a program informally (either on his or her initiative or with the advice and support of a sympathetic faculty member). Such a program would necessarily include courses sufficient to impart a thorough quantitative background (mathematics, physics, computer science) as well as a solid understanding of biology. As a rule, quantitative training should come first, because it is often difficult to develop expertise in quantitative approaches later in the undergraduate years. Exposure to intriguing ideas in biology (e.g., in popular lecture series) would also help to encourage interest in these directions.

Finally, an important issue at some universities is the fact that computer science departments and biology departments are located in different schools (school of engineering versus school of arts and sciences). As a result, biology majors may well face impediments to enrolling in courses intended for computer science majors, and vice versa. Such a structural impediment underlines both the need and the challenges for establishing a biological computing curriculum.

10.2.2.3 The BIO2010 Report

In July 2003, the National Research Council (NRC) released *Bio 2010: Undergraduate Education to Prepare Biomedical Research Scientists* (National Academies Press, Washington, DC). This report concluded that undergraduate biology education had not kept pace with computationally driven changes in life sciences research, among other changes, and recommended that mathematics, physics, chemistry, computer science, and engineering be incorporated into the biology curriculum to the point that interdisciplinary thinking and work become second nature for biology students. In particular, the report noted "the importance of building a strong foundation in mathematics, physical and information sciences to prepare students for research that is increasingly interdisciplinary in character."

The report elaborated on this point in three other recommendations—that undergraduate life sci-

[6] Mary Lidstrom, University of Washington, personal communication, August 1, 2003.
[7] See http://www.nsf.gov/pubs/2002/nsf02109/nsf02109.htm.

ence majors should be exposed to engineering principles and analysis, should receive quantitative training in a manner integrated with biological content, and should develop enough familiarity with computer science that they can use information technology effectively in all aspects of their research.

10.2.2.3.1 Engineering In arguing for exposure to engineering, the report noted that the notion of function (of a device or organism) is common to both engineering and biology, but not to mathematics, physics, or chemistry. Echoing the ideas described in Chapter 6 of this report, BIO2010 concluded:

> Understanding function at the systems level requires a way of thinking that is common to many engineers. An engineer takes building blocks to build a system with desired features (bottom-up). Creating (or re-creating) function by building a complex system, and getting it to work, is the ultimate proof that all essential building blocks and how they work in synchrony are truly understood. Getting a system to work typically requires (a) an understanding of the fundamental building blocks, (b) knowledge of the relation between the building blocks, (c) the system's design, or how its components fit together in a productive way, (d) system modeling, (e) construction of the system, and (f) testing the system and its function(s). . . . Organisms can be analyzed in terms of subsystems having particular functions. To understand system function in biology in a predictive and quantitative fashion, it is necessary to describe and model how the system function results from the properties of its constituent elements.

The pedagogical conclusion was clear in the report:

> Understanding cells, organs, and finally animals and plants at the systems level will require that the biologist borrow approaches from engineering, and that engineering principles are introduced early in the education of biologists. . . . Students should be frequently confronted throughout their biology curriculum with questions and tasks such as how they would design 'xxx,' and how they would test to see whether their conceptual design actually works. [For example,] they should be asked to simulate their system, determine its rate constants, determine regimes of stability and instability, investigate regulatory feedback mechanisms, and other challenges.

A second dimension in which engineering skills can be useful is in logistical planning. There are many areas in biology now where it is relatively easy to conceive of an important experiment, but drawing out the implications of the experiment involves a combinatorial explosion of analytical effort and thus is not practical to carry out. It is entirely plausible that many important biological discoveries will depend on both the ability to conceive an experiment and the ability to reconceive and restructure it logistically so that it is, in fact, doable. Engineers learn to apply their fundamental scientific knowledge in an environment constrained by nonscientific concerns, such as cost or logistics, and this ability will be critically important for the biologist who must undertake the restructuring described above. Box 10.1 provides a number of examples of engineering for life science majors.

10.2.2.3.2 Quantitative Training In its call for greater quantitative training, the BIO2010 report echoed that of other commentators.[8] Recognizing that quantitative analysis, modeling, and prediction play important roles in today's biomedical research (and will do so increasingly in the future), the report noted the importance to biology students of understanding concepts such as rate of change, modeling, equilibrium, and stability, structure of a system, and interactions among components, and argued that every student should acquire the ability to analyze issues arising in these contexts in some depth, using analytical methods (including paper-and-pencil techniques) and appropriate computational tools. As part of a necessary background, the report suggested that an appropriate course of study would include aspects of probability, statistics, discrete models, linear algebra, calculus and differential equations, modeling, and programming (Box 10.2).

[8] See for example, A. Hastings and M.A. Palmer, "A Bright Future for Biologists and Mathematicians," *Science* 299(5615):2003-2004, 2003, available at http://www.biosino.org/bioinformatics/a%20bright%20future.pdf.

Box 10.1
Engineering for Life Science Majors

One example of an engineering topic suitable for inclusion in a biology curriculum is the subject of long-range neuron signals. Introducing such a topic might begin with the electrical conductivity of salt water and of the lipid cell membrane, and the electrical capacitance of the cell membrane. It would next develop the simple equations for the attenuation of a voltage applied across the membrane at one end of an axon "cylinder" with distance down the axon, and the effect of membrane capacitance on signal dynamics for time-varying signals.

After substituting numbers, it becomes clear that amplifiers will be essential. On the other hand, real systems are always noisy and imperfect; amplifiers have limited dynamical range; and the combination of these facts makes sending an analog voltage signal through a large number of amplifiers essentially impossible.

The pulse coding of information overcomes the limitations of analog communication. How are "pulses" generated by a cell? This would lead to the power supply needed by an amplifier—ion pumps and the Nernst potential. How are action potentials generated? A first example of the transduction of an analog quantity into pulses might be stick-slip fraction, in which a block resting on a table, and pulled by a weak spring whose end is steadily moved, moves in "jumps" whose distance is always the same. This introduction to nonlinear dynamics contains the essence of how an action potential is generated.

The "negative resistance" of the sodium channels in a neuron membrane provides the same kind of "breakdown" phenomenon. Stability and instabilities (static and dynamic) of nonlinear dynamical systems can be analyzed, and finally the Hodgkin-Huxley equations illustrated.

The material is an excellent source of imaginative laboratories involving electrical measurements, circuits, dynamical systems, batteries and the Nernst potential, information and noise, and classical mechanics. It has great potential for simulations of systems a little too complicated for complete mathematical analysis, and thus is ideal for teaching simulation as a tool for understanding.

Other biological phenomena that can be analyzed using an engineering approach and that are suitable for inclusion in a biology curriculum include the following:

- The blood circulatory system and its control; fluid dynamics; pressure and force balance;
- Swimming, flying, walking, dynamical description, energy requirements, actuators, control; material properties of biological systems and how their structure relates to their function (e.g., wood, hair, cell membrane cartilage);
- Shapes of cells: force balance, hydrostatic pressure, elasticity of membrane and effects of the spatial dependence of elasticity; effects of cytoskeletal force on shape; and
- Chemical networks for cell signaling; these involve the concepts of negative feedback, gain, signal-to-noise, bandwidth, and cross-talk. These concepts are simple to experience in the context of how an electrical amplifier can be built from components.

SOURCE: Adapted from National Research Council, *BIO2010: Transforming Undergraduate Education for Future Research Biologists*, The National Academies Press, Washington, DC, 2003.

10.2.2.3.3 Computer Science Finally, the BIO2010 report noted the importance of information technology-based tools for biologists. It recommended that all biology majors be able to develop simulations of physiological, ecological, and evolutionary processes; to modify existing applications as appropriate; to use computers to acquire and process data; to carry out statistical characterization of the data and perform statistical tests; to graphically display data in a variety of representation; and to use information technology (IT) to carry out literature searches, locate published articles, and access major data-

Box 10.2
Essential Concepts of Mathematics and Computer Science for Life Scientists

Calculus

- Complex numbers
- Functions
- Limits
- Continuity
- The integral
- The derivative and linearization
- Elementary functions
- Fourier series
- Multidimensional calculus: linear approximations, integration over multiple variables

Linear Algebra

- Scalars, vectors, matrices
- Linear transformations
- Eigenvalues and eigenvectors
- Invariant subspaces

Dynamical Systems

- Continuous time dynamics—equations of motion and their trajectories
- Test points, limit cycles, and stability around them
- Phase plane analysis
- Cooperativity, positive feedback, and negative feedback
- Multistability
- Discrete time dynamics—mappings, stable points, and stable cycles
- Sensitivity to initial conditions and chaos

Probability and Statistics

- Probability distributions
- Random numbers and stochastic processes
- Covariation, correlation, and independence
- Error likelihood

Information and Computation

- Algorithms (with examples)
- Computability
- Optimization in mathematics and computation
- "Bits": information and mutual information

Data Structures

- Metrics: generalized "distance" and sequence comparisons
- Clustering
- Tree relationships
- Graphics: visualizing and displaying data and models for conceptual understanding

SOURCE: Reprinted from National Research Council, *BIO2010: Transforming Undergraduate Education for Future Research Biologists*, The National Academies Press, Washington, DC, 2003.

bases. From the perspective of this report, Box 10.3 describes some of the essential intellectual aspects of computer science that biologists must understand.

Recognizing that students might require competence at multiple levels depending on their needs, the BIO2010 report identified three levels of competence as described in Box 10.4.

Box 10.3
Essential Concepts of Computer Science for the Biologist

Key for the computer scientist is the notion of a field that focuses on information, on understanding of computing activities through mathematical and engineering models and based on theory and abstraction, on the ways of representing and processing information, and on the application of scientific principles and methodologies to the development and maintenance of computer systems—whether they are composed of hardware, software, or both.

There are many views of understanding the essential concepts of computer science. One view, developed in 1991 in the NRC report *Computing the Future*, is that the key intellectual themes in computing are algorithmic thinking, the representation of information, and computer programs.[1]

- An algorithm is an unambiguous sequence of steps for processing information. Of particular relevance is how the speed of the algorithm varies as a function of problem size—the topic of algorithmic complexity. Typically, a result from algorithmic complexity will indicate the scaling relationships between how long it takes to solve a problem and the size of the problem when the solution of the problem is based on a specific algorithm. Thus, algorithm A might solve a problem in a time of order N^2, which means that a problem that is 100 times as large would take $100^2 = 10,000$ times as long to solve, whereas a faster algorithm B might solve the same problem in time of order $N \ln N$, which means a problem 100 times as large would take $100 \ln 100 = 460.5$ times as long to solve. Such results are important because all computer programs embed algorithms within them. Depending on the functional relationship between run time and problem size, a given program that works well on a small set of test data may—or may not—work well (run in a reasonable time) for a larger set of real data. Theoretical computer science thus imposes constraints on real programs that software developers ignore at their own peril.
- The representation of information or a problem in an appropriate manner is often the first step in designing an algorithm, and the choice of one representation or another can make a problem easy or difficult, and its solution slow or fast. Two issues arise: (1) how should the abstraction be represented, and (2) how should the representation be structured properly to allow efficient access for common operations? For example, a circle of radius 2 can be represented by an equation of the form $x^2 + y^2 = 4$ or as a set of points on the circle ((0.00, 2.00), (0.25, 1.98), (0.50, 1.94), (0.75, 1.85), (1.00, 1.73), (1.25, 1.56), (1.50, 1.32), (1.75, 0.97), (2.00, 0.00)), and so on. Depending on the purpose, one or the other of these representations may be more useful. If the circle of radius 2 is just a special case of a problem in which circles of many different radii are involved, representation as an equation may be more appropriate. If many circles of radius 2 have to be drawn on a screen and speed is important, a listing of the points on the circle may provide a faster basis for drawing such circles.
- A computer program expresses algorithms and structure information using a "programming language." Such languages provide a way to represent an algorithm precisely enough that a "high-level" description (i.e., one that is easily understood by humans) can be translated mechanically ("compiled") into a "low-level" version that the computer can carry out ("execute"); the execution of a program by a computer is what allows the algorithm to be realized tangibly, instructing the computer to perform the tasks the person has requested. Computer programs are thus the essential link between intellectual constructs such as algorithms and information representations and the computers that perform useful tasks.

[1] The discussion below is adapted from Computer Science and Telecommunications Board, National Research Council, *Computing the Future: A Broader Agenda for Computer Science and Engineering*, National Academy Press, Washington, DC, 1992.

continued

Box 10.3 Continued

This last point is often misunderstood. For many outsiders, computer science is the same as computer programming—a view reinforced by many introductory "computer science" courses that emphasize the writing of computer programs. But it is better to understand computer programs as the specialized medium in which the ideas and abstractions of computer science are tangibly manifested. Focusing on the writing of the computer program without giving careful consideration to the abstractions embodied in the program is not unlike understanding the writing of a novel as no more than the rules of grammar and spelling.

Algorithmic thinking, information representation, and computer programs are themes central to all subfields of computer science and engineering research. They also provide material for intellectual study in and of themselves, often with important practical results. The study of algorithms is as challenging as any area of mathematics, and one of practical importance as well, since improperly chosen or designed algorithms may solve problems in a highly inefficient manner. The study of programs is a broad area, ranging from the highly formal study of mathematically proving programs correct to very practical considerations regarding tools with which to specify, write, debug, maintain, and modify very large software systems (otherwise called software engineering). Information representation is the central theme underlying the study of data structures (how information can best be represented for computer processing) and much of human-computer interaction (how information can best be represented to maximize its utility for human beings).

Finally, computer science is closely tied to an underlying technological substrate that evolves rapidly. This substrate is the "stuff" out of which computational hardware is made, and the exponential growth that characterizes its evolution makes it possible to construct ever-larger, ever-more-complex systems—systems that are not predictable based on an understanding of their individual components. (As one example, the properties of the Internet prove a rich and surprisingly complex area of study even though its components—computers, routers, fiber-optic cables—are themselves well understood.)

A second report of the National Research Council described fluency with information technology as requiring three kinds of knowledge: skills in using contemporary IT, foundational concepts about IT and computing, and intellectual capabilities needed to think about and use IT for purposeful work.[2] The listing below is the perspective of this report on essential concepts of IT for everyone:

- *Computers* (e.g., programs as a sequence of steps, memory as a repository for program and data, overall organization, including relationship to peripheral devices).
- *Information systems* (e.g., hardware and software components, people and processes, interfaces (both technology interfaces and human-computer interfaces), databases, transactions, consistency, availability, persistent storage, archiving, audit trails, security and privacy and their technological underpinnings).
- *Networks:* physical structure (messages, packets, switching, routing, addressing, congestion, local area networks, wide area networks, bandwidth, latency, point-to-point communication, multicast, broadcast, Ethernet, mobility), and logical structure (client/server, interfaces, layered protocols, standards, network services).
- *Digital representation of information:* concept of information encoding in binary form; different information encodings such as ASCII, digital sound, images, and video/movies; precision, conversion and interoperability (e.g., of file formats), resolution, fidelity, transformation, compression, and encryption; standardization of representations to support communication.
- *Information organization* (including forms, structure, classification and indexing, searching and retrieving, assessing information quality, authoring and presentation, and citation; search engines for text, images, video, audio).
- *Modeling and abstraction:* methods and techniques for representing real-world phenomena as computer models, first in appropriate forms such as systems of equations, graphs, and relationships, and then in appropriate programming objects such as arrays or lists or procedures. Topics include continuous and discrete

[2] Computer Science and Telecommunications Board, National Research Council, *Being Fluent with Information Technology*, National Academy Press, Washington, DC, 1999.

models, discrete time events, randomization, and convergence, as well as the use of abstraction to hide irrelevant detail.

- *Algorithmic thinking and programming:* concepts of algorithmic thinking, including functional decomposition, repetition (iteration and/or recursion), basic data organization (record, array, list), generalization and parameterization, algorithm vs. program, top-down design, and refinement.
- *Universality and computability:* ability of any computer to perform any computational task.
- *Limitations of information technology:* notions of complexity, growth rates, scale, tractability, decidability, and state explosion combine to express some of the limitations of information technology; connections to applications, such as text search, sorting, scheduling, and debugging.
- *Societal impact of information and information technology:* technical basis for social concerns about privacy, intellectual property, ownership, security, weak/strong encryption, inferences about personal characteristics based on electronic behavior such as monitoring Web sites visited, "netiquette," "spamming," and free speech in the Internet environment.

A third perspective is provided by Steven Salzberg, senior director of bioinformatics at the Institute for Genomic Research in Rockville, Maryland. In a tutorial paper for biologists, he lists the following areas as important for biologists to understand:[3]

- Basic computational concepts (algorithms, program execution speed, computing time and space requirements as a function of input size; really expensive computations),
- Machine learning concepts (learning from data, memory-based reasoning),
- Where to store learned knowledge (decision trees, neural networks),
- Search (defining a search space, search space size, tree-based search),
- Dynamic programming, and
- Basic statistics and Markov chains.

[3] S.L. Salzberg, "A Tutorial Introduction to Computation for Biologists," *Computational Methods in Molecular Biology,* S.L. Salzberg, D. Searls, and S. Kasif, eds., Elsevier Science Ltd., New York, 1998.

10.2.2.4 Graduate Programs

Graduate programs at the BioComp interface are often intended to provide B.S. graduates in one discipline with the complementary expertise of the other. For example, individuals with bachelor's degrees in biology may acquire computational or analytical skills during early graduate school, with condensed "retraining" programs that expose then to nonlinear dynamics, algorithms, and so on. Alternatively, individuals with bachelor's degrees in computer science might take a number of courses to expose them to essential biological concepts and techniques.

Graduate education at the interface is much more diverse than at the undergraduate level. Although there is general agreement that an undergraduate degree should expose the student to the component sciences and prepare him or her for future work, the graduate degree involves a far wider array of goals, focuses, fields, and approaches. Like undergraduate programs, graduate programs can be stand-alone departments, independent interdisciplinary programs, or certificate programs that require students to have a "home" department.

A bioinformatics program oriented toward genomics is very common. Virginia Tech's program, for example, has been renamed the program in "Genetics, Bioinformatics, and Computational Biology," indicating its strong focus on genetic analysis. In contrast, the Keck Graduate Institute at Claremont stresses the interdisciplinary skill set necessary for the effective management of companies that straddle the biology-quantitative science boundary. It awards a master's of bioscience, a professional degree

Box 10.4
Competence and Expertise in Computer Science for Biology Students

The BIO2010 report recommended that all biology students receive instruction in computer science, distinguishing among three levels of competency. From lowest to highest, these include the following:

- *Fluency*. Based on the NRC report *Being Fluent with Information Technology,* fluency refers to the ability of biology students to use information technology today and to adapt to changes in IT in the future. For example, they need a basic understanding of how computers work and of programming, and a higher degree of fluency in using networks and databases. Students should also be exposed to laboratory experiences using MEDLINE, GenBank, and other biological databases, as well as physiological and ecological simulations. For example, students could be asked to use computer searches to track down all known information about a given gene and the protein it encodes, including both structure and function. This would involve exploring the internal structure of the gene (exons, introns, promoter, transcription factor binding sites); the regulatory control of the gene; sequence homologues of the gene and the protein; the structure and function of the protein; gene interaction networks and metabolic pathways involving the protein; and interactions of the protein with other proteins and with small molecules.
- *Capability in program design for computational biology and genomics applications*. Students at this level acquire the minimal skills required to be effective computer users within a computationally oriented biology research team. For example, they would learn structured software development and selected principles of computer science, with applications in computational biology and allied disciplines, and would use examples and tutorials drawn from problems in computational biology.
- *Capability in developing software tools for use by the biology community*. At this sophisticated level, students need a grounding in discrete mathematics, data structures, and algorithms, as well as database management systems, information systems, software engineering, computer graphics, or computer simulation techniques. Students at this level would be able to design and specify database and information systems for use by the entire community. Of special interest will be tools that require background in graph theory, combinatorics, and computational geometry as applications in high-throughput genomics research and rational drug design become increasingly important.

SOURCE: Adapted from National Research Council, *BIO2010: Transforming Undergraduate Education for Future Research Biologists*, The National Academies Press, Washington, DC, 2003.

somewhat like an M.B.A. with a science requirement. Some programs, such as Stanford's, are administered by the medical school, leading to a focus on medical informatics as well as bioinformatics. This would include topics such as clinical trials and image analysis, which would not show up in a more traditional genomics-focused bioinformatics degree.

The Research Training Program of the Keck Center for Computational and Structural Biology is intended to develop one of two different kinds of expertise. Emerging from this program, a trainee would be a computational expert well versed in computer science and quantitative methods who would also be knowledgeable in at least one application area of biological significance, or an expert in some biological area (e.g., molecular biology) who would also be aware of the most advanced concepts in computing. Students entering from computational backgrounds take at least three courses in biology-biochemistry-biophysics areas, while students entering from biological backgrounds at least three courses in computational areas. In addition, all students take an introductory course in computational science. Dissertation research is supervised by a committee with faculty members as required by the student's home department, but with representation from the computational biology faculty at other Keck Center institutions as well. Research can be undertaken in areas including the visualization of biological complexes, the development of DNA and protein sequence analysis, and advanced simulations.

A challenge for a field as interdisciplinary as this is that incoming students will arrive with possibly completely non-overlapping backgrounds. Most programs accept a B.S. in computer science, biology, or math as a prerequisite; to produce a well-rounded computational biologist will require very different training programs. The University of Colorado's certificate program in computational biology requires incoming students to take preparatory classes in "Biology for Computer Scientists," "Computer Science for Bioscientists," or "Mathematics for Bioscientists," depending on what the student missed earlier in his or her education.

An advantage of graduate programs is that when communication among faculty of different disciplines is good, graduate projects provide an ideal opportunity for students to work in an interdisciplinary environment. In some cases, work with adjunct professors from industry can lead to exciting projects. On the other hand, if communication between faculty is poor (which may be possible for reasons described later in this chapter), a graduate student dependent on completing a project (e.g., a dissertation) can get caught in the middle of a dispute with no way to graduate.

10.2.2.5 Postdoctoral Programs

Postdoctoral programs at the BioComp interface are also varied. Some postgraduate programs are explicitly aimed at "conversion," that is, training a fully trained member of one field (usually biology) in the basic tenets of its complement. For example, the University of Pennsylvania's postdoctoral program in computational biology is a master's degree in computer and information systems, designed for those with Ph.D.s in biology who need the training. Other programs focus on involving the participant in research and laboratory work, in preparation for industry or a faculty position, just as in postdoctoral programs in other fields.

Some programs, such as Duke's Center for Bioinformatics and Computational Biology, are similar to graduate programs in that they focus on genome analysis. Others, like the Johns Hopkins' program in computational biology, are firmly grounded in genomics but are pointedly reaching out to larger questions of integrative biology and experimental biology.

In promoting postdoctoral programs at the interface of computing and biology, it will be necessary to take into account the very different traditions of the two fields. In biology, one or more postdoctoral fellowships are quite common (indeed, routine) before an individual strikes out on his or her own. By contrast, the most typical career path for a newly graduated Ph.D. in computer science calls for appointment to a junior faculty position or a position in industry—postdoctoral fellows in computer science are relatively rare (though not unheard of).

Two foundation-supported postdoctoral programs have been influential in stimulating interest at the BioComp interface: the Sloan-Department of Energy (DOE) program and the Burroughs-Welcome program.

10.2.2.5.1 The Sloan-DOE Postdoctoral Awards for Computational Molecular Biology[9] For 8 years, the Alfred P. Sloan Foundation and the U.S. Department of Energy (Office of Biological and Environmental Research) have jointly sponsored postdoctoral research awards for scientists interested in computational molecular biology. The purpose of these fellowships has been to catalyze career transitions into computational molecular biology by those holding doctorates in mathematics, physics, computer science, chemistry, engineering or other relevant fields who would like to bring their computational sophistication to bear on the complex problems that increasingly face molecular biology.

Operationally, the program was designed to offer computationally sophisticated young scientists an intensive postdoctoral opportunity in an appropriate molecular biology laboratory. In most cases, awardees had strong educational backgrounds in a computationally intensive field, although in a few

[9] See http://www.sloan.org/programs/scitech_postdoct.shtml.

instances, awardees had backgrounds from more traditional biological orientations without the computational dimension. Of particular interest to the Sloan-DOE program are important problems in structural biology and genome analysis, including analysis of protein and nucleic acid sequence, protein and nucleic acid structure, genome structure and maps, cross-species genome analysis, multigenic traits, and structure-function relationships where the structures are from genomes, genes, or gene products.

The Sloan-DOE postdoctoral award supports up to 2 years of research in an appropriate molecular biology department or laboratory in the United States or Canada selected by the awardee. In magnitude, the award provides for a total budget of $120,000 (including indirect and overhead costs), spread over a grant period of 2 years.

10.2.2.5.2 The Burroughs-Wellcome Career Awards at the Scientific Interface[10] The Burroughs-Wellcome Career Awards at the Scientific Interface are intended to foster the early career development of researchers with backgrounds in the physical and computational sciences whose work addresses biological questions and who are dedicated to pursuing a career in academic research.[11] Prospective awardees are expected to have Ph.D.-level training in a scientific field other than biology and are encouraged to describe potential collaborations with well-established investigators working on interface problems of interest.

The program provides $500,000 over 5 years to support up to 2 years of advanced postdoctoral training and the first 3 years of a faculty appointment. In general, an awardee is expected to accept a faculty position at an institution other than the one supporting the postdoc, a requirement that is likely to spread the philosophy of interface research embodied in the program more effectively than the publishing of papers or program descriptions.

In addition, the Burroughs-Welcome Fund (BWF) requires the faculty-hiring institution to make a significant commitment to the award recipient's career development, where "significant commitment" is demonstrated by the financial and professional situation offered. Tenure-track faculty appointments are strongly preferred, accompanied by salary support and/or support for starting up a laboratory. Awardees are required to devote at least 80 percent of their time to research-related activities. Furthermore, the faculty-hiring institution must offer the awardee to take an adjunct appointment in a second department and name at least one tenured faculty member in a discipline complementary to the awardee's primary discipline who is willing to serve as an active collaborator.

10.2.2.5.3 Keck Center for Computational and Structural Biology: The Research Training Program The W.M. Keck Center for Computational and Structural Biology is an interdisciplinary and interinstitutional organization, including Baylor College of Medicine, the University of Houston, Rice University, University of Texas Health Science Center, the M.D. Anderson Cancer Center, and University of Texas Medical Branch at Galveston. Subareas of focus include computational methods and tools, biomolecular structure and function, imaging and dynamics, mathematical modeling of biosystems, and medical and genomic informatics. The faculty include some 130 members, drawn from member institutions, and a

[10] See http://www.bwfund.org/programs/interfaces/career_awards_background.html.

[11] A previous Burroughs-Wellcome Fund (BWF) program, known as Institutional Awards at the Scientific Interface, has been discontinued. (Together with the Career Awards program, it constituted the BWF Interfaces in Science effort.) The purpose of the Institutional Awards program was to support U.S. and Canadian academic institutions in developing interdisciplinary graduate and postdoctoral training programs for individuals with backgrounds in the physical, computational, or mathematical sciences to pursue biological questions. For example, pre- and postdoctoral fellows at the La Jolla Consortium and the University of Chicago's Institute for Biophysical Dynamics had to propose research projects that required the participation of two mentors—one from the quantitative sciences and one from the biological sciences—before being awarded financial support. For more on the Institutional Awards program, see N.S. Sung, J.I. Gordon, G.D. Rose, E.D. Getzoff, S.J. Kron, D. Mumford, J.N. Onuchic, et al., "Science Education: Educating Future Scientists," *Science* 301(5639):1485, 2003, available at http://www.bwfund.org/programs/interfaces/institutional_main.html.

few dozen predoctoral and postdoctoral fellows. The Keck Center was established in 1990 by a $5 million grant from the Keck foundation and currently receives more than $20 million annually in grants, from agencies such as the National Institutes of Health, the National Science Foundation, the Department of Defense, and private sources.

The Keck Center's training program, supported by the W.M. Keck Foundation and the National Library of Medicine, seeks to cross-train new scientists in both computational science and a specialized area of biology so that they can shed new light on the cellular and molecular basis of biological processes.[12] Fellowships are supported for research in algorithm development, advanced computational methods, biomedicine, crystallography, electron cryomicroscopy and computer reconstruction, genome studies, imaging and visualization, mathematical modeling of biosystems, medical informatics, neuroscience, protein dynamics and design, robotics applications in molecular biology, and the structure and function of biomolecules. The fellowship provides trainees with cross-training in computational science and in biological applications, dual mentorship, and access to cutting-edge facilities.

10.2.2.6 Faculty Retraining in Midcareer

Faculty training or retraining can augment the above opportunities. In some cases, this means participation in workshops (given release time to allow for this investment), sabbaticals spent learning a new subject, or explicitly switching from one field to another. As a rule, funded release time will be necessary to provide a break from academic constraints and to offer the time and opportunity to see biological work up close. In some cases, a good way to develop cross-disciplinary expertise is to spend a sabbatical year in the laboratory of a colleague in another discipline.

The committee was unable to find programs specifically oriented toward retraining computer scientists to do biological research. However, the National Science Foundation (NSF) does support the Interdisciplinary Grants in the Mathematical Sciences program through its Mathematical and Physical Sciences Directorate whose objective is "to enable mathematical scientists to undertake research and study in another discipline so as to expand their skills and knowledge in areas other than the mathematical sciences, subsequently apply this knowledge in their research, and enrich the educational experiences and broaden the career options of their students."[13] Recipients spend a year full-time (in a 12-month period) in a nonmathematical academic science department or in an industrial, commercial, or financial institution, and the outcome is expected to be sufficient familiarity with another discipline on the part of the supported individual "to open opportunities for effective collaboration by the mathematical scientist with researchers in another discipline." Applicants must have a tenured or tenure-track academic appointment, and the proposal must include a co-principal investigator at the level of dean (or higher-level university official) at the submitting institution as well as a commitment from the host institution or department that the hosted individual will be treated as a regular faculty member within the host unit and that at least one senior person will be provided who will serve as institutional host.

In addition, the National Institutes of Health (NIH's) National Research Service Awards program for Senior Fellows (F33) supports scientists from any field with 7 or more years of postdoctoral research experience who wish to make major changes in the direction of their research careers or who wish to broaden their scientific background by acquiring new research capabilities. In most cases, these awards are used to support sabbatical experiences for established independent scientists in which they receive training to increase their scientific capabilities. Such training must be within the scope of biomedical, behavioral, or clinical research and must offer an opportunity for individuals to broaden their scientific background or extend their potential for research in health-related areas. The maximum annual stipend is considerably lower than senior scientists typically receive, but most awardees find supplements so that they may obtain their full salaries while pursuing studies in a new field. The guidelines for eligibil-

[12] See http://cohesion.rice.edu/centersandinst/keckcenter/training.cfm?doc_id=2368.
[13] See http://www.nsf.gov/pubs/2001/nsf01115/nsf01115.htm.

ity specifically do not include previous experience in biomedical research; and thus, computer scientists would be eligible for such a program.[14]

10.2.3 Academic Organizations

Typically, academic research is conducted in departments or in centers that draw on faculty from multiple departments. The descriptions of the three departments below are simply illustrative and not exhaustive (no inference should be drawn from the fact that any given department or center is not included below):

- Cornell University maintains four distinct programs in computational biology, three hosted by a parent discipline. Biological sciences, computer science, and mathematics all offer concentrations in computational biology (the math department calls it "mathematical biology"). The only stand-alone department is the Department of Biological Statistics and Computational Biology (BSCB), a part of the College of Agriculture and Life Sciences. BSCB was originally the Department of Biometry and Biostatistics, and in 2005, it has six tenure-track faculty (plus two emeritus professors), one nontenure-track lecturer, and four "adjunct" faculty. There are 2 postdoctoral associates, 26 graduate students, and 65-70 undergraduate students. The department focuses mainly on biological statistics, computational biology, and statistical genomics. Research interests of the faculty include statistical genomics, Bayesian statistics, population genetics, epidemiology, modeling, molecular evolution, and experiment design.
- The University of California at Santa Cruz has a Department of Biomolecular Engineering, an interdisciplinary department that contains research programs in bioinformatics and experimental systems biology, among others. The bioinformatics program was originally administered by the computer engineering department. In 2005, the program has nine core tenure-track faculty members, and one affiliated faculty member. The bioinformatics curriculum includes a core of bioethics, Bayesian statistics, molecular biology, biochemistry, computational analysis of proteins, and computational genomics. Electives are drawn from biology, chemistry, computer science, and applied mathematics and statistics.
- Carnegie Mellon University (CMU) has offered programs in computational biology (through its computer science, biology, mathematics, physics, and chemistry departments) since 1989. In 2005, CMU's Department of Biological Sciences had 5 faculty involved in both computational biology and bioinformatics and genomics, proteomics, and systems biology. The department offers a B.S. in computational biology, which consists largely of a traditional biological curriculum augmented with math, programming, and computer science classes. In addition, students (B.S. or Ph.D.) can participate in the interdepartmental Merck Computational Biology and Chemistry Program, which requires students to have a home department in biology, computer science, statistics, math, or chemistry. This program was established in 1999 with a grant from the Merck Company Foundation.

Centers are often created without specific departmental affiliation because the number of departments that might plausibly contribute expertise is large. In these instances, absent a center, it is difficult to unify and coordinate research and educational activities or to convey to the outside world what the university is doing in the area. Centers are intended to be focal points for research at the BioComp interface (most often with a bioinformatics or computational biology flavor), and they usually work with departments to make new faculty appointments and provide a single point for students to learn about university programs.

Four university-based centers are described below, simply as illustrative:

[14] For more information, see http://grants.nih.gov/grants/guide/pa-files/PA-00-131.html.

- The University of California-Berkeley's Center for Integrative Genomics was founded in December 2002, supported by the Gordon and Betty Moore Foundation.[15] Its mission is to bring tools from many disciplines to bear on problems at the intersection of evolution and developmental biology. The enabling technology for new progress in this field will be acceleration of the sequencing of species genomes, and it is hoped to sequence 100 genomes of various species in the next 5 years.[16] The faculty includes 20 researchers drawn from molecular cellular biology, integrative biology, statistics, plant and microbial biology, mathematics, computer science, bioengineering, physics, paleontology, and the Lawrence Berkeley National Laboratory. The center also plans to serve an educational role, teaching or supporting the teaching of genomic science to computer science students and computer topics to biology students, as well as providing a center for graduate and postgraduate work.
- The Vanderbilt Institute for Integrative Biosystem Research and Education (VIIBRE) at Vanderbilt University (Nashville, Tennessee) was begun with an initial grant from Vanderbilt's Academic Venture Capital Fund.[17] VIIBRE has also received project-specific funding and other support from NSF, DARPA, NIH, and other institutions, enabling it to create centers of bioengineering education technologies and to begin research in cellular instrumentation and control, biomedical imaging, technology-guided therapy, biological applications of nanosystems, cellular and tissue bioengineering and biotechnology, and bioengineering education technologies. Engineers, scientists, doctors, and mathematicians conduct research for VIIBRE; more than 20 biological physics and bioengineering faculty in Vanderbilt's College of Arts and Science and the Schools of Engineering and Medicine participate in the program. VIIBRE is also developing a postdoctoral training program for physical scientists and engineers who wish to direct their careers toward the interface between biology, medicine, engineering, and the physical sciences.
- The Computational and Systems Biology Initiative (CSBi) at the Massachusetts Institute of Technology (MIT) is a campus-wide education and research program that links biologists, computer scientists, and engineers in a multidisciplinary approach to the systematic analysis of complex biological phenomena.[18] CSBi places equal emphasis on computational and experimental methods and on molecular and systems views of biological function. CSBi includes about 80 faculty members from more than 10 academic units in science, engineering, and management. Overall, membership in CSBi is self-determined, based on a self-identified interest in systems biology, and it is offered to faculty and principal investigators, postdoctoral fellows, graduate students, and research staff.
- The Institute for Biophysical Dynamics at the University of Chicago[19] is focused on interdisciplinary study of biological entities and is supported by the BWF program of Institutional Awards at the Scientific Interface. Drawing on the biological and physical science divisions of the university, the institute focuses on RNA-DNA structure, function, and regulation; protein dynamics, folding, and engineering; cytoskeleton, membranes, and organelles; hormones and cell signaling; and cell growth, death, and multicellular function. Physical scientists at the institute have expertise in macromolecular-scale manipulation via optical tweezer and chemical means; biologically relevant model systems; measurement of dynamics of macromolecules and assemblies on scales from femtoseconds to seconds; theoretical and simulation methods; soft condensed matter theory of complex and analysis of nonlinear dynamic phenomena. Part of the institute's mission is to establish cross-disciplinary training programs for students. The essential feature of the program is the placement, on a competitive basis, of predoctoral fellows with backgrounds in the physical sciences into biological science research groups, thereby

[15] See http://www.moore.org/grantees/grant_summaries_content.asp?Grantee=ucb_cig.

[16] G. Shiffrar, "New Center for Integrative Genomics to Study Major Evolutionary Changes," *College News*; see http://ls.berkeley.edu/new/02/cig.html.

[17] See http://www.vanderbilt.edu/viibre/av-goal.html and http://www.physics.vanderbilt.edu/oldpurplesite/whatshot/newsletterwinter0102.html.

[18] For more information, see http://csbi.mit.edu/whatis.

[19] For more information, see http://ibd.uchicago.edu/.

formalizing interdisciplinary connections. Fellows participate in new "translational core courses," establishing a common culture, and select an individualized program of additional coursework tailored to their research and career goals. They also take a lead role in a weekly seminar-discussion program.

Finally, in some cases centers are not associated with a specific university at all. Their purpose can be to consolidate resources on a larger scale or simply to provide a congenial intellectual home for like-minded individuals. Three nonuniversity centers are described below, again as illustrations only:

- Cold Spring Harbor Laboratory (CSHL) is a private research institution on Long Island, New York, that employs more than 800 people (300 classified as scientists) and has an annual budget of over $120 million. CSHL was established in 1889 with missions in biological research and education. In 1993, it began the annual Cold Spring Harbor Symposium on Quantitative Biology. As of 1998, it offers a Ph.D. program. Its prime research focus is on cancer biology, although it also has strong programs in plant genetics, genomics and bioinformatics, and neurobiology. In genomics, its researchers are investigating genome structure, sequencing, pattern recognition, gene expression, prediction of protein structure and function, and other related topics. A large portion of its funding comes from revenue, such as publications, intellectual property licensing, and events fees.
- The Institute for Systems Biology (ISB) is a private nonprofit institution founded in 2000 in Seattle, Washington, by Leroy Hood, Alan Aderem, and Ruedi Aebersold.[20] With a mission of applying systems biology to problems of human health such as cancer, diabetes, and diseases of the immune system, its 11 faculty members and 170 staff have expertise in fields such as immunity, proteomics, genomics, computer science, biotechnology, and biophysics. Since its founding, ISB has received its funding predominantly from federal grants, although also including private, corporate, and foundation support and industrial collaboration.[21] ISB has also spun out a number of companies to pursue commercialization opportunities around cell sorting and cancer therapies, in addition to cooperating in a multiventure capital firm-backed incubator for new biotechnology start-ups.[22] Of particular significance is the report that Hood left the University of Washington after he failed to convince it to establish a systems biology research center; he later said that he thought "the university culture and bureaucracy just could not have sufficient flexibility" to respond to the opportunity that post-Human Genome Project systems biology presented.[23]
- The Sloan-Swartz Centers for Theoretical Neurobiology were created in 1994 under the auspices of the Sloan Foundation.[24] Located at Brandeis University, California Institute of Technology, New York University, Salk Institute, and University of California, San Francisco, the Swartz Foundation also made major grants to these centers in 2000. These centers place experimentalists and theoreticians from physics, mathematics, and computer sciences in experimental brain research laboratories, where they learn about neuroscience and apply their vantage point and nontraditional skills to cooperative lines of inquiry. The centers have investigated topics such as gain fields and gain control in nerve circuits, neural coding and information theory, neural population coding and response, natural field analysis, and short-term memory.

[20] See http://www.systemsbiology.org.

[21] L. Timmerman, "Progress, Not Profit: Nonprofit Biotech Research Groups Grow in Size, Influence," *Seattle Times*, August 4, 2003.

[22] J. Cook, "Accelerator Aims to Lure, Nurture Best Ideas in Biotech," *Seattle Post-Intelligencer*, May 23, 2003.

[23] "Under Biology's Hood," *Technology Review*, September 2001, available at http://www.techreview.com/articles/01/09/qa0901.asp.

[24] See http://www.swartzneuro.org/research_a.asp and http://www.sloan.org/programs/scitech_supresearch.shtml.

10.2.4 Industry

Industrial interest in the BioComp interface is driven by the prospect of potentially very large markets in the life sciences—especially medicine. Information-enabled bioscience is further expected to create large markets for information technologies customized and adapted to the needs of life scientists—accounting in substantial measure for the interest of some large IT companies in this area. Indeed, according to the International Data Corporation, life science organizations will spend an estimated $30 billion on technology-related purchases in 2006, up from $12 billion in 2001.[25]

Life science companies (e.g., pharmaceuticals) view information technology as a (or perhaps the) key enabler for drug design and treatments that can in principle be customized to groups as small as a single individual. Consider, for example, the specific problem of finding useful organic compounds, such as drugs, to treat or reduce the effects of disease. One approach is based on the use of combinatorial methods in chemistry, genetic engineering, and high-throughput screening technology. Such an approach relies on trial-and-error to sift candidate compounds on a large scale to sidestep the complexities of data in a search for compounds with sufficient potential to be worth the effort of laboratory testing for useful outcomes; similar techniques can be used for strain improvement and natural product synthesis.[26]

A second approach is to use computational modeling and simulation. Data mining (Section 4.4.8) can be used in addition to empirical screening to identify compounds that are likely to have a desired pharmacological effect. Moreover, what the combinatorial and high-throughput empirical approach gains in expediency, it may lose in insight. For example, causality in combinatorial approaches is often difficult to attribute; and thus, it is difficult to generalize these results to other systems. Combinatorial methods are less likely to find solutions when the desired functionality is complex (e.g., when the biosynthetic route to a product is complicated or when a disease treatment relies on the inhibition, without side effects, of various pathways). Also, of course, from the standpoint of basic science, predictive understanding is at a premium. Computational simulation is thus used as the screening tool for promising compounds—a cell's predicted functional response to a given compound is used as that compound's measure of promise for further (empirical) testing. Thus, although granting drug approvals on the basis of simulations makes little sense, simulations may be able to predict with an adequate degree of reliability what drugs should not advance to expensive in vivo clinical trials.[27] Many believe that information-enabled bioscience and biotechnology have the potential to be as revolutionary as information technology was a few decades ago.

[25] E. Frauenheim, "Computers Replace Petri Dishes in Biological Labs," *CNET News.com,* June 2, 2003, available at http://news.com.com/2030-6679_3-998622.html?tag=fd_lede2_hed.

[26] See, for example, C. Khosla, and R.J. Zawada, "Generation of Polyketide Libraries via Combinatorial Biosynthesis," *Trends in Biotechnology* 14(9):335-341, 1996; C.R. Hutchinson, "Combinatorial Biosynthesis for New Drug Discovery," *Current Opinion in Microbiology* 1(3):319-329, 1998; A.T. Bull, A.C. Ward, and M. Goodfellow, "Search and Discovery Strategies for Biotechnology: The Paradigm Shift," *Microbiology in Molecular Biology Review* 64(3):573-606, 2000; Y. Xue and D.H. Sherman, "Biosynthesis and Combinatorial Biosynthesis of Pikromycin-related Macrolides in *Streptomyces venezuelae*," *Metabolic Engineering* 3(1):15-26, 2001; and L. Rohlin, M. Oh, and J.C. Liao, "Microbial Pathway Engineering for Industrial Processes: Evolution, Combinatorial Biosynthesis and Rational Design," *Current Opinion in Microbiology* 4(3):330-335, 2001.

[27] For example, the Tufts Center for the Study of Drug Development estimates the cost of a new prescription drug at $897 million, a figure that includes expenses of project failures (e.g., as those drugs tested that fail to prove successful in clinical trials). Since clinical trials—occurring later in the drug pipeline—are the most expensive parts of drug development, the ability to screen out drug candidates that are likely to fail in clinical trials would have enormous financial impact and would also reduce the many years associated with clinical trials. See Tufts Center for the Study of Drug Development news release, "Total Cost to Develop a New Prescription Drug, Including Cost of Post-Approval Research, Is $897 Million," May 13, 2003, available at http://csdd.tufts.edu/NewsEvents/RecentNews.asp?newsid=29. Of particular interest is a finding reported by DiMasi that if preclinical screening could increase success rates from the current 21.5 percent to 33 percent, the cost per approved drug could be reduced by $230 million (J.A. DiMasi, "The Value of Improving the Productivity of the Drug Development Process: Faster Times and Better Decisions," *PharmacoEconomics* 20(S3):1-10, 2002).

As in the case of academic organizations, specific company names provided below are illustrative and hardly exhaustive, and no inference should be drawn from the fact that any given company is not included.

10.2.4.1 Major IT Corporations

As a fast-growing, (comparatively) well-funded, and high-profile sector, life sciences research and business represents an irresistible target to large IT vendors. As such, companies such as HP and IBM have both developed suites of products and services customized for the consumption of research labs as well as the biotechnology and pharmaceutical sectors. These services are not necessarily substantially different from those that vendors provide to other sectors—a disk drive is a disk drive—but are bundled with useful software or interfaces designed with the life sciences in mind.

IBM established its Life Sciences Business Unit in 1998, incorporating hardware, consulting services, and an aggressive alliance program that includes many major vendors of bioinformatics and related software. In addition, it provides DiscoveryLink, a customized front end to IBM's successful DB/2 relational database product. Among other features, DiscoveryLink allows single-application views and queries into multiple back-end databases, providing a convenient answer to a very common situation in bioinformatics, which often deals with many databases simultaneously.

Of higher profile are IBM's research activities in computational biology. One of these is Blue Gene, the architectural successor to Deep Blue, the IBM-designed supercomputer that beat chess champion Gary Kasparov in 1997. Blue Gene, announced in 1999 as a $100 million, 5-year project, is projected to be 1 petaflop (10^{15} floating point operations per second), a thousand times more powerful than Deep Blue, and 30 times more powerful than the NEC Earth-Simulator/5120. Blue Gene is designed in part to be able to simulate the molecular forces that occur during protein folding, in order to better understand how a large protein shape emerges from a peptide sequence.[28]

Blue Gene is only one project, albeit the best known, of IBM Research's Computational Biology Center. This is a group of approximately 35 researchers who are investigating computational techniques in molecular dynamics, pattern discovery, genome annotation, heterogeneous database techniques, and so forth.

Hewlett-Packard also maintains a life sciences division, and aggressively sells hardware, software, and services to genomics research organizations, pharmaceutical companies, and agribusiness.[29] HP has had good success in winning high-profile clients.

10.2.4.2 Major Life Science Corporations

Genomic bioinformatics, and more generally the use of information technology to support research and development, has become one of the central pillars of the modern biotechnology industry, especially the pharmaceutical sector. A wave—some say a boom—of investment in bioinformatics in the late 1990s and early 2000s has tapered off, however, due to disappointing returns amid mounting costs. While few in the industry doubt the eventual impact of computational techniques, the more significant effects may not be felt for years. Even in 2002, however, corporate spending in bioinformatics was estimated to be $1 billion.[30]

The first wave of biotechnology firms, established in the 1970s, has grown into multibillion dollar operations. These firms—Amgen, Biogen, Chiron, Genentech, and Genzyme—were all founded with

[28] F. Allen, G. Almasi, W. Andreoni, D. Beece, B.J. Berne, A. Bright, J. Bruheroto, et al., "Blue Gene: A Vision for Protein Science Using a Petaflop Supercomputer," *IBM Systems Journal* 40(2):310-327, 2001, available at http://www.research.ibm.com/journal/sj/402/allen.html.

[29] See http://www.hp.com/techservers/life_sciences/overview.html.

[30] See http://www.redherring.com/investor/2002/0419/dealflop.html.

the idea of capitalizing on progress in genetic technologies. Yet because they predate the bioinformatics boom, they were often late to the game, catching up by heavy investment or by outright purchasing of other firms that had organically grown the bioinformatics capability. For example, in December of 2001, Amgen announced that it was buying the bioinformatics-rich biotech company Immunex Corp for $16 billion.[31] Genentech highlights its own bioinformatics capabilities as a key part of the research portfolio.[32] However, while these firms and the pharmaceutical giants are clearly great consumers of bioinformatics software and human resources, it is less clear to what extent they are performing original computational biology research.

A second wave of companies was founded in the 1990s, in the era of the Human Genome Project and the increase in availability of information technology. Millennium Pharmaceuticals, for example, was founded in 1993 with the goal of being a science- and technology-driven pharmaceutical company, with a capability for target discovery based on the human genome information being published. However, most of Millennium's drugs on the market have come from acquisitions, and the goal of real rational drug discovery remains challenging. Millennium does have a high-profile leader in charge of bioinformatics and uses IT for three main functions: bioinformatic inference making, such as identifying likely functions of novel proteins or the existence of gene expression patterns that correlate with disease states; chemoinformatics, searchable databases of chemical structure and biological activity; and computational analysis to predict drug candidates' physiological qualities such as absorption rates, distribution, metabolism, excretion, and toxicity.

Of higher profile is Celera, which Craig Venter founded in 1998 to compete with the publicly funded Human Genome Project. While genomics experts still argue over his methods, he certainly found innovative uses for computational and analytic techniques in stitching together the results of his "shotgun" sequencing method. Regardless of its scientific success, however, Celera has had little commercial success[33] as it turned from sequencing to the potentially more lucrative field of drug discovery. It still makes money by offering access to its proprietary databases to other biotechnology and pharmaceutical companies, but its has given up on its efforts to commercialize its software platform, selling the Celera Discovery System to sister company Applied Biosystems (both Celera and Applied Biosystems are owned by Applera Corporation). In addition to the Celera Discovery System, a subscription-based database, Applied Biosystems offers an array of software for gene sequencing, laboratory information management, and gene analysis (as well as a variety of instrumentation and reagents).

10.2.4.3 Start-up and Smaller Companies

The area still receives some attention from venture capital firms such as Flagship Ventures, Kleiner Perkins Caufield Byers, Atlas Ventures, and Alloy Ventures. However, the emphasis seems to be shifting from bioinformatics to a stronger emphasis on biology, including medical devices and drug discovery. Even companies that once positioned themselves as bioinformatics companies now describe themselves as being in the drug discovery business,[34] most notably Celera but also many smaller companies. For companies that concentrate primarily or exclusively on informatics, times are very difficult, in large part due to the same sort of bubble collapse as mainstream IT faced from 2000 onward.

Analysts blame overinvestment in the area, leading to more companies than the space can support; companies founded by IT players with insufficient biological knowledge; and increasing competition from big players such as IBM and HP.

Midsize companies such as Gene Bank and Incyte have a similar business model to Celera, offering access to proprietary databases, which often contain patented gene sequences. One model that seems to

[31] See http://www.informationweek.com/story/IWK20011221S0038.
[32] See http://www.genentech.com/gene/research/biotechnology/bioinformatics.jsp.
[33] See http://www.fool.com/portfolios/rulebreaker/2002/rulebreaker020423.htm.
[34] See http://www.bizjournals.com/washington/stories/2002/07/08/newscolumn5.html.

be more successful than others is "in silico" simulation of various biological and biomedical processes, such as offered by AnVil Informatics.[35] Beyond Genomics develops proprietary algorithms that look for large-scale biological systems such as pathways in gene and protein bioinformatics and experimental data. A second seemingly successful model is a focus on providing information about pathways and networks; Ingenuity, Cytoscape, GeneGO, PathArt, are companies that have sought to exploit this niche.

Beyond bioinformatics, vendors are attempting to develop or customize for life sciences customers a number of IT solutions, including applications for knowledge management, laboratory information management, and tracking clinical trials (including sophisticated statistical analysis).

A leading example of real computer science research being applied to biology problems is the application of distributed or grid computing to extremely computation-intensive tasks such as protein folding simulation. While many IT vendors are developing and pushing their grid platform, Stanford has been running Folding@Home, a screen-saver that anyone can download and run on a home computer, which calculates a tiny piece of the protein folding problem.[36]

10.2.5 Funding and Support

Both the federal government and private foundations support research at the BioComp interface. (The latter can be regarded as an offshoot of the historically extensive foundation support for biology research.)

10.2.5.1 General Considerations

10.2.5.1.1 The Role of Funding Institutions Funding institutions obviously exert a great deal of control and influence over the nature and direction of research. That is, researchers tend to gravitate toward research problems for which funding is available. Funding agencies can also influence the development of new talent in the field by encouraging faculty development, as illustrated non-exhaustively below:

- Release time to design new curricula and collect successful course material. However, as in peer-reviewed scientific research, the fruits of these efforts should be made public, and their successes or limitations should be openly available (e.g., as online courses or published material).
- Supervision of undergraduate special projects or research at the BioComp interface. Special projects for one or a few undergraduates (e.g., summer student projects, undergraduate theses) can be undertaken with minimal risk, and facilitating early exposure to a variety of ideas would benefit both students and faculty.
- Support for individuals who wish to make the transition to research at the BioComp interface early in their careers. Such individuals may lack the publication track record that would enable more senior researchers to undertake such a transition. Thus, support dedicated to such people may facilitate early career transitions and all of the accompanying benefits.

10.2.5.1.2 The Review Process A central dimension of funding institutions is the review process they employ to decide what research to support. Different institutions have different styles, but they all face the same types of issues.

- *Excellence*. No institution wants to support mediocre research. But as suggested below in Section 10.3.1, definitions of excellence are in many ways field-specific. Thus, an effective review process must find ways of managing this tension when proposals cross disciplinary lines.

[35] See http://www.anvilinformatics.com.

[36] See http://folding.stanford.edu. Perhaps the most famous of such distributed applications is SETI@Home, a program that supports the data processing underlying the search for extraterrestrial life.

- *Potential impact.* All else being equal, institutions would prefer to support research in which the potential impact of success is large. However, as a rule, claims of large impact are much more speculative than other claims, simply because the long-term ramifications of any given discovery are difficult to underscore in any convincing manner before the fact.
- *Technical risk.* A research investigation may or may not be successful. Research that presents the lowest technical risk (i.e., the lowest risk of failure or of being unsuccessful) is most often very closely tied to some existing and successful research. Thus, as a rule, research that is of low technical risk tends also to be of lesser potential impact.
- *Personnel risk.* Research is performed by people, and any given research effort can be executed more or less effectively depending on the people involved. Established track records of success are an important dimension of the teams proposed to undertake research but cannot be the only dimension taken into account if new researchers with good ideas are to be welcomed.
- *Budget.* Institutions with a fixed level of support to offer investigators can support a larger number of inexpensive research proposals or a smaller number of more expensive ones. All else being equal, inexpensive proposals will tend to be favored over expensive ones.

Proposals for research must weigh each of these factors and make trade-offs among them. For example, a lower budget may mean greater technical or personnel risk; a high-impact project may have greater technical risk. Funding agencies must assess the plausibility of the trade-offs that a prospective research team has made.

These notions suggest that review panels need a wide range of expertise and experience to judge the merits of new proposals effectively or to carry out peer review of scientific papers. In principle, the requisite range of expertise can be obtained through the use of a set of individual disciplinary experts whose collective expertise is adequately broad. An alternative is to use a few individuals who themselves have interdisciplinary expertise. The disadvantage of the first model is that for practical purposes it may reproduce forums in which the difficulties of cross-disciplinary understanding are manifested. The disadvantage of the second model is that such individuals may be few in number and thus difficult to enlist.

10.2.5.2 Federal Support

A variety of federal agencies support work at the BioComp interface, and this support has grown over time.

10.2.5.2.1 The National Institutes of Health For computational biology (i.e., the computing-to-biology side of the BioComp interface), the main actor in the U.S. government is the National Institutes of Health, part of the Department of Health and Human Services.

A notable instance of bioinformatics work at NIH is the National Center for Biotechnology Information (NCBI), a part of the National Library of Medicine. Established in 1988, it is NCBI that created and maintains GenBank (see Chapter 3).

The NIH's National Institute of General Medical Sciences (NIGMS) manages the Biomedical Information Science and Technology Initiative, or BISTI. BISTI represents an NIH-wide collaboration and coordination program between its many institutes and centers, as computational biology and bioinformatics activity is spread throughout the organization. In addition, NIGMS also runs the Center for Bioinformatics and Computational Biology, which focuses on theoretical and methodological infrastructure, such as modeling, simulation, theory, and analysis tools in biological networks.[37] The NIH's Center for Information Technology, in addition to providing IT services to the rest of NIH, also main-

[37] See http://www.nigms.nih.gov/news/releases/cbcb.html.

tains the Division of Computational Bioscience, which includes activities in high-performance computing and molecular modeling; it is staffed mostly by computer scientists rather than biologists and appears to focus on the computer science aspects of problems.

In addition, the National Center for Research Resources (NCRR) is a center within NIH whose mission is to create new research technologies and provide researchers access to resources such as high-end instrumentation, animal models, and cell line repositories. In FY 2004, it had a budget of slightly over a billion dollars, in large part dedicated to funding research centers, as well as individual predoctoral, postdoctoral, and career awards. NCRR's 2004-2008 strategic plan includes a number of computational biology activities within its funding programs. This includes support for software and algorithm development, mathematical modeling, and simulation. NCRR, through its Research Infrastructure Division, also supports the creation of networks to promote cross-institutional collaboration, including virtual laboratories and shared databases for a variety of specific clinical research programs. This includes the Biomedical Informatics Research Network (BIRN), an Internet2 project first funded in 2002 and slated to expand in 2004. NCRR also supports cross-discipline training at all levels of a researcher's career—for example, supporting the entry into biology of individuals with backgrounds in technical fields such as computer science, and retraining established researchers in appropriate fields.

The National Institute for Biomedical Imaging and Bioengineering (NIBIB) is the newest institute at NIH, and is unusual for its mission of assessing and developing technological capabilities for health and medical research. Its research goals and portfolio include support for a number of activities at the BioComp interface, including bioinformatics, simulation and computational modeling, image processing, brain-computer interfaces, and telemedicine. More broadly, its support for interdisciplinary training and research that draw on engineering, as well as physical and life sciences, mark it as another instrument for encouraging the development of researchers and scientists having experience with and exposure to computational science.

In addition to these institutional entities, NIH has created a set of programmatic initiatives to promote quantitative, interdisciplinary approaches to biomedical problems that involve the complex, interactive behavior of many components.[38] One initiative consists of a variety of programs to develop human capital, including those for predoctoral training for life scientists in bioinformatics and computational biology,[39] support for short courses on mathematical and statistical tools for the study of complex phenotypes and complex systems,[40] postdoctoral fellowships in quantitative biology,[41] and support for a period of supervised study and research for professionals with quantitative scientific and engineering backgrounds outside of biology or medicine who have the potential to integrate their expertise with biomedicine and develop into productive investigators.[42] The National Library of Medicine supported awards for predoctoral and postdoctoral training programs in informatics research oriented toward the life sciences (originally medical informatics but moving toward biomedical informatics in its later years).[43]

A second group of programs is targeted toward specific problems involving complex biomedical systems. This group includes an R01 program focused on genetic architecture, biological variation, and complex phenotypes (including human diseases);[44] another on quantitative approaches to the analysis of complex biological systems, with a special focus on research areas in which systems approaches are likely to result in the determination of the system-organizing principles and/or the system dynamics;[45] and still another on evolutionary mechanisms in infectious diseases.[46]

[38] See http://www.nigms.nih.gov/funding/complex_systems.html.

[39] See http://grants.nih.gov/grants/guide/pa-files/PAR-99-146.html.

[40] See http://grants.nih.gov/grants/guide/pa-files/PA-98-083.html.

[41] See http://grants.nih.gov/grants/guide/pa-files/PA-98-082.html.

[42] See http://grants.nih.gov/grants/guide/pa-files/PA-02-127.html.

[43] See http://grants.nih.gov/grants/guide/rfa-files/RFA-LM-01-001.html.

[44] See http://grants.nih.gov/grants/guide/pa-files/PA-02-110.html.

[45] See http://grants.nih.gov/grants/guide/pa-files/PA-98-077.html. This program includes P01 program project awards as well.

[46] See http://grants.nih.gov/grants/guide/pa-files/PA-02-113.html. This program includes P01 program project awards as well.

A third group of programs is institutional in nature. One program establishes new academic Centers of Excellence in Complex Biomedical Systems Research[47] that promote the analysis of the organization and dynamic behaviors of complex biological systems through the development of multi-investigator teams capable of engaging biomedical complexity with a scope of activities not possible with other funding mechanisms, including research, training, workshops, symposia, and other forms of outreach. Typical areas of interest include computationally based modeling of processes such as the cell cycle; pattern formation during embryogenesis; the flux of substrates and intermediates in metabolism; and the application of network analysis to understanding the integrated systemic host responses to trauma, burn, or other injury. A second program on Integrative and Collaborative Approaches to Research[48] encourages collaborative and integrative approaches to research on multifaceted biological problems for individual investigators with existing support who need to attract and coordinate expertise in different disciplines and approaches and require access to specialized resources, such as computational facilities, high-throughput technologies, and equipment. A third program[49] supports new quantitative approaches to the study of complex, fundamental biological processes by encouraging nontraditional collaborations across disciplinary lines through supplements to existing R01, R37, or P01 NIGMS grants to support the salary and expenses of collaborating investigators such as physicists, engineers, mathematicians, and other experts with quantitative skills relevant to the analysis of complex systems.

Finally, a major contributor to research that includes biology and computation is the NIH Roadmap. The Roadmap is a broad set of funding opportunities and programs dealing with research issues that, due to their complexity, scope, or interdisciplinary nature, could not be addressed adequately by a single NIH institute or center. Relevant BioComp programs described by the Roadmap include molecular libraries, which in part seek to develop large databases of "small molecules," and structural biology, which includes research to develop algorithmic tools for analyzing and predicting protein structure.

The most significant BioComp initiative within the Roadmap, however, is the Bioinformatics and Computational Biology program. This program seeks to create and support a National Program of Excellence in Biomedical Computing (NPEBC), a national network of software engineering and grid resources to support cutting-edge biomedical research. The prime components of the NPEBC are the National Centers for Biomedical Computing (NCBCs), seven 5-year U54 grants that total approximately $120 million, along with a larger number of R01 and R21 individual grants to support collaboration opportunities with the NCBCs.

The NCBCs are intended as more than merely well-funded research centers; their missions of training, tool creation and dissemination, community support, and liberal intellectual property policies for software and data are designed to create national networks and communities of researchers organized around BioComputational research. The structure of the grant process required the identification of three different research thrusts (or "cores"): a core of computational research, responsible for performing original work in algorithms and computer science; a core of biomedical research, or "driving biological projects," and a core Biocomputing engineering, responsible for both interfacing between computation and biomedical research, and creating the concrete tools and software systems to actualize the research.

The recipients of the first round of NCBC funding were announced in September of 2004, covering four centers. The second round, expected to fund an additional three centers, will be announced in 2005. The centers funded in the first round include:

- The Stanford Center for Physics-based Simulation of Biological Structures, an effort that seeks to create common software and algorithmic representation for modeling and simulation, addressing prob-

[47] See http://grants.nih.gov/grants/guide/rfa-files/RFA-GM-03-009.html.
[48] See http://grants.nih.gov/grants/guide/pa-files/PA-00-099.html.
[49] See http://grants.nih.gov/grants/guide/pa-files/PA-98-024.html.

lems of how to integrate models that may have widely different physical scales, have discrete or continuous approximations, or work at very different levels of abstraction. The driving biological problems for this center include RNA folding, myosin dynamics, neuromuscular dynamics, and cardiovascular mechanics.

- The National Alliance for Medical Image Computing (NAMIC) is a center based at Brigham and Women's Hospital in Boston that includes partners from universities and research centers around the country. The goal of NAMIC is to develop computational tools for analysis and visualization of image data, especially in integrating data from many different imaging technologies (e.g., magnetic resonance imaging, electroencephalography, positron emission tomography, etc.) with genomic and clinical data. The initial driving biological projects for NAMIC are various forms of neurological abnormality associated with schizophrenia.
- The Center for Computational Biology at UCLA is also investigating questions of imaging, concentrating on the production of "computational atlases," database-like structures that allow sophisticated queries of large-scale data. The computational research includes mathematics of volumes and geometry, and the driving biological projects are language development, Alzheimer's, multiple sclerosis, and schizophrenia.
- The Center for Informatics for Integrating Biology and the Bedside (I2B2), organized by a consortium of Boston-area universities, hospitals, and medical insurance providers, seeks to develop techniques to integrate and present huge sets of clinical data in ways appropriate for research into the genetic bases of disease and, thus, helping to identify appropriate targeted therapies for individual patients. This involves the development of statistical and algorithmic techniques for analyzing protein structure, as well as population dynamics. The driving biological projects include airways diseases such as asthma, hypertension, Huntington's disease, and diabetes.

10.2.5.2.2 The National Science Foundation The National Science Foundation provides a great deal of support for research at the BioComp interface through its programs of individual and institutional grants. The NSF's Directorate of Biological Sciences (BIO) formerly offered a funding program in computational biology activities. The BIO directorate ended this program in 1999,[50] not because the research no longer deserved funding, but because computational biology had "mainstreamed" to become an important part of many other biological research activities, particularly environmental biology, integrative biology, and molecular and cellular biosciences. NSF does, in its Biological Infrastructure Division, maintain a biological databases and informatics program that funds direct research into the creation of tools and datasets.

In its 2003 report *Science and Engineering Infrastructure for the 21st Century: The Role of the National Science Foundation*, NSF concludes that its support for science and engineering infrastructure (cyberinfrastructure), in which it includes next-generation computational tools and data analysis and interpretation toolkits (along with a great deal of other infrastructure elements), should increase from 22 percent of its total budget to 27 percent; it also recommends strengthening its support for cross-disciplinary fields of research. Both of these recommendations are likely to improve the funding climate for computational biology and bioinformatics, although of course they will still be competing with a number of other important infrastructure programs.

Many existing NSF funding programs emphasize interdisciplinary research and thus are effective vehicles for supporting BioComp research, although not exclusively. For example, the Integrative Graduate Education and Research Traineeship (IGERT) Program offers 5-year, $3 million grants to universities to support interdisciplinary graduate student training.[51] Many of the existing programs funded by IGERT work at the BioComp interface, such as bioinformatics, computational neuroscience, computa-

[50] See http://www.nsf.gov/pubs/1999/nsf99162/nsf99162.htm.

[51] See http://www.nsf.gov/pubs/2005/nsf05517/nsf05517.htm.

tional phylogenetics, functional genomics, and so forth. Within biology, the Frontiers in Integrative Biological Research (FIBR) program is designed to fund research projects using innovative approaches that draw on many fields, including information sciences, to attack major unanswered questions in biology.[52] It funds projects for five years at $1 million per year, and the 2005 round will fund eight projects. Also, a funding program for postdoctoral training in bioinformatics is funded at $1 million.[53]

A central and challenging application in BioComp research is an attempt to construct the entire historic phylogenetic Tree of Life. NSF is supporting this research through its Assembling the Tree of Life program, funded at $29 million; databases will contain molecular, morphological, and physiological evidence for placing taxa in relationship to other taxa. Current algorithms and data structures do not scale well at the number of taxa and data points necessary, so both computational and biological research is necessary to achieve this grand challenge.

The NSF participates with other government agencies in coordinating research agendas and programs. Of particular note is the joint initiative between the NSF Directorate for Mathematics and Physical Sciences and NIGMS to support research in mathematical biology.[54] Work supported under this initiative is expected to impact biology and advance mathematics or statistics, and the competition is designed to encourage new collaborations between the appropriate mathematical and biological scientists as well as to support existing ones. The Office of Science and Technology Policy (OSTP) included research into “molecular-level understanding of life processes” in a list of the government’s top priorities for science and engineering research.[55] NSF is supporting this goal through its CAREER funding program, which is aimed at faculty members early in their careers.[56]

Finally, NSF sponsors a Small Grants Exploratory Research Program that supports high-risk research on a small scale. According to NSF, proposals eligible for support under this program must be for “small-scale, exploratory, high-risk research in the fields of science, engineering and education normally supported by NSF may be submitted to individual programs. Such research is characterized as preliminary work on untested and novel ideas; ventures into emerging research ideas; application of new expertise or new approaches to ‘established’ research topics; efforts having a severe urgency with regard to availability of, or access to data, facilities, or specialized equipment, including quick-response research on natural disasters and similar unanticipated events; or efforts of a similar character likely to catalyze rapid and innovative advances.”[57] Typically, grants provided under this program are less than $200,000.

10.2.5.2.3 Department of Energy The Department of Energy played a key role in the initiation of the Human Genome Project. Its scientific interest was first motivated by a need to understand the biological effects of ionizing radiation, which it viewed as part of the science mission surrounding its stewardship of the nation’s nuclear weapons program. Furthermore, DOE scientists have had considerable experience with advanced computation in the design and manufacturing process for nuclear weapons, a fact that DOE leveraged to investigate the genome.

Today, the Department of Energy is a major supporter of 21st century biology, because it believes that biological approaches may help it to meet its missions of energy production, global climate change mitigation, and environmental cleanup.

- For energy production, renewable energy from plants requires the design of plants with biomass that can be transformed efficiently to fuels. However, a limiting factor in developing such plants is the

[52] See http://www.nsf.gov/pubs/2004/nsf04596/nsf04596.htm.
[53] See http://www.nsf.gov/pubs/2004/nsf04539/nsf04539.html.
[54] See http://www.nsf.gov/pubs/2002/nsf02125/nsf02125.htm.
[55] See FY 2004 *Interagency Research and Development Priorities*, http://www.ostp.gov/html/ombguidmemo.pdf.
[56] See http://www.nsf.gov/pubs/2002/nsf02111/nsf02111.htm.
[57] See http://www.nsf.gov/pubs/2004/nsf042/dcletter.htm.

lack of understanding about their metabolic pathways, and knowledge of these pathways may lead to more efficient strategies for converting biomass to fuels.

- For mitigating climate change, reduction in the buildup of greenhouse gases (specifically CO_2) would be desirable. One approach to this problem is to alter natural biological cycles to store extra carbon in the terrestrial biomass, soils, and biomass that sinks to ocean depths—a sequestration approach. Research continues on the best ways to achieve large-scale carbon sequestration, and one method under investigation is tied to microbial metabolism and activities that may lead to new ways to store and monitor carbon.
- For environmental cleanup, microbes may provide a means to degrade or immobilize contaminants and accelerate the development of new, less costly strategies for cleaning up a variety of DOE waste sites. For example, microbes may be developed that can consume waste materials and degrade them or concentrate them in a form that is easier to clean up.

To address these missions, DOE supports a number of programs. Perhaps the best known is the Genomes-to-Life (GTL) program, a large research grant-providing program with four major scientific goals: (1) identification of systems of interacting proteins at the microbial level ("protein machines"), (2) characterization of gene regulatory networks, (3) exploration of microbial communities and ecosystems, and (4) development of the computational capability for modeling biological systems. To pursue these goals, the GTL program combines large experimental datasets with advanced data management, analysis, and computational simulations to create predictive simulation models of microbial function and of the protein machines and pathways that embody those behaviors. The program identifies specific challenges for computer science:[58] automated gene annotation; software to support protein expression-proteomics analysis; the ability to meaningfully and automatically extract meaning from biological technical papers; simulation for cellular networks; and model and system interoperability. These will require advances in data representation, analysis tools, integration methods, visualization techniques, models, standards, and databases. The program has funded five major projects (three at DOE labs and two at academic institutions) for a total of $103 million over the period from 2002 to 2007. In the project descriptions of the winners, four included "computational models" as part of their charge.[59]

A second DOE effort is the Microbial Genome program, which spun off from the Human Genome Project in 1994. The Microbial Genome program exploits modern sequencing technologies to sequence completely the genomes of microbes, primarily prokaryotes, based on their relevance for energy, the global carbon cycle, and bioremediation. As of April 2003, the genomes of about 100 microbes had been sequenced, most of them by the Joint Genome Institute,[60] and placed in public databases. Microbial genomics presents some particularly interesting science in that for newly sequenced microbial genomes, a large fraction of the genes identified (about 40 percent) have unknown functions and biological value. In addition, most of what is known about microbes involves microbes that are easy to culture and study or that cause serious human and animal diseases. These constitute only a small minority of all microbes living in natural environments. Most microbes are part of communities that are very difficult to study but play critical roles in Earth's ecology, and a genomic approach to understanding these microbes may be one of the only paths toward developing an understanding of them.

A third component of DOE's efforts is in structural biology. The purpose of this program is to understand the function of proteins and protein complexes that are key to the recognition and repair of DNA damage and the bioremediation of environmental contamination by metals and radionuclides. Research supported in this program focuses on determining the high-resolution three-dimensional

[58] See http://www.doegenomestolife.org/pubs/ComputerScience10exec_summ.pdf.

[59] See http://doegenomestolife.org/research/2002awards.htm.

[60] The Joint Genome Institute, established in 1997, is a consortium of scientists, engineers, and support staff from DOE's Lawrence Berkeley, Lawrence Livermore, and Los Alamos National Laboratories. See http://www.jgi.doe.gov/whoweare/index.html.

structures of key proteins; understanding the changes in protein structure related to interaction with molecules such as DNA, metals, and organic ligands; visualization of multiprotein complexes that are essential to understand DNA repair and bioremediation; prediction of protein structure and function from sequence information, and modeling of the molecular complexes formed by protein-protein or protein-nucleic acid interactions.

10.2.5.2.4 Defense Advanced Research Projects Agency With a reputation for engaging in "high-risk, high-return" research, DARPA has been a key player in the development of applications that utilize biomolecules as information processing, sensing, or structural components in anticipation of reaching the limits of Moore's law. This research area, largely supported under DARPA's biocomputation program,[61] was described in Section 8.4. Managed out of DARPA's Information Processing Technology Office (IPTO), the biocomputation program has also supported the BioSPICE program, a computational framework with analytical and modeling tools that can be used to predict and control cellular processes (described in Chapter 5 (Box 5.7)). Finally, the biocomputation program has supported work in synthetic biology (i.e., the design and fabrication of biological components and systems that do not already exist in the natural world) as well as the redesign and fabrication of existing biological systems (described in Section 8.4.2.2).

IPTO also supports a number of programs that seek to develop information technology that embodies certain biological characteristics.[62] These programs have included the following:

- *Software for distributed robotics*, to develop and demonstrate techniques to safely control, coordinate, and manage large systems of autonomous software agents. A key problem is to determine effective strategies for achieving the benefits of agent-based systems, while ensuring that self-organizing agent systems will maintain acceptable performance and security protections.
- *Mobile autonomous robot software*, to develop the software technologies needed for controlling the autonomous operation of singly autonomous, mobile robots in partially known, changing, and unpredictable environments. In this program, ideas from robot learning and control are extended, including soft computing, robot shaping, and imitation.
- *Taskable agent software kit*, to codify agent design methodology as a suite of control and decision mechanisms, to devise metrics that characterize the conditions and domain features that indicate appropriate design solutions, and to explain and formalize the notion of emergent behavior.
- *Self-regenerative systems*, to develop core technologies necessary for making computational systems able to continue operation in the face of attacks, damage, or errors. Specific avenues of investigation include biological metaphors of diversity, such as mechanisms to automatically generate a large number of different implementations of a given function that most of them will not share a given flaw; immune systems; and human cognitive models.
- *Biologically inspired cognitive architectures*, to codify a set of theories, design principles, and architectures of human cognition that are specifically grounded in psychology and neurobiology. Although implementation of such models on computers is beyond the scope of the current project, it is a natural extension once sufficiently complete models can be created.

DARPA's Defense Sciences Office (DSO) supports a variety of programs that connect biology to computing in the broad sense in which this report uses the term. These programs have included the following:

[61] See http://www.darpa.mil/ipto/programs/biocomp/index.htm.
[62] See http://www.darpa.mil/ipto/Programs/programs.htm.

- *Bio:Info:Micro.* In collaboration with DSO, IPTO and the Microsystems Technology Office, the Bio:Info:Micro program supports research in neuroprocessing and biological regulatory networks. These research thrusts seek to develop devices for interrogating and manipulating living brains and brain slices (in the neuroprocessing track) and single cells or components thereof (in the regulatory network track), and the computational tools needed to analyze and interpret information derived from these devices. Thus, neural decoding algorithms for neural spikes and local field potentials, and methods for representing spatial components in distributed systems and using decision theoretic approaches for decoding brain signals are of interest to the neuroprocessor track, and algorithms that can automatically detect patterns and networks given appropriate data and models for networks that govern cell growth and death are of interest to the regulatory track.
- *Biological input/output systems.* Focused on the design and assembly of molecular components and pathways that can be used to sense and report the presence of chemical or biological analytes, this program seeks to develop technologies to enable the facile engineering and assembly of functional biological circuits and pathways in living organisms, thereby enabling such organisms to serve as remote sentinels for those analytes. The essential notion is that the binding of an analyte to an engineered cytoplasmic or cell surface receptor will lead to regulated and specific changes in an organism, which might then be observed by imaging, spectroscopy, or DNA analysis.
- *Simulation of biomolecular microsystems.* Biological or chemical microsystems in which biomolecular sensors are integrated with electronic processing elements offer the potential for significant improvements in the speed, sensitivity, specificity, efficiency, and affordability of such systems. This program seeks to develop data, models, and algorithms for the analysis of molecular recognition processes; transduction of molecular recognition signals into measurable optical, electrical, and mechanical signals; and on-chip fluidic-molecular transport phenomena. The ultimate goal is to produce advanced computer-aided design (CAD) tools for routine analysis and design of integrated biomolecular microsystems.
- *Engineered biomolecular nanodevices and systems.* This program is focused on hybrid (biotic-abiotic) nanoscale interface technologies that enable direct, real-time conversion of biomolecular signals into electrical signals. Success in this area would enable engineered systems to exploit the high sensory sensitivity, selectivity, and efficiency that characterize many biological processes. The objective of this research is to develop hybrid biomolecular devices and systems that use biological units (e.g., protein ion channels or nanopores, g-protein-coupled receptors) for performing a sensing function but use silicon circuitry to accomplish the signal processing. Ultimately, this research is intended to lay the foundation for advanced "biology-to-digital" converter systems that enable direct, real-time conversion of biological signals into digital information.
- *Biologically inspired multifunctional dynamic robots.* This program seeks to exploit biological approaches to propulsion mechanisms for multifunctional, dynamic, energy-efficient, and autonomous robotic locomotion (e.g., running over multiple terrains, climbing trees, jumping and leaping, grasping and digging); recognition and navigation mechanisms that enable biological organisms to perform terrain following, grazing incidence landings, target location and tracking, plume tracing, and hive and swarm behavior; and the integration of these capabilities into demonstration robotic platforms.
- *Compact hybrid actuators program.* This program seeks to develop electromechanical and chemomechanical actuators that perform the same functions for engineered systems that muscle performs for animals. The performance goal is that these new actuators must exceed the specific power and power density of traditional electromagnetic- and hydraulic-based actuation systems by a factor of 10.
- *Active biological warfare sensors.* This program seeks to develop technology to place living cells with similar behavior to human cells onto chips, so that their health and behavior can be monitored for the presence of harmful chemical or biological agents.
- *Protein design processes.* This program is using two specific challenge problems to motivate research into technologies for designing novel proteins for specific biological purposes. Such design will require advances in computational models, as well as knowledge of molecular biology. The challenge

problems include tasks of designing specific proteins in less than a day that can catalyze specific chemicals or inactivate an influenza virus.

As a vehicle for pursuing its mission, DARPA typically uses Broad Agency Announcements. Some are focused on achieving a specific technical capability (e.g., a program that will develop technology for synthesizing, within 24 hours, an arbitrary 10,000-oligonucleotide sequence in quantity), whereas others are more broadly cast. In many cases, these programs target private industry as well as the more engineering-oriented academic institutions.

10.3 BARRIERS

Because work at the BioComp interface draws on different disciplines, there are barriers to effective cooperation between practitioners from each field. (In some cases, "each field" is more properly cast as the contrast between practitioners of systems biology and practitioners of empirical or experimental biology.) This section describes some of these barriers.[63]

10.3.1 Differences in Intellectual Style

It is almost axiomatic that substantial progress in any area of intellectual inquiry depends on the excellence of work undertaken in that area. On the other hand, differences in intellectual style will affect what is regarded as excellence, and computer scientists and biologists often have very different intellectual styles.

The existence of shared intellectual styles tends to increase the mutual understanding of colleagues working within their home disciplines, a fact that leads to more efficient communication and to shared epistemological understanding and commitments. However, when working across disciplines, lack of a shared intellectual style increases the difficulties for both parties in making meaningful progress.

What are some of the differences involved? While it is risky (indeed, foolhardy) to assert hard and fast differences between the disciplines, an examination of the intellectual traditions and histories associated with each discipline suggests that practitioners in each are generally socialized and educated with different styles.[64] Over time, these differences may moderate as biology becomes a more quantitative discipline (indeed, a premise of this report is that such evolution is to be encouraged and facilitated).

10.3.1.1 Historical Origins and Intellectual Traditions

Many differences in intellectual style between the two fields originate in their histories.[65] Computer science results from a marriage between mathematics and electrical engineering—although it has evolved far from these beginnings. The mathematical thread of computer science is based on formal problem statements, formulating hypotheses (conjectures) based on those statements, and generating formally correct proofs of those hypotheses. Most importantly, a single counterexample to a conjecture invalidates the conjecture. Note also that formal proofs often entail problems that are far from reality, because many real problems are simply too complex to be represented as formal problem statements that are at all comprehensible. Research in mathematics (specifically, applied mathematics) often con-

[63] An early perspective on some of these barriers can be found in K.A. Frenkel, "The Human Genome Project and Informatics: A Monumental Scientific Adventure," *Communications of the ACM* 34:40-51, 1991.

[64] An interesting ethnographic account of life in an academic biology laboratory is provided in J. Owen-Smith, "Managing Laboratory Work Through Skepticism: Processes of Evaluation and Control," *American Sociological Review* 66(3):427-452, 2001.

[65] Some of this discussion is inspired by G. Wiederhold, "Science in Two Domains," Stanford University, March 2002, updated February 2003. Unpublished manuscript.

sists of finding solutions to abstractly formulated problems and then finding real-world problems to which these solutions are applicable.

The engineering thread of computer science is based on finding useful and realizable solutions to real-world problems. The space of possible solutions is usually vast and involves different architectures and approaches to solving a given problem. Problems are generally simplified so that only the most important aspects are addressed. Economic, human, and organizational factors are at least as important as technological ones, and trade-offs among alternatives to decide on the "best" approach to solve a (simplified) problem often involve art as much as science.

Biology—the study of living things—has an intellectual tradition grounded in observation and experiment. Because biological insight has often been found in apparently insignificant information, biologists have come to place great value on data collection and analysis. In contrast to the theoretical computer scientist's idea of formal proof, biologists and other life scientists rely on empirical work to test hypotheses.

Because accommodating a large number of independent variables in an experiment is expensive, a common experimental approach (e.g., in medicine and pharmaceuticals) is to rely on randomized observations to eliminate or reduce the effect of variables that have not explicitly been represented in the model underlying the experiment. Subsequent experimental work then seeks to replicate the results of such experiments.

A biological hypothesis is regarded as "proven" or "validated" when multiple experiments indicate that the result is highly unlikely to be due to random factors. In this context, the term "proven" is somewhat misleading, as there is always some chance that the effect found is a random event. A hypothesis "validated" by experimental or empirical work is one that is regarded as sufficiently reliable as a foundation for most types of subsequent work. Generalization occurs when researchers seek to extend the study to other conditions, or when investigation is undertaken in a new environment or with more realism. Under these circumstances, the researcher is investigating whether the original hypothesis (or some modification thereof) is more broadly applicable.

Within the biological community (indeed, for researchers in any science that relies on experiment), repetition of an experiment is usually the only way to validate or generalize a finding, and replication plays a central role in the conduct of biological science. By contrast, reproducing the proof of a theorem is done by mathematicians and computer scientists mostly when a prior result is suspicious. Although there is an honored tradition of seeking alternative proofs of theorems even if the original proof is not at all suspicious, replication of results is not nearly as central to mathematics as it is to biology.

Finally, biology is constrained by nature, which makes rules (even if they are not known a priori to humans), and models of biological phenomena must be consistent with the constraints that those rules imply. By contrast, computer science is a science of the artificial—more like a game in which one can make up one's own rules—and the only "hard" constraints are those imposed by mathematical logic and consistency (hence data for most computer scientists have a very different ontological role than for biologists).

10.3.1.2 Different Approaches to Education and Training

The first introduction to computer science for many individuals involves building a computer program. The first introduction to biology for many individuals is to watch an organism grow (remember growing seeds in Dixie cups in grade school?). These differences continue in different training emphases for practitioners in computer science and biology in their undergraduate and graduate work.

To characterize these different emphases in broad strokes, formal training in computer science tends to emphasize theory, abstractions, problem solving, and formalism over experimental work (indeed, computer programming—core to the field—is itself an abstraction). Moreover, as with many mathematically oriented disciplines, much of the intellectual content of computer science is integrated and, in that sense, cumulative. By contrast, data and experimental technique play a much more central role in a biologist's education. Traditionally, mathematics (apart from statistics) is not particularly important to biology education; indeed many biologists have entered the field because they wish to pursue science that does not involve a great deal of math. Although there is a common core of knowledge among most biologists, there is an enormous amount of highly specialized knowledge that is not tightly integrated.

A second issue, often encountered in conversion programs, is the difficulty of expanding one's horizons to choose intellectual approaches or tools appropriate to the nature of the problem. Disciplinary training in any field entails exposure to the tools and approaches of that field, which may not be the best techniques for addressing problems in another field. Thus, successful researchers and practitioners at the BioComp interface must be willing to approach problems with a wide array of methodologies and problem-solving techniques. Computer scientists often may be specialists in some specific methodology, but biological research often requires the coordination of multiple approaches. Conversely, biological labs or groups that address a wide range of questions may be more hospitable to computational researchers, because they may provide more opportunities in which computational expertise is relevant.

10.3.1.3 The Role of Theory

Theory plays a very different role and has a very different status in the two fields. For computer scientists, theoretical computer science is essentially mathematics, with all of the associated rigor, certainty, and difficulties. Of particular interest in theoretical computer science is the topic of algorithmic complexity. The most important practical results from algorithmic complexity indicate the scaling relationships between how long it takes to solve a problem and the size of the problem when its solution is based on a specific algorithm. Thus, algorithm A might solve a problem in a time of order N^2, which means that a problem that is 3 times as large would take $3^2 = 9$ times as long to solve, whereas a faster algorithm B might solve the same problem in time of order $N \log N$ (that is, $O(N \log N)$), which means that a problem 3 times as large would take $3 \log 3 = 3.29$ times as long to solve. (A specific example is that when asked to write a program to sort a list of numbers in ascending order, one of the most common programs written by novice programmers involves an $O(N^2)$ algorithm. It takes a somewhat greater degree of algorithmic sophistication to write a program that exhibits $O(N \log N)$ behavior—which can be proven to the best that is possible.)

Such results are important to algorithm design, and all computer programs embody algorithms. Depending on the functional relationship between run time and problem size, a given program that works well on a small set of test data may—or may not—work well (i.e., run in a reasonable time) for a larger set of real data. Theoretical computer science thus imposes constraints on real programs that software developers ignore at their own peril.

Computer scientists and mathematicians derive satisfaction and pleasure from elegance of reasoning, logic, and structure. Being able to explain a phenomenon or account for a dynamical behavior with a simple model is highly valued. The reason for this penchant is clear: the simpler the model, the more likely it is that the tools of analysis can be used to dissect and understand the model fully.

This sometimes means that a tendency to oversimplify overwhelms the need for preserving realistic features, to the dissatisfaction or derision of biologists. Computer scientists, of course, may well perceive a biologist's dissatisfaction as a lack of analytical or theoretical sophistication and an unwillingness to be rigorous, and often fail to recognize the complexity inherent in biological systems. In other cases, the love of elegance leads to fixation with elegant, but irrelevant, models far beyond their value outside the field, simply because the inherent model is clean and simple. In still other cases, the lack of training of computer scientists in eliciting from users the precise nature of their problems has led computer scientists to develop good solutions to problems that are not interesting to most biologists or relevant to real biological phenomena.

By contrast, many—perhaps most—biologists today have a deep skepticism about theory and models, at least as represented by mathematics-based theory and computational models. For example, theoretical biology has a very different status within biology and has often been a poor stepchild to mainstream biology. Results from theoretical biology are often irrelevant to specific biological systems such as a particular species, and even the simplest biological organism is so complex as to render virtually impossible a theoretical analysis based on first principles. Indeed, most biologists have a long-ingrained suspicion of theoretical models that they regard as vastly oversimplified (i.e., almost all of them) and are skeptical of any purported insights that emerge from such models. (Box 10.5 provides some examples of misleading computational and mathematical models of biological phenomena.)

Box 10.5
Some Examples of Oversimplified and/or Misleading Computational and Mathematical Models in Biology

- The Turing reaction-diffusion theory for pattern formation in developmental biology—first suggested by Turing in 1952, and largely dormant until the mid-1970s, this theory, based on an activator-inhibitor system, became a focus of partial differential equations research. Initially, attempts were made to show that diffusion and reaction of the activator-inhibitor type are responsible for the development of real structures in real embryos (stripes or spots, positions of limbs and digits, etc.) However, later work has shown that the biological solution to the pattern formation problem is inelegant and "kludgy", with many "redundant" or "inefficient" parts.[1]
- A senior computer scientist faced the issue of how one might infer the structure of a genetic regulatory network from data on the presence or absence of transcription factors. In a cell, a set of genes interact to produce a protein—and the transcription factors (themselves proteins) influence the rate at which that protein is produced. His initial model of this network was a Boolean circuit, in which the presence or absence of certain factors led to the production of the protein. A typical experimental procedure in a biology lab to probe the nature of this circuit is to observe its behavior by inhibiting the production of some transcription factor and to observe whether or not the protein is produced. The analogous action in the Boolean circuit would be cutting a wire in that circuit. However, this simple analogy failed to model the actual behavior of the biological system because, in many cases, the inhibition of one transcription factor results in another set of proteins that do the same job. Thus, the notion of simple perturbation experiments that can be viewed as analogous to just snipping a wire in a logic circuit is obvious for computer scientists—but turns out to be not particularly relevant to this particular phenomenon.
- The problem of genome sequence assembly involves piecing together a large number of short sequences (fragments) into the correct master sequence. The initial computer scientist formulation of this problem was to find the shortest sequence that would contain a given set of sequences as a consecutive piece. But this formulation of the problem was completely wrong for two reasons. First, the available information on the fragments is sometime erroneous—that is, the data might indicate that a fragment would have a certain base at a given location, but in reality it would have a different base at that location. Second, DNA molecules have a great deal of repeated structure (i.e., the same sequence is typically found multiple times). Thus, the shortest sequence is not biologically plausible because that repeated structure is ignored.
- Amino acids are represented by codons (i.e., triplets of nucleotide bases). Because there are 4 nucleotides, the number of possible codons is 4^3, or 64. But for a long time, only 20 amino acids were known that occur in nature. It turns out that by assuming that the codons overlapped each other and requiring that the coding be unambiguous, only 20 codons are possible. Because of this match, a natural assumption was that an overlapping code was operative in DNA coding. However, experimental data dispelled this notion, indicating instead that multiple codons can represent the same amino acid and further that the codons were not overlapping.

[1] See, for example, G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, "The Segment Polarity Network Is a Robust Developmental Module," *Nature* 406(6792):188-192, 2000. At the same time, the reaction-diffusion approach appears to have nontrivial utility in explaining other biological phenomena, such as certain aspects of microtubule organization (C. Papaseit, N. Pochon, and J. Tabony, "Microtubule Self-organization Is Gravity-dependent," *Proceedings of the National Academy of Sciences* 97(15):8364-8368, 2000).

A related point is that computer scientists tend to assume that universal statements have no exceptions, whereas biologists have learned that there are almost always exceptions to rules. For example, if a biologist says that all crows are black, and one asks about albino crows, the answer will be, "Oh, sure, albino crows are white, but all normal crows are black." The biologist is describing the average case—all standard crows are black—but keeps in the back of his or her mind the exceptional cases. By contrast, the computer scientist wants to know if the crow database he or she is building needs to accommodate anything other than a black crow—and thus, when a computer scientist makes a biological generalization, the biologist will often jump immediately to the exceptional case as a way of dismissing the generalization.

These comments should not be taken to imply that biologists do not use theory at all; in fact, biologists use theory and models in their everyday work. The theory of evolution is among the most powerful of all scientific theories, in the sense that it underlies the scientific understanding of all natural biological phenomena. But because the outcomes of evolutionary processes are driven by a myriad of environmental and chance influences, it is difficult to make measurable or quantitative predictions about specific biological phenomena. In this context, evolution is more of an organizing principle than a predictive formalism.

Perhaps a fairer statement is that many biologists remain to be persuaded of the value of quantitative theory and abstraction on a global basis, although they accept their value in the context of specialized hypothesis, individual probes, or inquiries on a biological process. Biological researchers are beginning to see the potential explanatory value of computational and mathematical approaches—a potential that is less apparent than might be expected because of the very success of an empirical approach to biology that has been grounded in experiment and observation for many decades.

10.3.1.4 Data and Experimentation

As mentioned above, computer scientists and biologists also view data quite differently. For the computer scientist, data usually result from measurements of some computational artifact in use (e.g., how long it takes for a program to run, how many errors a program has). Because these data are tied to artifacts that have been made by human beings, they are as ephemeral and transient as the underlying artifact, which may indeed change in the next revision or release. Because computer science is a science of the artificial, the intellectual process of the computer scientist does not begin with data, but rather with an act of artifact creation, after which measurements can be taken.[66]

Indeed, for the computer scientist, the term "experimental computer science" refers to the engineering and creation of new or improved computational artifacts—hardware or software—as the central objective of intellectual efforts.[67] Engineering has intellectual biases toward model reduction, extracting key elements, and understanding subsystems in isolation before assembling larger structures. The engineering approach also rests on the idea that basic units (e.g., transistors, silicon chips) have repeatable, predictable behavior; that "modules" with specific capability (e.g., switches, oscillators, and filters) can be made from such units; and that larger systems with arbitrary complexity are, in turn, made of such modules.

In contrast, biology today is a data-driven science—and theories and models are created to fit the data. Data, presuming they are accurate, impose "hard" constraints on the biologist in much the same way that results from theoretical computer science impose hard constraints on the computer scientist. Because of the central role that data play in biology, biologists pay a great deal of attention to experi-

[66] This is not to deny that computer scientists often work with large datasets. For example, computer scientists may work with terabytes of textual or image data. But these data are the subjects of manipulation and processing, rather than being tied directly to the performance of the hardware and software artifacts of the computer scientist.

[67] National Research Council, *Academic Careers for Experimental Computer Scientists and Engineers*, National Academy Press, Washington, DC, 1994.

mental technique and laboratory procedure and instrumentation—much more so than most computer scientists pay to the comparable areas in computer science. Thus, a computer scientist with insufficient awareness of experimental design may not be accustomed to or even aware of techniques of formal model or simulation validation.

In addition, biology has not traditionally looked to engineering for insight or inspiration. For example, proteins come in an endless variety with many variations and do not necessarily have straightforward analogues to engineering parts. Experimental biologists often focus on discovering new pieces of cellular machinery and on how defective behavior stems from broken or missing pieces (e.g., mutations). Experimental work is aimed at proving or disproving specific hypotheses, such as whether or not a particular biochemical pathway is relevant to some cellular phenomena.

The training that computer scientists receive also emphasizes general solutions that give guarantees about events in terms of their worst-case performance. Biologists are interested in specific solutions that relate to very particular (although voluminous) datasets. (A further complication is that biological data are often erroneous and/or inconsistent, especially when collected in large volume.) By recognizing and exploiting special characteristics of biologically significant datasets, special-purpose solutions can be crafted that function much more effectively than general-purpose solutions. For example, in the problem of genomic sequence assembly, it turns out that by exploiting the information available concerning the size of fragments, the number of choices for where a fragment might fit is sharply restricted.

The central role that experimental data plays in biology is responsible for the fact that, to date, computer scientists have been able to make their most important contributions in areas in which the details of some biological phenomena can be neglected to some important extent. Thus, the abstraction of DNA as merely a string of characters derived from a four-letter alphabet is a very powerful notion, and considerable headway in genomics can be made knowing little else. To be sure, there are experimental errors to take into account, and a model of the noisiness of the data must be developed, but the underlying problem is pretty clear to a computer scientist.

On the other hand, as the discussion in Section 4.4.1 makes clear, there are limits to this abstraction that arise from just such "details." Also, proteomics—in which the three-dimensional structure of a protein, rather than the linear sequence, determines its function—presents even greater challenges. To understand the geometry of a three-dimensional structure, discrete mathematics—the stock in trade of the computer scientist—is far less useful than continuous mathematics.[68] Furthermore, the properties and characteristics of the specific amino acids in a protein matter a great deal to its structure and function, whereas the various nucleotide bases are more or less equivalent from an informational standpoint. In short, proteomics involves a much more substantial body of domain knowledge than does genomics.

One illustration related by a senior computer scientist working in biology is his original dream that, with enough data,[69] it would be computationally straightforward to understand the mechanisms of gene regulation. That is, with sufficient data on regulatory pathways, cascades, gene knockouts, expression levels, and their dependencies on environmental factors, how genetic regulatory networks work would become reasonable clear. With the hindsight of several years, he now believes that this dream was hopelessly naïve in that it did not account for the myriad exceptions and apparent special cases inherent in biological data that make the biologist's intellectual life very complicated indeed.

Finally, consider that many biologists are suspicious—or at least not yet persuaded—of the value and importance of high-throughput measurement of biological systems (Section 7.2). Because many biologists were educated and worked in an era in which data were scarce, experiments in biology have historically been oriented toward hypothesis testing. High-throughput data collection drives in the opposite direction,

[68] The reason is that geometric descriptions naturally involve continuous variables such as lengths and angles, and functions of those variables.

[69] Richard Karp, University of California, Berkeley, personal communication, July 29, 2002.

and relies on the ability to sift through and recognize patterns in large volumes of data whose meaning can then be inferred. Of course, predictions that emerge from the analysis of large volumes of data must still be verified one at a time, and science is today far from the point at which such analysis would, by itself, provide reliable biological conclusions. Nevertheless, such analysis can play an important role in suggesting interesting hypotheses and thus expand the options available for biological exploration.

10.3.1.5 A Caricature of Intellectual Differences

A number of one-liners that can be used to encapsulate the differences described above, though as with all one-liners, there is considerable oversimplification. Here are four:

- The goal of computer science (CS) is to develop solutions that can be useful in solving many problems, while the goal of biology is to look for solutions to individual and specific problems.
- Computer science is driven by the development of method and technique, while biology is driven by experiment and data.
- Computer scientists are trained to search for boundary conditions and constraints, whereas biologists are trained to seek signal in the noise of their experimental data.
- Computer scientists are trained to take categorical statements literally, whereas biologists use them informally.

10.3.2 Differences in Culture

Another barrier at the BioComp interface is cultural. Each field has its own cultural style, and what seems obvious to practitioners in one field may not be obvious to those in the other. Consider, for example, differences between computer science and biology. Before PowerPoint became ubiquitous to both fields, computer scientists tended to use overhead transparencies in visiting lectures, while biologists tended to use 35 mm slides. Computer science, as a discipline, can often be pursued while working at home, whereas biological lab work requires being "in the office" to a far greater extent—a computer scientist who is away from the lab may well be seen by biologists as "not being around enough" or "not being a team player." Computer scientists are accustomed to having their own office space, while biologists (especially postdoctoral associates) work out of their labs and rarely have their own offices until they achieve an appropriate seniority.

Such differences are in some sense trivial, but they do suggest the reality of different cultures, and it is helpful to explore some other differences that are not so trivial. One of the most important differences is that of intellectual style: the discussion in Section 10.3.1 would suggest that biologists (especially those untrained in quantitative sciences) may well distrust the facile approaches and oversimplified models of computer scientists or mathematicians unfamiliar with the complexities of living things, and the computer scientist may well regard the biologist as obsessed with details and molecular parts lists rather than the qualitative or quantitative whole. This section explores some issues that lie outside the domain of intellectual style.

10.3.2.1 The Nature of the Research Enterprise

When practitioners from two fields collaborate, each brings to the table the values that characterize each field. Given the importance that biologists place on the understanding of specific biological phenomena of interest, they place the highest value on answers that are specific to those phenomena. Biologists want "the answer," and they are interested in details of a computational model only insofar as they have an effect on the answer; for the most part, they care far less about a hypothetical biological phenomenon than about explaining the data obtained from experiment. Computer scientists and mathematicians, in contrast, are interested in the parameters of a model or a solution and in ways to improve it, characterize it, understand it better, or make it more generally applicable to other problems. Thus, the biologist will likely be interested in the results of a model run on the single dataset of interest, while the computer scientist will want to run hundreds or thousands of datasets to better analyze the behavior of the model, and mathematicians will want to explore the limits of a model's applicability.[70]

An example of this cultural difference is illustrated in the history of the Gene Ontology (GO) discussed in Chapter 4. Begun in 1998 as a collaboration between researchers responsible for three model organism databases (FlyBase [Drosophila], the Saccharomyces Genome Database, and the Mouse Genome Database), GO collaborators sought to develop structured, controlled vocabularies that describe gene products in terms of their associated biological processes, cellular components, and molecular functions in a species-independent manner. In their work, these researchers have apparently not made extensive use of the (mostly domain-independent) theoretical contributions of computer science from the last 20 years, but rather have reinvented much of that work on their own. The reason for this reinvention, offered by one knowledgeable observer, is that they were unable to find computer scientists with appropriately specialized experience who were willing to sacrifice their quest for general applicability to develop a functional, usable system.[71]

A related point is that in academia, research computer scientists have very little motivation to take a software implementation beyond the prototype stage. That is, they may have developed a powerful algorithm that is likely to be useful in many biological contexts, implemented a prototype software system based on this algorithm, and convincingly demonstrated its utility in a few cases. But because most of the intellectual credit inheres in the prototype (e.g., papers for publication and promotions), research computer scientists have little motivation to move from the prototype system, which can generally be used only by those familiar with the quirks of its operation, to a more robust system that can be used by the broader community at large. Because going from prototype to broadly usable system is generally a time-intensive process, many powerful methods are not available to the biology community.

Similar considerations apply in the biology community with respect to data curation. Intellectual credit for academic biologists inheres in the publication of primary data, rather than in any long-term follow-up to ensure that the data are useful to the broader community. (Indeed, if the data are not made useful to the broader community, the researcher originally responsible for the data gains the competitive advantage of being the only one, or one of a few, able to use them.) This suggests that cultural incentives for data curation (or the lack thereof) have to be altered if data curation is to become a more significant activity in the research community.[72]

[70] These differences in perspective are also found at the interface of medical informatics and bioinformatics. For example, Altman notes that "the pursuit of bioinformatics and clinical informatics together is not without some difficulties. Practitioners in clinical medicine and basic science do not instantly understand the distinction between the scientific goals of their domains and the transferability of methodologies across the two domains. *They sometimes question whether informatics investigators are really devoted to the solution of scientific problems or are simply enamored of computational methodologies of unclear significance* [emphasis added]." To reduce these tensions, Altman argues—similarly to the argument presented in this report—that "informatics investigators (and their students) be able to work collaboratively with physicians and scientists in a manner that makes it clear that the creation of excellent, well-validated methods for solving problems in these domains is the paramount goal." See R.B. Altman, "The Interactions Between Clinical Informatics and Bioinformatics: A Case Study," *Journal of the American Medical Informatics Association* 7(5):439-443, 2000.

[71] Russ B. Altman, Stanford University, personal communication, December 16, 2003.

[72] One approach that has been used to support data annotation and curation activities is the data jamboree. In November 1999, the Celera Corporation hosted an invitation-only event ("the jamboree") in which participants worked for two weeks at annotating and correcting data from the *Drosophila melanogaster* genome. By all accounts a successful event that resulted in the publication of the complete sequence as well as appropriate annotations (see M.D. Adams, S.E. Celniker, R.A. Holt, C.A. Evans, J.D. Gocayne, P.G. Amanatides, S.E. Scherer, et al., "The Genome Sequence of *Drosophila melanogaster*," *Science* 287(5461):2185-2195, 2000) the event featured a very informal atmosphere that promoted social connection and interaction as well as a work environment conducive to the task. The emergence of some level of community curation on Amazon and eBay may also provide some useful hints on how to proceed. In these efforts, community assessment is allowed, but there's no overall review of the quality of the assessment. Nonetheless, users have access to a diverse collection of assessments of and can do their own meta-quality control by deciding which of the reviewers to believe. This model does scale with increasing database size, although consistent curation is hardly guaranteed. It is an open question worth some investigation as to whether community commentary (perhaps supported with an appropriate technological infrastructure) could result in meaningful data curation.

10.3.2.2 Publication Venue

Although both biologists and computer scientists in academia seek to make their work known to their respective peer communities, they do so in different ways. For many areas within computer science, refereed conferences are the most prestigious places to publish leading research.[73] In biology, by contrast, conference publications are often considered less prestigious, and biologists tend to prefer journal publications. Computer scientists often write abstracts in such a way as to entice the reader to read the full paper, whereas biologists often write abstracts in such a way that the reader need not read the full paper. In publishing work that refers to a new tool, a computer scientist may be more likely to reference a Web site where the tool can be found, while a biologist may be more likely to reference a paper describing the tool.

This difference in publication venues strongly affects attempts at collaboration. Academic biologists often do not understand refereed conferences, and computer scientists often think of journals as mere repositories of papers, rather than a means of communicating results. Given that publication is the primary output of academic research, this disagreement can be very disturbing and can have important inhibitory effects on collaboration.

10.3.2.3 Organization of Human Resources

While it is clear to all parties from the start that that the professional expertise of the biologist is needed to do good work in computational biology, a view that equates computer science with programming can lead biologists to underestimate the intellectual capabilities on the computer science side necessary for computation-intensive biology problems. Thus, many biologists who do see rewards in bridging the interdisciplinary gap (especially in academia) tend to prefer doing so in their own labs, by hiring postdoctoral fellows from physics or computer science to work on their problems, keeping these ventures "in the family," rather than by establishing partnerships with more established computer scientists.

Such an approach has advantages and disadvantages. An advantage is that postdoctoral fellows with good quantitative and computational background can be exposed to the art of biological experimentation and interpretation as a part of their postdoctoral training, the result of which can be the nurturing of young interdisciplinary scientists. One disadvantage is that by engaging individuals at the beginning of their careers, the biologist is deprived of the intellectual maturity and insight that generally accompanies more seasoned computer scientists—and such maturity and insight may be most necessary for making headway on complex problems.

The integration of computational expertise into a biological research enterprise can be undertaken in different ways. For example, in some instances, a group of computer scientists can work with a group of biologists, each bringing its own computational approach to the biological problem. In other cases, a single individual with computational expertise (e.g., a postdoc) can work in an otherwise "pure" biological group, offering expertise in math, modeling, and programming. A second dimension of integration is that the bioinformaticist can play a role as a team member who engages equally with everyone else on the research team or as a "bridge" that serves as intermediate between practitioners of various disciplines. These two roles are not mutually exclusive, although the first seems to be more common.

10.3.2.4 Devaluing the Contributions of the Other

A sine qua non of interdisciplinary work is that intellectual credit be apportioned appropriately. In some cases known to the committee, the expertise of scientists from a nonbiological discipline has

[73] Such a practice arose because computer science is a fast-moving field, with a tradition of sharing discovery by online demonstration and discussion. Conferences were originally formed as a way to talk together, in person, and with relatively fast publication of results for the requirements of academia. In this context, journal publication would have been much too slow.

been used, and yet joint authorship or due credit withheld. In one story, a respected biologist held regular discussions with an excellent mathematician colleague. The biologist assigned a theoretical project on a topic already worked out by the mathematician to an in-house physics postdoctoral fellow instead of pursuing joint work with the mathematician. The results of this in-house work fell short of what was possible or desirable but displaced other serious attempts at theoretical analysis of an interesting problem.

This example suggests a view of mathematics and computer science that is ancillary and peripheral to the "real" substance of biology. The fact that computing and mathematics have developed powerful tools for the analysis of biological data makes it easy for biologists to see the computer scientist as the data equivalent of a lab technician. However, although programming is an essential dimension of most computer scientists' backgrounds, it does not follow that the primary utility of the computer scientist is to do programming. Algorithm design, to take one example, is not programming, but because algorithms must be implemented as a computer program, it is easy to confuse the two.

In other cases known to the committee, the expertise of biological scientists has been denigrated by those in computing. For example, computer scientists sometimes view a successful biological experiment as one that "merely" produces more data and do not appreciate the fundamental creative act required to devise the appropriate experiment. This attitude suggests a view of biology in which the "real" science resides in the creation of a theory or a computational model, and data are merely what is needed to populate the model.

What accounts for such attitudes? The committee believes one contributing factor is not much different than loyalty to one's discipline. Professionals in one discipline quite naturally come to believe that the ways in which they have learned to see their discipline have inherent advantages (if they did not, they would not be part of the discipline), and challenges to the intellectual paradigms they bring to the subject may well be met with a certain skepticism.

A second point to consider is that interdisciplinary work is not necessarily symmetric. This is especially true in the mix of academic research activity vis a vis applied or technical support activity. That is, it is often possible to identify one field as being the side where research advances are occurring and the other as applying some kind of support. In some cases, Ph.D.-level research in computer science can be enriched by what is routinely taught in undergraduate classes in biology, and vice versa.

For example, individuals pursuing cutting-edge research in database design may be interested in finding data models to exercise their design. They are interested in finding domain experts to help them better understand the complexities of a certain interesting problem domain, such as biology, but these database researchers see the data and the insights coming from the biologist as helping to define the problem, but as having little to do with finding the solution. Similarly, biologists may be investigating a new topic in biology and need quantitative or logistical or algorithmic help to accomplish the research, but they feel the real intellectual contribution—to biology—comes from their insights on the biological side.

The primary exception to these scenarios is where a research group in computer science gets teamed up with a research group in biological science. In such instances, the relationship can be truly symmetrical. Both parties benefit from a symbiotic relationship. Both yield practical value to the other, while gaining theoretical value for themselves. Both operate at an equivalent level of intellectual contribution. Both gain an equivalent level of real research coming out of the activity.

10.3.2.5 Attitudinal Issues

Biology laboratories are increasingly dependent on various forms of information technology. High-throughput instrumentation generates large volumes of data very quickly. Computer-based databases are the only way to keep track of a biological literature that is growing at exponential rates. Computer programs are increasingly needed to assemble and understand biological data derived from experiments or resident in databases.

With such dependence on IT, it would not be surprising if individuals who are especially knowledgeable about information technology were necessary to keep these laboratories running at high efficiency. However, computer scientists are very wary of being put into the role of technician or programmer. Computer science researchers, facing this prospect from various research disciplines, can be sensitive about wanting respect for the fundamental research advances they bring to the table.[74]

The roles of intellectual collaborator and co-equal partner are largely incompatible with the role of technician, and it is understandable that a computer scientist would want to be treated as a coequal. At the same time, a certain amount of humility and respect is also necessary. That is, the computer scientist must refrain from jumping to conclusions, must be willing to learn the facts and contemplate biological data seriously, and must not work solely on the refined abstraction problem. It may well be necessary for the computer scientist to do some mundane things to earn the confidence of the biologist partner before being able to do more interesting things.

The biologist has a role to play in facilitating partnership as well. For example, the biologist must understand that the computer scientist (especially one at the beginning of his or her career) wants to do work of publishable quality as well—work that will earn the respect of colleagues in computer science. As suggested above, programming generally does not meet this test. A second important point is to recognize without condescension the fact that many (most?) computer scientists have very little experience or familiarity with either biological concepts or data. Still a third point is the recognition that while primary data generation and experiment remain important to the life sciences, analytical work on existing data can be every bit as valuable—bioinformatics is not simply "taking someone else's data." This last point suggests a more subtle risk in partnerships—that a person with specialized skills may be regarded as a technician or a stand-alone consultant rather than as a true collaborator.

10.3.3 Barriers in Academia

One important venue for research at the BioComp interface is academia. Universities can provide infrastructure for work in this area, but institutional difficulties often arise in academic settings for work that is not traditional or easily identified with existing departments. These differences derive from the structure and culture of departments and disciplines, and lead to scientists in different disciplines having different intellectual and professional goals and experiencing different conditions for their career success. Collaborators from different disciplines must find and maintain common ground, such as agreeing on goals for a joint project, but also respect one another's separate priorities, such as having to publish in primary journals to present at particular conferences or to obtain tenure in their respective departments according to those departmental criteria. Such cross-pressures and expectations from the home departments and disciplinary colleagues remain even if the participants develop similar goals for a project.

10.3.3.1 Academic Disciplines and Departmental Structure

Universities are structured around disciplinary departments and often have considerable difficulty in supporting and sustaining interdisciplinary work. Neither fish nor fowl, the interdisciplinary researcher is often faced with the formidable task of finding an intellectual home within the university that will take the responsibility for providing tenure, research space, start-up funding, and the like. The essential problem is that a researcher working at the interface between fields X and Y is often doing work that does not fall clearly within the purview of either Department X or Department Y. When budgets are expanding and

[74] It is useful to note that research laboratories in both biology and computer science employ technicians and programmers, and such individuals serve very useful functions in each kind of laboratory. But the role of a lab technician in a biology laboratory or programmer in a computer laboratory is quite different from the role of the senior scientist who directs the biology or computer laboratory.

resources flush, it is easy for Department X or Department Y to take a risk on an interdisciplinary scholar. But as is more often the case today, when resources are scarce, each department is much more likely to want someone who fits squarely within its traditional departmental definitions, and any appointment that goes to an interdisciplinary researcher is seen as a lost opportunity.

For example, tenure letters may be requested from traditional researchers in the field for an interdisciplinary worker; despite great success, the tenure letters may well indicate that they were unfamiliar with the candidate's work. Graduate students seeking interdisciplinary training but nominally housed in a given department may have difficulty taking that department's qualifying exam, because their training is significantly different from mainstream students.

Another dimension of this problem is that publication venues often mirror departmental structures. Thus, it may be difficult to find appropriate venues for interdisciplinary work. That is, the forms of output and forums of publication for the interdisciplinary researcher may be different than for either Department X or Department Y. For example, even within computer science itself, experimental computer scientists that focus on system building often lack a track record of published papers in refereed journals, and tenure and promotion committees (often university-wide) that focus on such records for most other disciplines in the university have a hard time evaluating the worthiness of someone whose contributions have taken the form of software that the community has used extensively or presentations at refereed conferences. Even if biologists are aware in principle of such "publication" venues, they may not be aware that such conferences are heavily refereed or are sometimes regarded as the most prestigious of publication venues. Also, prestigious journals known for publishing biology research are often reluctant to devote space to papers devoted to computational technique or methodology if it does not include specific application to an important biological problem (in which case the computational dimensions are usually given a peripheral rather than primary status).

Further, the academic tenure and promotion system is biased toward individual work (i.e., work on a scale that a single individual can publish and receive credit for). However, large software systems—common in computer science and bioinformatics—are constructed by teams. Although small subsystems can be developed by single individuals, it is the whole system that provides primary value, and university-based research that is usually driven by a single-authored Ph.D. thesis or single faculty members is not very well suited to such a challenge.[75]

Finally, in most departments, it is the senior faculty that are likely to be the most influential with regard to the allocation of resources—space, tenure, personnel and research assistant support, and so on. If these faculty are relatively uninformed or disconnected from ongoing research at the BioComp interface, the needs and intellectual perspectives of interface researchers will not be fully taken into account.

10.3.3.2 Structure of Educational Programs

Stovepiping is also reflected in the structure of educational programs. Stovepiping refers to the tendency of individual disciplines to have different points of view on what to teach and how to teach it, without regard for what goes on in other disciplines. In some cases, the methods of the future are still undeveloped, or are undergoing revolution, so that suitable texts or syllabi are not yet available. Further, like individual researchers, departments tend to be territorial, protective of their realms, and insistent on ever-growing specialized course load requirements for their own students. This discourages or precludes cross-discipline shopping. Novel training creates a need for reeducation of faculty to change the design of old curricula and modernize the teaching. These changes take time and energy, and require release time from other academic burdens, whether administrative or teaching.

[75] C. Koch, "What Can Neurobiology Teach Computer Engineers?," Division of Biology and Division of Engineering and Applied Science, California Institute of Technology, January 31, 2001, position paper to National Research Council workshop, available at http://www7.nationalacademies.org/compbio_wrkshps/Christof_Koch_Position_Paper.doc.

Related to this point is the tension between breadth and depth. Should an individual trained in X who wishes to work at the intersection of X and Y undertake to learn about Y on his or her own, or seek to collaborate with an individual trained in Y? Leading-edge research in any field requires deep knowledge. But work at the interface of two disciplines draws on both of them, and it is difficult to be deep in both fields; thus, Ph.D.-level expertise in both computer science and biology may be unrealistic to expect. As a result, collaboration is likely to be necessary in all but extraordinary cases.

Thus, what is the right balance to be struck between collaboration and multiskilling of individuals? There is no hard-and-fast answer to this question, but the answer necessarily involves some of both. Even if "collaboration" with an expert in Y is the answer, the individual trained in X must be familiar enough with Y to be able to conduct a constructive dialogue with the expert in Y, asking meaningful questions and understanding answers received. At the same time, it is unlikely that an expert in X could develop in a reasonable time expertise in Y comparable to that of a specialist in Y, so some degree of collaboration will inevitably be necessary.

This generic answer has implications for education and research. In education, it suggests that students are likely to benefit from presentations by both types of expert (in X and in Y), and the knowledge that each expert has of the other's field should help to provide an integrated framework for the joint presentations. In research, it suggests that research endeavors involving multiple principal investigators (PIs) are likely to be more successful on average than single-PI endeavors.

Stovepiping can also cause problems for graduate students who are interested in dissertation work, although for graduate students these problems may be less severe than for faculty. Some universities make it easier for graduate students to do interdisciplinary work by allowing a student's doctoral work to be supervised by a committee composed of faculty from the relevant disciplines. However, in the absence of a thesis supervisor whose primary interests overlap with the graduate student's work, it is the graduate student himself or herself who must be the intellectual integrator. Such integration requires a level of intellectual maturity and perspective that is often uncommon in graduate students.

The course of graduate-level education in computing and in biology is different in some ways. In biology, students tend to propose thesis topics earlier in their graduate careers, and then spend the remainder of their time doing the proposed research. In computer science (especially more theoretical aspects), in contrast, proposals tend to come later, after much of the work has been done. Computer science graduates do not usually obtain postdoctoral positions, more commonly moving directly to industry or to a tenure-track faculty position. Receiving a postdoctoral appointment is often seen as a sign of a weak graduate experience in computer science, making postdoctoral opportunities in biology seem less attractive.

10.3.3.3 Coordination Costs

In general, the cost of coordinating research and training increases with interdisciplinary work. When computer scientists collaborate with biologists, they also are likely to belong to different departments or universities. The lack of physical proximity makes it harder for collaborators to meet, coordinate student training, and share physical resources, and studies indicate that distance has especially strong effects on interdisciplinary research.[76]

Recognizing the importance of reducing distances between collaborators, Stanford University's Bio-X program is designed specifically to foster communication campus-wide among the various disciplines in biosciences, biomedicine, and bioengineering. The Clark Center houses meeting rooms, a shared visualization chamber, low-vibration workspace, a motion laboratory, two supercomputers, the

[76] J. Cummings and S. Kiesler, *KDI Initiative: Multidisciplinary Scientific Collaborations*, report to National Science Foundation, 2003, available at http://netvis.mit.edu/papers/NSF_KDI_report.pdf; R.E. Kraut, S.R. Fussell, S.E. Brennan, and J. Seigel, "Understanding Effects of Proximity on Collaboration: Implications for Technologies to Support Remote Collaborative Work," pp. 137-162 in *Distributed Work,* P.J. Hinds and S. Kiesler, eds., MIT Press, Cambridge, MA, 2002.

small-animal imaging facility, and the Biofilms center. Other core shared facilities available to the Stanford research community include a bioinformatics facility, a magnetic resonance facility, a microarray facility, a transgenic animal facility, a cell sciences imaging facility, a product realization lab, the Stanford Center for innovation in in vivo imaging, a tissue bank, and facilities for cognitive neuroscience, mass spectrometry, electron microscopy, and fluorescence-activated cell sorting.[77]

Interdisciplinary projects are often bigger than unidisciplinary projects, and bigger projects increase coordination costs. Coordination costs are reflected in delays in project schedules, poor monitoring of progress, and an uneven distribution of information and awareness of what others in the project are doing. Coordination costs also reduce people's willingness to tolerate logistical problems that might be more tolerable in their home contexts. Furthermore, they increase the difficulty of developing mutual regard and common ground, and they lead to more misunderstandings.[78]

Coordination costs can be addressed in part through changes in technology, management, funding, and physical resources. But they can never be reduced to zero, and learning to live with greater overhead in conducting interdisciplinary work is a sine qua non for participants.

10.3.3.4 Risks of Retraining and Conversion

Retraining or conversion efforts almost always entail reduced productivity for some period of time. This fact is often viewed with dread by individuals who have developed good reputations in their original fields, and who may worry about sacrificing a promising career in their home field while entering at a disadvantage in the new one. These concerns are especially pronounced when they involve individuals in midcareer rather than recently out of graduate school.

Such fears often underlie the failure of individuals seeking to retool themselves to commit themselves fully to their new work. That is, they seek to maintain some degree of ties to their original fields—some research, some keeping up with the literature, some publishing in familiar journals, some going to familiar conferences, and so on. These efforts drain time and energy from the retraining process, but more importantly they may inhibit the necessary mind-set of success and commitment in the new domain of work. (On the other hand, keeping a foot in their old fields could also be viewed as a rational hedge against the possibility that conversion may not be successful in leading to a new field of specialization. Moreover, maintaining the discipline of continual output is a task that requires constant practice, and one's old field is likely to be the best source of such output.)

10.3.3.5 Rapid But Uneven Changes in Biology

Biology is an enormously broad field that contains dozens of subfields. Over the past few decades, these subfields have not all advanced or prospered equally. For example, molecular and cell biology have received the lion's share of biological funding and prestige, while subfields such as animal behavior or ecology have faired much less well. Molecular and cell biology (and more recently genomics, proteomics, and neuroscience) have swept through as departments modernize, in a kind of "bandwagon" effect, leaving some of the more traditional subfields to lie fallow because promising young scholars in those subfields are unable to find permanent jobs or establish their careers due to these shifts.

Moreover, prospering subfields are highly correlated with the use of information technology. Such a close association of IT with prospering fields is likely to exacerbate lingering resentments from non-prospering subfields toward the use of information technology.

[77] For more information see http://biox.stanford.edu/.

[78] J. Cummings and S. Kiesler, "Collaborative Research Across Disciplinary and Institutional Boundaries," *Social Studies of Science*, in press, available at http://hciresearch.hcii.cs.cmu.edu/complexcollab/pubs/paperPDFs/cummings_collaborative.pdf.

10.3.3.6 Funding Risk

Tight funding environments often engender in researchers a tendency to behave conservatively and to avoid risk. That is, unless special care is taken to encourage them in other directions (e.g., through special programs in the desired areas), researchers seeking funding are likely to pursue avenues of intellectual inquiry that are likely to succeed. Such researchers are therefore strongly motivated to pursue work that differs only marginally from previous successful work, where paths to success can largely be seen even before the actual research is undertaken. These pressures are likely to be exacerbated for senior researchers with successful and well-respected groups and hence many mouths to feed. This point is addressed further in Section 10.3.5.3.

10.3.3.7 Local Cyberinfrastructure

Section 7.1 addressed the importance of cyberinfrastructure to the biological research enterprise taken as a whole. But individual research laboratories need to be able to count on the local counterpart of community-wide cyberinfrastructure. Institutions generally provide electricity, water, and library services as part of the infrastructure that serves individual resident laboratories. But information and information technology services are increasingly as important to biological research as these more traditional services, and thus it makes sense to consider that they might be provided as a part of the local infrastructure.

On the other hand, regarding computing and information services as part of local infrastructure has institutional implications. For example, one important issue is providing centralized support for decentralized computing. Useful scientific computing must be connected to a network, and networks must interact and must be run centrally, but nonetheless, scientific computing must be accomplished in the way scientific instruments are used, that is, very much under the control of the researcher. How can institutions develop a computing infrastructure that delivers the cost effectiveness and the robustness and the reliability of well-run centralized systems while at the same time delivering the flexibility necessary to support innovative scientific use? In many research institutions, managers of centralized computing regard researchers as cowboys uninterested in exercising any discipline for the larger good, while researchers regard the managers of centralized computing as bureaucrats who are disinterested in the practice of science. Though neither of these caricatures is correct, these divergent views of how computing should effectively be deployed in a research organization will continue to exist unless the institution takes steps to reconcile them.

10.3.4 Barriers in Commerce and Business

10.3.4.1 Importance Assigned to Short-term Payoffs

In a time frame roughly coincident with the dot-com boom, commercial interest in bioinformatics was very high—perhaps euphoric in retrospect. Large, established, biotech-pharmaceutical companies, genomics-era drug discovery companies, and tiny start-ups all believed in the potential for bioinformatics to revolutionize drug design and even health care, and these beliefs were mirrored in very high stock prices.

More recently, market valuations of biotech firms have dropped along with the rest of the technology sector, and these more recent negative trends have affected the prevailing sentiment about the value of bioinformatics for drug design, at least for the short term. Although the human genome sequencing is complete, only a handful of drugs now in the pipeline stemmed from bioinformatic analysis of the genome. Bioinformatics does not automatically lead to marketable "blockbuster" drugs, and drug companies have realized that the primary bottlenecks involve biological knowledge: not enough is known of the overall biological context of gene expression and gene pathways. In the words of one person at a 2003 seminar, "This is work for [biological] scientists, not bioinformaticists." For this reason, further large-scale business investment in bioinformatics—and indeed for any research with a long time horizon—is difficult to justify on the basis of relatively short-term returns and thus is unlikely to occur.

These comments should not be taken to imply that bioinformatics and information technology have not been useful to the pharmaceutical industry. Indeed, bioinformatics has been integrated into the entire drug development process from gene discovery to physical drug discovery, even to computer-based support for clinical trials. Also, there is a continuing belief that bioinformatics (e.g., simulations of biological systems in silico and predictive technologies) will be important to drug discovery in the long term.

10.3.4.2 Reduced Workforces

The cultural differences between life scientists and computer scientists described in Section 10.3.2 have ramifications in industry as well. For example, a sense that bioinformatics is in essence technical work or programming in a biological environment leads easily to the conclusion that the use of formally trained computer scientists is just an expensive way of gaining a year or two on the bioinformatics learning curve. After all, if all of the scientists in the company use computers and software as a matter of course and can write SQL (Structured Query Language) queries themselves, why should the company have on its payroll a dedicated bioinformaticist to serve as an interface between scientists and software? In a time of expansion and easy money, perhaps such expenditures are reasonable, but when cash must be conserved, such a person on staff seems like an expensive luxury.

10.3.4.3 Proprietary Systems

In all environments, there is often a tension between systems built in a proprietary manner and those built in an open manner, and the bioinformatics domain is no exception. Proprietary systems are often not compatible or interoperable with each other, and yet vendors often think that they can maximize revenues through the use of such systems. This tendency is particularly vexing in bioinformatics where integration and interoperability have so much value for the research enterprise. Standards and open application programming interfaces are one approach to addressing the interoperability problem. But as is often the case, many vendors support standards only to the extent that they are already incorporated into existing product lines.

10.3.4.4 Cultural Differences Between Industry and Academia

As a general rule, private industry has done better than academia in fostering and supporting interdisciplinary work. The essential reason is that disciplinary barriers tend to be lower and teamwork is emphasized when all are focused on the common goals of making profits and developing new and useful products. By contrast, the coin of the realm in academic science is individual recognition for a principal investigator as measured by his or her publication record.

This difference appears to have consequences in a variety of areas. For example, expertise related to laboratory technique is important to many areas of life sciences research. In an industrial setting, this expertise is highly valued, because individuals with such expertise are essential to the implementation of processes that lead to marketable products. These individuals receive considerable reward and recognition in an industrial setting. Although such expertise is also necessary for success in academic research, lab technicians rarely—if ever—receive rewards that are comparable to the rewards accrued by the principal investigator.

Related to this is the matter of staffing a laboratory. In today's job environment, it is common for a newly minted Ph.D. to take several postdoctoral positions. If in those positions an individual does not develop a sufficient publication record to warrant a faculty position, he or she is for all intents and purposes out of the academic research game—a teaching position may be available, but taking a position that primarily involves teaching is not regarded as a mark of success. However, it is exactly individuals with such experience that are in many instances the backbone of industrial laboratories and provide the continuity that is needed for a product's life cycle.

The academic drive for individual recognition also tends to inhibit collaboration. Academic research laboratories can and do work together, but it is most often the case that such arrangements have to be negotiated very carefully. The same is true for large companies that collaborate with each other, but such companies are generally much larger than a single laboratory and intracompany collaboration tends to be much easier to establish. Thus, the largest projects involving the most collaborators are found in industry rather than academia.

Even "small" matters are affected by the desire for individual recognition. For example, academic laboratories often prepare reagents according to a lab-specific protocol, rather than buying standardized kits. The kit approach has the advantage of being much less expensive and faster to put into use, but often does not provide exactly the functionality that custom preparation offers. That is, the academic laboratory has arranged its processes to require such functionality, whereas an industrial laboratory has tweaked its processes to permit the use of standardized kits.

The generalization of this point is that because academic laboratories seek to differentiate themselves from each other, the default position of such laboratories is to eschew standardization of reagents, or of database structure for that matter. Standardization does occur, but it takes a special effort to do so. This default position does not facilitate interlaboratory collaboration.

10.3.5 Issues Related to Funding Policies and Review Mechanisms

As noted in Section 10.2.5.2, a variety of federal agencies support work at the BioComp interface. But the nature and scale of this support vary by agency, in terms of the procedures for making decisions about what proposals are worthy of support.

10.3.5.1 Scope of Supported Work

For example, although the NIH does support a nontrivial amount of work at the BioComp interface, its approach to most of its research portfolio, across all of its institutes and centers, focuses on hypothesis-testing research—research that investigates well-isolated biological phenomena that can be controlled or manipulated and hypotheses that can be tested in straightforward ways with existing methods. This focus is at the center of reductionist biology and has undeniably been central to much of biology's success in the past several decades.

On the other hand, the nearly exclusive focus on hypothesis testing has some important negative consequences. For example, experiments that require breakthrough approaches are unlikely to be directly supported. Just as importantly, advancing technology that could facilitate research is almost always done as a sideline. This has had a considerable chilling effect in general on what could have been, but the impact is particularly severe for implementation of computational technologies in biological sciences. That is, in effect as a cultural aspect of modern biological research, technology development to facilitate research is not considered real research and is not considered a legitimate focus of a standard grant. Thus, even computing research that would have a major impact on the advancement of biological science is rarely done (Box 10.6 provides one example of this reluctance).

It is worth noting two ironies. First, it was the Department of Energy, rather than the NIH, that supported the Human Genome Project. Second, the development of technology to conduct polymerase chain reaction (PCR)—a technology that is fundamental to a great deal of biological research today and was worthy of a Nobel Prize in 1993—would have been ineligible for funding under traditional NIH funding policy.

Box 10.6
Agencies and High-risk, High-payoff Technology Development

An example of agency reluctance to support technology development of the high-risk, high-payoff variety is offered by Robert Mullan Cook-Deegan:[1]

> In 1981, Leroy Hood and his colleagues at Caltech applied for NIH (and NSF) funding to support their efforts to automate DNA sequencing. They were turned down. Fortunately, the Weingart Institute supported the initial work that became the foundation for what is now the dominant DNA sequencing instrument on the market. By 1984, progress was sufficient to garner NSF funds that led to a prototype instrument two years later. In 1989, the newly created National Center for Human Genome Research (NCHGR) at NIH held a peer-reviewed competition for large-scale DNA sequencing. It took roughly a year to frame and announce this effort and another year to review the proposals and make final funding decisions, which is a long time in a fast-moving field. NCHGR wound up funding a proposal to use decade-old technology and an army of graduate students but rejected proposals by J. Craig Venter and Leroy Hood to do automated sequencing. Venter went on to found the privately funded Institute for Genomic Research, which has successfully sequenced the entire genomes of three microorganisms and has conducted many other successful sequencing efforts; Hood's groups, first at Caltech and then at the University of Washington, went on to sequence the T cell receptor region, which is among the largest contiguously sequenced expanses of human DNA. Meanwhile, the army of graduate students has yet [in 1996, eds.] to complete its sequencing of the bacterium *Escherichia coli*.

[1] R. Mullan Cook-Deegan, "Does NIH Need a DARPA?," *Issues in Science and Technology* XIII:25-28, Winter 1996.

To illustrate the consequences in more concrete but future-oriented terms, the list below suggests some of the activities that would be excluded under a funding model that focuses only on hypothesis-testing research:

- Developing technologies that enable data collection from a myriad of instruments and sensors, including real-time information about biological processes and systems, that permit us to refine and annotate this information and incorporate it into accessible repositories to facilitate scientific study or biomedical procedures;
- Flexible database systems that allow incorporation of multiscale, multimodal information about biological systems by enabling the inclusion (by data federation techniques such as mediation) of information distributed in an unlimited number of other databases, data collections, Web sites and so on;
- Acquisition of "discovery-driven" data (discovery science, as described in Chapter 2) to populate datasets useful for computational analytical methods, or improvements in data acquisition technology and methodology that serve this end;
- Development of new computational approaches to meet challenges of complex biological systems (e.g., improved algorithmic efficiency, development of appropriate signal processing or signal detection statistical approaches to biological data); and
- Data curation efforts to correct and annotate already-acquired data to facilitate greater interoperability.

These considerations suggest that expanding the notion of hypothesis may be useful. That is, the discussion above regarding hypothesis testing refers to *biological* hypotheses. But to the extent that the kinds of research described in the immediately preceding list are in fact part of 21st century biology, nonbiological hypotheses may still lead to important biological discoveries. In particular, a plausible and well-supported *computational* hypothesis may be as important as a biological one and may be instrumental in advancing biological science.

Today, a biological research proposal with excellent computational hypotheses may still be rejected because reviewers fail to see a clearly articulated biological hypothesis. To guard against such situations, funding agencies and organizations would be well served by including in the review process reviewers with the expertise to identify plausible and well-supported computational hypotheses that may aid their biological colleagues in reaching a sound and unbiased conclusion about research proposals at the BioComp interface.

More generally, these considerations involve changing the value proposition for what research dollars should support. At an early point in a research field's development, it certainly makes sense to emphasize very strongly the creation of basic knowledge. But as a field develops and evolves, it is not surprising that a need to consolidate knowledge and make it more usable begins to emerge. In the future, a new balance will have to be struck between the creation of new knowledge and making that knowledge more valuable to the scientific community.

10.3.5.2 Scale of Supported Work

In times of limited resources (and times of limited resources are always with us), unconventional proposals are suspect. Unconventional proposals are even more suspect when they require large amounts of money. No better example can be found than the reactions in many parts of the life sciences research community to the Human Genome Project when it was first proposed—with a projected price tag in the billions of dollars, the fear was palpable that the project would drain away a significant fraction of the resources available for biological research.[79]

Work at the BioComp interface, especially in the direction of integrating state-of-the-art computing and information technology into biological research, may well call for support at levels above those required for more traditional biology research. For example, a research project with senior expertise in both biology and computing may well call for support for co-principal investigators. Just as biological laboratories generally require support for lab technicians, a BioComp project could reasonably call for programmers and/or system administrators. (A related point is that for a number of years in the recent past [i.e., during the dot-com boom years] computer scientists commanded relatively high salaries.)

In addition, some areas of modern life sciences research, such as molecular biology, rely on large grants for the purchase of experimental instruments. The financial needs for instrumentation and laboratory equipment to collect the data necessary for undertake the data-intensive studies of 21st century biology are significant, and are often at a scale that is unaffordable to all but a small number of academic institutions. Although large grants are not unheard of in computer science, the across-the-board dependence of important subfields of biology on experiment means that a larger fraction of biological research is supported through such mechanisms than is true in computer science.

To the extent that proposals for work at the BioComp interface are more costly than traditional proposals and supported by the same agencies that fund those traditional proposals, it will not be surprising to find resistance when they are first proposed.

What is the scale of increased cost that might be associated with greater integration of information technology into the biological research enterprise? If one believes, as does the committee, that information technology will be as transformative to biology as it has been to many modern businesses, IT will affect the way that biological research is undertaken and the discoveries that are made, the infrastructure necessary to allow the work to be done, and the social structures and organizations necessary to support the work appropriately.

Similar transformations have occurred in fields such as high finance, transportation, publishing, manufacturing, and discount retailing. Businesses in these fields tend to invest 5-10 percent of their gross revenues in information technology,[80] and this is with data that is well structured and understood. It is thus not unreasonable to suggest that a full integration of information technology into the biological research enterprise might have a comparable cost. Today, there is federal support for only a very small fraction of that amount.

[79] See, for example, L. Roberts, "Controversial from the Start," *Science* 291(5507):1182-1188, 2001.

[80] See, for example, http://www.bain.com/bainweb/publications/printer_ready.asp?id=17269.

10.3.5.3 The Review Process

Within the U.S. government, there are two styles of review. In the approach relying mainly on peer review (used primarily by NIH and NSF), a proposal is evaluated by a review panel that judges its merits, and the consensus of the review panel is the primary factor that influencing a decision that a proposal does or does not merit funding. When program budgets are limited, as they usually are, the program officer decides on actual awards from the pool of proposals designated as merit-worthy. In the approach relying on program officer judgment (used primarily by DARPA), a proposal is generally reviewed by a group of experts, but decisions about funding are made primarily by the program officer.

The dominant style of review mechanism in agencies that support life sciences research is peer review. Peer review is intended as a method of ensuring the soundness of the science underlying a proposal, and yet it has disadvantages. To quote an NRC report,[81]

> The current peer-review mechanism for extramural investigator-initiated projects has served biomedical science well for many decades and will continue to serve the interests of science and health in the decades to come. NIH is justifiably proud of the peer review mechanism it has put in place and improved over the years, which allows detailed independent consideration of proposal quality and provides accountability for the use of funds. However, any system that focuses on accountability and high success rates in research outcomes may also be open to criticism for discriminating against novel, high-risk proposals that are not backed up with extensive preliminary data and whose outcomes are highly uncertain. The problem is that high-risk proposals, which may have the potential to produce quantum leaps in discovery, do not fare well in a review system that is driven toward conservatism by a desire to maximize results in the face of limited funding resources, large numbers of competing investigators, and considerations of accountability and equity. In addition, conservatism inevitably places a premium on investing in scientists who are known; thus there can be a bias against young investigators.

Almost by definition, peer review panels are also not particularly well suited to considering areas of research outside their foci. That is, peer review panels include the individuals that they do precisely because those individuals are highly regarded as experts within their specialties. Thus, an interdisciplinary proposal that draws on two or more fields is likely to contain components that a review panel in a single field is not able to evaluate as well as those components that do fall into the panel's field.

A number of proposals have been advanced to support a track of scientific review outside the standard peer review panels. For example, the NRC report recommended that NIH establish a special projects program located in the office of the NIH director, funded at a level of $100 million initially to increase over a period of 10 years to $1 billion a year, whose goal would be to foster the conduct of innovative, high-risk research. Most importantly, the proposal calls for a set of program managers to select and manage the projects supported under this program. These program managers would be characterized primarily by an outstanding ability to develop or recognize unusual concepts and approaches to scientific problems. Review panels constituted outside the standard peer review mechanisms and specifically charged with the selection of high-risk, high-payoff projects would provide advice and input to program managers, but decisions would remain with the program managers. Research initially funded through the special projects program that generated useful results would be handed off after 3-5 years for further development and funding through standard NIH peer review mechanisms. Whether this proposal, or a similar one, will be adopted remains to be seen.

Different agencies also have different approaches to the proposals they seek. For example, agencies differ in the amount of detail that they insist potential grantees provide in these proposals. Depending on the nature of the grant or contract sought, one agency might require only a short proposal of a few pages and minimal documentation, whereas another agency might require many more pages, insisting on substantial preliminary results and extensive documentation. An individual familiar with one kind

[81] National Research Council, *Enhancing the Vitality of the National Institutes of Health: Organizational Change to Meet New Challenges*, The National Academies Press, Washington, DC, 2003, p. 93.

of approach may not be able to cope easily with the other, and the overhead involved in coping with an unfamiliar approach can be considerable.

As one illustration, the committee heard from a professor of computer science, accustomed to the NSF approach to proposal writing, who reported that while many biology departments have grant administrators who provide significant assistance in the preparation of proposals to NIH (e.g., telling the PI what is required, drafting budgets, filling out forms, submitting the proposal), his department (of computer science) was unable to provide any such assistance—and indeed lacked anyone at all with expertise in the NIH proposal process. As a result, he found the process of applying for NIH support much more onerous than he had expected.

10.3.6 Issues Related to Intellectual Property and Publication Credit

Issues related to intellectual property (IP) are largely outside the scope of this report. However, it is helpful to flag certain IP issues that are particularly likely to be relevant in advancing the frontiers at the intersection of computer science and biology. Specifically, because information technology enables the sensible use of enormous volumes of biological data, biological findings or results that emerge from such large volumes are likely to involve the data collection work of many parties (e.g., different labs). Indeed, biology as a field recognizes as significant, and even primary, the generation of good experimental data about biological phenomena. By contrast, multiparty collaborations on a comparable scale are unusual in the world of computer science, and datasets themselves are less significant. Thus, computer scientists may well be taken aback by the difficulties in negotiating permissions and credit.

A second issue arises that is related to tensions between open academic research and proprietary commercialization of intellectual advantages. Because of the potential that advances in bioinformatics will have great commercial value, there are incentives to keep some research in bioinformatics proprietary (hence, not easily accessible to the peer community, less amenable to peer review, and less relevant to professional development and advancement). In principle, this is not particularly different at the BioComp interface than in any other research area of commercial value. Nevertheless, the fact that traditions and practices from two different disciplines (disciplines that are at the forefront of economic growth today) are involved rather than just one may exacerbate these tensions.

A third point is the potential tension between making data publicly available and the intellectual property rights of journal publishers. For example, some years ago a part of the neuroscience community sought to build a functional positron emission tomography database. In the course of their efforts, they found that they needed to add substantial prose commentary to the image database to make it useful. Some of the relevant neuroscience journals were reluctant to give permission to use large extracts from publications in the database. To the extent that this example can be generalized, it suggests that efforts to build a far-reaching cyberinfrastructure for biology will have to identify and deal with intellectual property issues as they arise.[82]

[82] In responding to this report in draft, a reviewer argued that by taking collective action, the major research institutions could exert strong leverage on publishers to relax their copyright requirements. Today, many top-rated journals require as a condition of publication the transfer of all copyright rights from the author to the publisher. Given the status of these journals, this reviewer argued that it is a rare researcher who will take his or her paper from a top-rated journal to a secondary journal with less stringent requirements in order to retain copyright. However, the researcher's home institution could adopt a policy in which the institution retained the basic copyright (e.g., under the work-for-hire provisions of current copyright law) but allowed researchers to license their work to publishers but not to transfer the copyright on their own accord. Under such circumstances, goes the argument, journal publishers would be faced with a situation of rejecting work not just from one researcher but from all researchers at institutions with such a policy—a situation that would place far more pressure on journal publishers to relax their requirements and would improve the ability of researchers to share their information through digital resources and databases. The committee makes no judgment about the wisdom of this approach, but believes that the idea is worth mention.

11

Conclusions and Recommendations

11.1 DISCIPLINARY PERSPECTIVES

11.1.1 The Biology-Computing Interface

The committee began this study with two key notions. First, it hoped to identify a field of intellectual inquiry associated with the biology-computing interface that drew equally and bilaterally on computing and biology. Second, it hoped to explicate a symmetry between computing and biology in which the impact of computing on biology was increasingly deep and profound and in which biology would have a comparable effect on computing.

Both of these notions proved unfounded in certain important ways. From the standpoint of applications, technology, and practical utility, the committee saw substantial asymmetry. Computing has had a huge transformational impact on biology and will span virtually all areas of life sciences research, but the impact of biology on computing is likely to be much more targeted (i.e., affecting specific problem domains within computing), and large-scale, biology-based technology changes for computing are in the relatively distant future if they occur at all. At the same time, the committee did find that the epistemological and conceptual frameworks of each field may have in the future some substantial influence on the other. The committee believes that an engineering and computational view (as discussed in Chapter 6) will increasingly be recognized as an important way of looking at biological systems. In a parallel though somewhat more speculative vein, the committee also believes that insight into biological mechanisms may have important impact on how certain problems in computing can be approached (as discussed in Chapter 8).

The reason for the deep and transformational impact of computing on biology is that insight into the vast and heterogeneous datasets of 21st century biology will be possible only through the application of computing to analyze and manage those data. (This is not to deny that many quantitative sciences will contribute to biology, although this report has focused primarily on the computing dimensions.) Views among biologists about where best to deploy computing resources will surely differ, but the main contributions of computing to biology will come from new ideas for solving complex biological problems and new models for testing hypotheses; from delivering cyberinfrastructure for biology research, providing ever more computing power, distributed computing and storage, complex software, fault-tolerant computing, and so forth; and from training fearless scientists who can find the right collaborators for whatever difficulties arise at the frontier. That is, specific computing-to-biology "tech transfer" of intellectual ideas will have some impact, but the greatest impact of computing on biology will come from an overall acceleration of the pace of progress.

To fulfill the promise of 21st century biology, research scientists from both computer and biological science need to work together more extensively, more often, and more closely than ever before. As quantitative methods are increasingly adopted within the biological sciences, it will be possible to answer a new range of scientific questions, not just to accelerate research progress. Uncovering the meaning implicit in the complete sequence of the human genome to deliver on the promises of the project for society is an obvious case.

A revitalized enterprise driven by this newly trained cadre of interdisciplinary scientists and maintained through a balance of individual investigator-initiated and group projects along with continued technology and computational advances, will be able (1) to address fundamental questions in biology such as the relationship of structure to function and the basis for homeostasis; (2) to integrate biological knowledge across the vast scales of time, space, and organizational complexity that characterize biology; (3) to translate basic biology to preventive, predictive, and personalized medicine and to extend biological knowledge to engineering soft materials and other industrial nanobiotechnology contributions; and (4) to uncover how biology can contribute to energy production and environmental restoration. The Committee on the Frontiers at the Interface of Computing and Biology believes that such a vision for 21st century biology is realistic, and that the implementation of its recommendations would ensure decades of exponential progress and a major transformation of our understanding of life.

On the other side of the interface, biological inspiration for new approaches to computing continues to be important, in the sense that biology provides existence proofs that information-processing technology based on biochemistry rather than on silicon electronics is possible. For areas of computing that are generally complex and unwieldy in the associated technologies available so far to address them, or areas lacking in empirical and/or theoretical knowledge, inspiration from whatever source is welcome—and biological inspiration is most likely to be valuable in these areas. (For other areas of computing, whose intellectual terrain is well explored and for which a solid base of empirical and theoretical knowledge is available, biological inspiration is both unnecessary and less interesting, because good and useful solutions are available without any kind of biological connection at all.)

Furthermore, computer scientists tend to be most interested in the general applicability of their work and are often less interested in work that is relevant to only one problem domain. Individuals from this perspective should thus understand the key difference between applications-driven research and applications-specific research. That is, problems in the life sciences can be important drivers of computer science research, and in many cases the knowledge developed in seeking solutions to these problems will be applicable in other domains.

Finally, it is worth noting one possible domain of symmetry between the two fields, although it is a symmetry of ignorance rather than one of knowledge. Both computing and biology provide objects of enormous complexity whose behavior is not well understood—consider the Internet and a cell. It may well turn out that studying each of these objects as systems can yield insights useful in understanding the other—and the same kinds of (yet-to-be-developed) formalism may apply to both—but the jury is still out on this possibility.

11.1.2 Other Emerging Fields at the BioComp Interface

Apart from computing-enabled biology and biologically inspired computing, a number of other new areas of inquiry are also emerging at the BioComp interface, although in addition to biology and computing they draw from chemistry, materials science, bioengineering, and biochemistry. Some of these efforts can be characterized loosely as different flavors of biotechnology, and three of the most important are analytical biotechnology, materials biotechnology, and computational biotechnology.

1. Analytical biotechnology describes the application of biotechnological tools for the creation of chemical measurement systems. Examples include the creation of sensors from DNA-binding proteins for the detection of trace amounts of arsenic and lead in ground waters, and the development of nanoscale DNA cascade switches that can be used to identify single molecular events. Significant challenges for analytical biotechnology arise in proteomics, glycomics, and lipidomics.

2. Materials biotechnology entails the use of biotechnological methods for the fabrication of novel materials with unique optical, electronic, rheological, and selective transport properties. Examples include novel polymers created from genetically engineered polypeptide sequences and the formation of nanowires and circuits from metal nanoparticles attached to a DNA backbone.

3. Computational biotechnology focuses on the potential replacement of silicon devices with nanoscale biomolecular-based computational systems. Examples include the creation of DNA switches from hairpin structures and the programmable self-assembly of DNA tiles for the creation of memory circuits.

A common feature of many of the three new biotechnology application areas is that they all require the production of well-characterized, functional biopolymer nanostructures. The molecular precision and specificity of the enzymatic biochemical pathways employed in biotechnology can often surpass what can be accomplished by other chemical or physical methods—a point that is especially relevant to the problem of nanoscale self-assembly. It is this fine control of nanoscale architecture exhibited in proteins, membranes, and nucleic acids that researchers hope to harness with these applied biotechnologies.

An important enabler of the production of such nanostructures, especially on a large scale, is the availability of increasingly standardized and increasingly automatable fabrication techniques. In some ways, the status of fabrication technologies for these nanostructures is similar to the status of integrated circuit fabrication technology several decades ago, which evolved from a laboratory activity with trial-and-error doping of individual devices to a large-scale automated enterprise driven by design automation software over a period of 20 years beginning in the early 1960s.

Although they draw on biology and computing (along with other disciplines), the tools of these parent disciplines are being applied by researchers in these new biotechnological areas to a different and unrelated set of scientific interests and goals, and these areas often attract scientists with no interests in or ties to traditional biology or computing research. Indeed, these researchers are likely to find intellectual homes in areas such as neuroscience, robotics, and space exploration.

These new areas also have obvious relevance to computing. For example, computational biotechnology is relevant to computing in the same way that lithographic silicon fabrication technologies are today—underpinning these latter technologies are understandings of fundamental physics and well-developed electrical engineering techniques and approaches. Similarly, computational biotechnology will draw on materials science and biochemistry as well as biology as it seeks to create highly regular DNA nanoparticles, mate DNA with submicron electronic structures fabricated in silicon, and create networks of interconnecting nanostructures with unique enzyme communication paths. Analytical and materials biotechnologies are also relevant for enabling MEMS—microelectromechanical systems that interact with the physical world (taking in data through various sensors and affecting the world through various actuators).

11.2 MOVING FORWARD

The committee believes that the most important barriers today impeding the broader integration of computing and information technology into life sciences research are cultural barriers. Twenty-first century biology will not entail a diminution of the central role that traditional empirical or experimental research plays, but it will call for the whole-hearted embrace of a style of biology that integrates reductionist biology with systems biology research. At the same time, computing and physical science practitioners must be wary of underestimating the true complexity of biological systems and, in particular, of inappropriately applying their traditional intellectual paradigms for simplicity to biology.

Over the long run, a change in the culture of academic life sciences research is required to sustain the approaches needed for 21st century biology, due to the increased need for disparate skill sets and collaborative approaches, a change that emphasizes interdisciplinary teams that integrate biology and computing expertise. For this reason, the main focus of this chapter's conclusions and recommendations concern actions that can accelerate the required cultural shift. By contrast, reflecting the committee's view that the impact of biological research is likely to be more modest in scope and scale, the conclusions and recommendations place less emphasis on biology's impact on computing. (In this light, Chapters 4-8 of this report should not be seen as laying out a research agenda for computing-enabled biology or for biology-inspired computing, but rather as suggesting some of the areas in which the frontiers of the interface have been pushed—and that still hold considerable intellectual interest.)

11.2.1 Building a New Community

The most important target of promoting cultural change is people. Thus, it should be a key objective of science policy makers to create a large, multitalented population of individuals who can act as the intellectual translators and mediators along the frontier, a group that will directly foster interdisciplinary research and technology development. True for any discipline or research area involving disparate skill sets, such an approach is especially critical at the interface between the fields of biology and computing because these areas are enjoying the most rapid growth and intellectual progress. Both junior and senior talent must be cultivated, the former to be the basis of a next generation ready to develop and exploit the technology and conduct the science, and the latter to serve in mentorship and leadership roles.

This message is not a new one—indeed, private programs such the Burroughs-Wellcome Foundation Interfaces in Sciences have avidly sought the development of community. Nevertheless, it remains true that despite many studies, reports, and proclamations, universities and federal funding agencies have fallen short of the goal of fully facilitating a range of interdisciplinary science and minimizing the birth pains associated with new hypotheses and directions.[1]

An essential aspect of this community is the ability to build on each other's work. Indeed, the most advanced and sophisticated cyberinfrastructure imaginable will be ineffective if different laboratories and researchers are not motivated or are unwilling to work together or to share data and other information. Formal collaborations between individual laboratories or researchers do exist, of course, but these exist entirely on the basis of individually negotiated arrangements between consenting parties. A different, and complementary, model of working together is one in which individuals researchers contribute to and draw from an entire research community. In spirit, this model is the familiar one of publishing research articles and supporting information (data, software) for others to cite and use as appropriate in their own research—and the dominant ethos of the new community should be one of sharing rather than withholding.

This section provides some core principles on how individuals and institutions might help to support and nurture such work. The core principles described here may come across as "motherhood and apple pie," but it is often the case that such motherhood is not honored as fully as one might think appropriate. The committee does recognize the centrality of providing appropriate incentives for hon-

[1] For example, a report was prepared by the National Institutes of Health (NIH) and the National Science Foundation (NSF) in August 2001 addressing many of the cultural issues described in Chapter 10. This report on training in bioengineering and bioinformatics, *Assessing Bioengineering and Bioinformatics Research Training, Education, and Career Development*, recommended that measures be taken to (1) increase the number of fellowships and institutional training grants at all career levels that include quantitative, computational biology and integrative systems modeling; (2) include funds to support faculty with complementary expertise (e.g., computer scientists to teach biologists); and (3) support the development of curricula. In the intervening 2 years, the importance of continued efforts in these areas has not diminished.

oring these principles. Both institutions and funding agencies have important roles to play in providing incentives for change when these principles are not honored and for continuity when they are. Thus, the core principles for institutions (Section 11.2.3) and for funding agencies (Section 11.4.1) should be seen partly in this light.

11.2.2 Core Principles for Practitioners

The following items are offered as advice to current and prospective researchers at the BioComp interface. These workers include those seeking to retrain themselves to work at the BioComp interface (e.g., a postdoctoral fellow with a computer science background working in a biology laboratory), those facilitating such retraining (e.g., the director of a biology laboratory employing such a postdoc), and those who collaborate as peers with others (e.g., a tenured professor of computer science working with a tenured professor of biology on some interesting problem). Practitioners should:

- *Respect their partners*. Neither the biologist who sees the computer scientist only as a craftsman writing computer programs for data analysis nor the computer scientist who sees the biologist as a provider of dirty and unreliable data shows respect for the other. Scientists with quantitative backgrounds and scientists with biomedical backgrounds must work as peers if their collaborations are to be successful.
- *Have reasonable expectations*. One's intellectual partners in an interdisciplinary endeavor will have differing and often unfamiliar intellectual paradigms. Both vocabulary and epistemology will be different, and a respect for other ways of looking at the world reflects an understanding that paradigms can be different for very sound reasons.
- *Avoid hype*. In the quest for funding and attention, practitioners need to maintain a high degree of questioning to avoid hype, unrealistic expectations, or empty promises.
- *Don't complain*. Complaining to close colleagues about the apparently poor science practiced by other disciplines further reinforces xenophobic arrogance and chauvinism. When the other parties sense such arrogance, the trust needed to achieve scientific collaboration is no longer available.[2]
- *Seek new techniques and intellectual inspiration everywhere*. Both biology and computer science have traditions of applying other disciplines to their problems. For example, Leeuwenhoek's optical microscope led to the discovery of cells, electrical recording devices revealed the voltage-gated channels in neuronal signaling, and knowledge of crystallography uncovered the helical structure and code of DNA. Computer science, originating from a marriage between electrical engineering and mathematics, continues to maintain close intellectual connections to these disciplines.
- *Nurture young talent*. The key to long-term growth of a new field is the ability to sustain and nurture young scientists working in that field. To the extent that attention can be focused on young scientists (e.g., targeting this generation with well-placed, exciting, and novel funding opportunities), problems of competing for funds with senior groups working on classical topics can be reduced.

The committee understands that these principles will have different meaning to researchers at different stages of their careers. For those early in their careers, these recommendations should be taken as a checklist of things to keep in mind as they engage with colleagues and seek support. However, these items are also relevant to senior researchers who serve as role models for their younger colleagues.

[2] G. Wiederhold, "Science in Two Domains," unpublished working paper, Department of Computer Science, Stanford University, March 2002, updated February 2003.

11.2.3 Core Principles for Research Institutions

The following items are offered as advice to institutions that are supporting work at the BioComp interface. These institutions include academic laboratories, research centers, and departments, as well as business or commercial operations with a research component. Collectively, these items are based on the barriers to collaboration and community discussed in Chapter 10. However, no attempt has been made to specifically align recommendations with barriers because in most cases, the correspondence is many-to-many rather than many-to-one or one-to-many.

Relevant institutions should:

- *Attract and retain professionals with quantitative, computational, and engineering skills to work in biological fields.* As a rule, recruitment and retention will require reasonable career tracks that hold the promise of long-term stability and upward mobility. If good individuals are to be attracted to and retained in any enduring interdisciplinary area, they must have career opportunities that offer the potential for growth. For example, these individuals must be assured that their intellectual work at the interface will be fairly evaluated. Such issues are matters of academic survival for many young faculty, and if processes are not put into place explicitly that ensure an appropriately rigorous but still fair evaluation process, promising faculty may well have strong disincentives to pursue research at the interface. A corollary is that traditional departments often see considerable opportunity cost in supporting (and granting tenure to) individuals who do not fit squarely in their centers of gravity (Section 10.3.3); thus, independent support for researchers with interdisciplinary interests, or support that cannot be converted to individuals with traditional interests, helps to remove the threat that departments may see.
- *Support retraining efforts.* Because much of the computing talent required at the BioComp interface will have to come from individuals with substantial prior experience in computing, retraining will be an essential part of efforts to build the talent base. Individuals considering retraining will be more motivated to do so if funding agencies and tenure and promotion committees wishing to support these faculty members recognize that retooling takes some time to be successful and do not penalize them for lowered productivity during such periods.
- *Develop curricula for interdisciplinary teaching of quantitative, computational, and engineering sciences made relevant to the BioComp interface.* Note the desirability of such curricula being made available in multiple formats—online versus in class, 2-week courses versus semester-length courses, and so on—as well as on multiple topics. Over the long run, it is likely that immersion in these curricula will become a natural part of the educational process for all budding biologists, but today, obtaining this background requires some special effort.
- *Facilitate networking.* Especially for newcomers to a line of work, intellectual connection to others plays an important role in their integration into the new community. An institution can promote informal knowledge exchange and the establishment of social relationships on campus through on-site seminars for like-minded individuals. It can also facilitate off-campus connections by providing support for travel to tutorials, workshops, and seminars.
- *Nurture partnerships.* It is desirable for senior scientists from different intellectual backgrounds to work at the interface and for peer relationships between biologists and computer scientists to develop. Partnerships are best undertaken in close proximity with intense interaction, and even small issues such as office arrangements (e.g., whether or not a computer scientist has an office or a desk in the laboratory of a collaborator or partner) can seriously inhibit the development of close partnerships.[3] Many scenarios could promote partnerships, such as sabbatical visits and the establishment of positions at cen-

[3] For example, a computer scientist developing software to aid in the analysis of biological data would be well advised to spend enough time in the laboratory to understand the actual needs of his or her biologist colleagues. Software delivered "over the transom" is unlikely to be used easily, a point suggesting that there is more to software design than the development of an appropriate algorithm.

ters of excellence. Partnerships with industry could ensure sabbaticals in complementary work environments and stimulate knowledge dissemination to commercial applications.

- *Recognize collaborative work.* A corollary of partnerships is that experts from disparate disciplines will collaborate in publication. Institutions thus have a responsibility to provide fair and appropriate evaluation measures for tenure and promotion cases in which the individuals involved have undertaken large amounts of collaborative work. For example, departments may have to be induced to expand their definitions of tenurable work, or universities may have to establish extradepartmental mechanisms for granting and holding tenure outside of traditional departments.
- *Maintain excellence.* Research at the BioComp interface is inherently interdisciplinary, and evaluation of such research faces all of the problems described above. Nevertheless, problem domains that are at the interface of two disciplines can attract not only highly talented individuals who see interesting and important problems but also individuals of lesser talent who are unable to meet the exacting standards of one discipline and are seeking a home where the standards of acceptance are lower. Individuals in the first category are to be sought and cherished—individuals in the second category ought to be shunned.
- *Provide mentors.* Mentors play a strong role in the success of any retraining effort. However, mentoring individuals who have an established track record of success in another field is different. For example, such individuals may be less able to work autonomously and more likely to flail or drift without an activist mentor than someone with a background in the same field. Shared mentorships may make particular sense in these circumstances, as illustrated by the Burroughs-Wellcome requirement that fellowship awardees have a mentor from outside the department of primary appointment.
- *Reward good behavior.* It has been observed that behavior that is rewarded institutionally is behavior that tends to take hold and to be internalized. The institutions with which individual researchers are associated can play important roles in providing such rewards, especially with respect to the principles described in Section 11.2.2.

11.3 THE SPECIAL SIGNIFICANCE OF EDUCATIONAL INNOVATION AT THE BIOCOMP INTERFACE

The pursuit of 21st century biology will require a generation of biologists who can appreciate fundamental statistical approaches, evaluate computational tools and use them appropriately, and know how to choose the best collaborators from the quantitative sciences as a whole. To support the education of this generation, an integrative education, whether formal or informal, will be needed.

Many reports have acknowledged a need for broader training.[4] Increasingly, bioinformatics programs at both the undergraduate and the graduate level do entail study in mathematics, computer science, and the natural sciences.

11.3.1 Content

The committee fully supports these trends and encourages them further, with the strong caveat that an appropriate curriculum to deal with the interface of computing and biology should not simply be the union of course requirements from multiple departments. Courses and other work that deal explicitly with the integrative issues are necessary, and one of the most important skills that such interdisciplinary courses can teach is the ability to communicate among the relevant disciplines. This does not entail simply learning the jargon of each one (though this is, of course, essential), but also interleaving the training in such a way that the student continually sees and explores various parallels between the

[4] See, for example, National Research Council, *Bio2010: Undergraduate Education to Prepare Biomedical Research Scientists*, The National Academies Press, Washington, DC, 2003.

different fields of study. Later, in a collaboration, the ability to identify, explain, and exploit these parallels will be valuable.

Cultural barriers should be discussed and addressed specifically. Where it seems easy to dismiss some math or physics as irrelevant to biology, case studies can be assembled to show successes and contrast these with failures or counterproductive avenues. Where it seems easy to dismiss biology as too detail oriented and reductionistic, similar case studies showing the need to understand minute details of the living machinery are also necessary.

It is broadly agreed that an essential element of 21st century biology is the (re)introduction of quantitative science to the biological science curriculum. The committee recognizes, however, that such reintroduction should not be equated with an abstract, theoretical approach devoid of experimentation or phenomenology, and educational programs for 21st century biology must provide sound footing in quantitative science alongside a clear understanding of the intricacies of biology.

In light of the discussions in Chapter 6 regarding the view of biological organisms as engineered entities, the committee believes that students of 21st century biology would benefit greatly from some study of engineering as well. In this view, the committee emphasizes most strongly its support for the recommendations of the BIO2010 report for exposure to engineering principles (discussed in Chapter 10), at the earliest possible time in the training of life scientists. Just as engineers must construct physical systems to operate in the real world, nature also must operate under these same constraints—physical laws—to "design" successful organisms. Despite this fundamental similarity, biology students rarely learn the important analysis, modeling, and design skills common to engineering curricula nor a suite of topics such as engineering thermodynamics, solid and fluid dynamics, control theory, and so forth, that are key to the engineer's (and nature's) ability to design physical systems.

The particular area of engineering (electrical, mechanical, computer, and so forth) is probably much less relevant than exposure to essential principles of engineering design: the notion of trade-offs in managing competing objectives, control systems theory, feedback, redundancy, signal processing, interface design, abstraction, and the like (Box 11.1). Ready intellectual access to such notions is likely to enable researchers in this area to search for higher-level order in the data forest. Indeed, as biology continues to examine the system-wide functioning of a large number of interacting components, engineering skills may become necessary for successful biological research.

11.3.2 Mechanisms

The committee believes that the availability of individuals with significant computing expertise is an important limiting factor for the rate at which the biological sciences can absorb such expertise.[5] The field, to include both basic and applied life sciences research, is extraordinarily large and dwarfs most other fields outside of engineering itself; thus, influx from other fields is not likely to result in large-scale infusion of computing expertise. Only integrated education of new researchers, along with some retraining of existing researchers, can bring benefits of the computing to a large segment of that world, and previous calls from groups such as Biomedical Information Science and Technology Initiative (BISTI) that a new generation of 21st century researchers must be trained remain compelling, true, and overdue.

Given this perspective, it is appropriate to offer educational opportunities across a broad front. Educational opportunities should span a range in several dimensions, including the following:

- *Time and format.* Monthly lectures or seminars, short-duration workshops (of several weeks), survey courses, undergraduate minors, undergraduate majors, graduate degrees, and postdoctoral

[5] This belief is not based on the existence of a "shortage" or "scarcity" in the sense that economists generally recognize. Rather, it is rooted in the premise that most of biology could benefit intellectually from the integration of significant computing expertise, and the observation that such integration is more the exception than the rule when taken across biology writ large.

Box 11.1
Some Engineering Ideas and Concepts That Biologists May Find Useful

- Control theory (feedback, optimization, game theory)
- Model design
- Signal processing (gain, signal-to-noise, cross-talk)
- Engineering thermodynamics and energy
- Optimal design theory
- Modularity (and protocols)
- Robustness
- Multiscale and large-scale stochastic simulation
- Network theory or graph theory
- Fluid and solid dynamics or mechanics
- "Collective behavior" from physics
- Reverse engineering
- Computational complexity (decidability, P-NP)
- Information theory, source and channel coding
- Dynamical systems: dynamics, bifurcation, chaos
- Statistical physics: phase transitions, critical phenomena

(re)training focusing on the BioComp interface can serve to motivate interest (when they require little investment or time commitment) or to serve a strong professional interest (when the time commitment required is substantial).[6] Short-term opportunities for cross-disciplinary "pollination" workshops that bring together fields from both sides of the interface and provide a vehicle for tutorials and other educational exchanges are particularly useful in that they have a low cost of entry for participants; thus, those who are dabbling can be enticed more easily.

- *Content.* Although genome informatics is perhaps the most obvious topic, computational techniques and approaches will become increasingly relevant to all aspects of biological research—and educational opportunities should target a wide range of subfields in biology.
- *Target audience.* Given the need for more computing expertise in biology, it is appropriate to provide instruction at multiple levels of sophistication in different fields. Some research biologists have substantial informal computing experience but would benefit greatly from more formal exposure; such

[6] In this regard, the model of statistics as a discipline may offer a good example for the way in which bioinformatics might become a discipline of its own. Many universities offer three types of programs in statistics. The first and most formal program is designed for those aspiring to become professional, academic statisticians—that is, those aspiring to become researchers in the field of statistics. This program usually culminates in the Ph.D. degree and establishes an absolutely sound theoretical understanding of the foundations of statistics. The second program is intended for individuals who intend to become professional applied statisticians—that is, those who will work in industry, perhaps in research, and whose primary responsibilities will involve carrying out statistical analyses using established statistical methods. Often, individuals pursuing this degree track will stop with a master's degree and in some cases even with a bachelor's degree. The third program involves a set of courses intended for individuals who will be getting a degree in another field, but who have a need for significant understanding of statistical methods so that those methods might be applied in the individual's home field. In very large universities, sometimes these third-track courses are specialized even further so that we might see courses in business statistics, biological statistics, or even medical statistics. Similarly, it seems reasonably clear that in the field of bioinformatics there will always be a need for researchers, whose primary interest will be in devising new algorithms, new models, and new methods in bioinformatics. There will also be a need for applied bioinformaticians—those whose primary responsibility will be in applying established bioinformatics methods to current projects in biology and biotechnology. It also seems reasonable to suppose that there will be those whose careers in other disciplines will be enhanced by some knowledge of bioinformatics. (These latter two program types may have to provide the biological knowledge to those trained primarily in computation or the computational overview to those trained primarily in biology.)

individuals are in an obviously different situation than those whose only exposure to computing is spreadsheets and word processors.

The development of such educational opportunities generally requires resources, such as release time; assistance in compiling lecture notes, assembling readings, or grading; funding for developing online courses, travel to workshops, and so on. Furthermore, it is desirable to share the outcomes of such development with the academic community (e.g., in the form of online courses, published books, and open commentary about successes and failures). Funding agencies can also provide incentives for such cooperative efforts by giving higher funding priority to research proposals that are put forward in partnerships between or among universities.

11.4 RECOMMENDATIONS FOR RESEARCH FUNDING AGENCIES

The committee believes that it is possible—and feasible—for agencies to support work at the BioComp interface that serves to develop simultaneously (1) fundamental knowledge that enables broad advances in biology; (2) technical innovations that help to improve the quality of life and enhance industrial competitiveness; and (3) the creation and sustenance of a critical mass of talented scientists and engineers intellectually capable and professionally positioned to work creatively at the BioComp interface and to train new generations effectively.

Funding agencies and nongovernmental supporters of research have traditionally been able to influence the course of research through the allocation of resources to particular research fields, and the committee believes that funding at the biology-computing interface is no exception. This support has made important contributions in the past, and the committee urges that such support be continued and expanded.

11.4.1 Core Principles for Funding Agencies

Recognition of the importance in focusing on the BioComp interface amplifies earlier agency-centered studies and reflects its unprecedented richness. Responding to the opportunities, the scientific community, private foundations, and the federal government have taken the first steps in recognizing this enormous intellectual opportunity.

However, no single agency—let alone any individual program, directorate, institute, center, or office—owns the science or the excitement and promise at the interface between computing and biology. Neither can a single agency by itself establish and sustain a process to realize the grand opportunities. In their growing commitment to this frontier science effort, the Defense Advanced Research Projects Agency (DARPA), the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Department of Energy (DOE) each have unique objectives and existing expertise. To exploit the potential fully, the agencies, more than ever before, will have to collaborate and also seek (formal or informal) partnerships with private foundations and industry. Extensive interactions including fully open, joint planning exercises and shared support for technical workshops will be central to true coordination at the agency level.

As is the case for individuals and institutions, a number of core principles provide good desiderata for the funding policies and practices of agencies. Again, these core principles are not particularly new—but remain essential to realizing goals at the BioComp interface. Of course, how these principles are instantiated is key.

To obtain maximum impact, funding agencies and foundations should pay appropriate attention to the following items. Agencies and foundations should:

- *Support awards that can be used for retraining purposes.* While a number of agencies have supported such awards for individuals at early stages of their careers, these programs are fewer in number than in the past. Also, to the best of the committee's knowledge, there are no programs that explicitly target senior faculty for retraining at the BioComp interface, although, as noted in Section 10.2.2.6, NIH does support a retraining program open to scientists of many backgrounds to undertake biomedical research. To the extent that such programs continue to exist, agencies should seek to publicize them beyond their usual core constituencies.

- *Balance quality and excellence against openness to new ideas in the review process.* Intellectual excellence is central. Yet especially in interdisciplinary work, it is also important to invest in work that challenges existing assumptions about how research in the field "should" be conducted—and the problem is that traditional review mechanisms often have a hard time distinguishing between proposals for such work and proposals for work that simply does not meet any reasonable standard of excellence. This point suggests that agencies wishing to support work at the BioComp interface would be wise to find review mechanisms that can draw on individuals who collectively have the relevant interdisciplinary expertise and, as importantly, an appropriate forward-looking view of the field.
- *Encourage team formation*. It is important not to discriminate against team-researched articles in individual performance evaluations and to provide incentives for universities to reward multiple members of cross-disciplinary teams of investigators. Under today's arrangements, work performed by an individual as part of a team often receives substantially less credit than work performed by an individual working alone or with graduate students.
- *Provide research opportunities for investigators at the interface who are not established enough to obtain funding on the strength of their track record alone*. In these instances, balance must be struck between taking a chance on an unproven track record and shutting down nonfruitful lines of inquiry. One approach is to set time limits (a few years) on grants made to such individuals, requiring them to compete on their own against more established investigators after the initial period. (As in other fields, the duration of "a few years" is established by the fact that it is unreasonable to expect significant results in less time, and norms of regular funding set an upper limit for this encouragement of work outside the boundaries.)
- *Use funding leverage to promote institutional change.* That is, agencies can give priority or differential advantages to proposals that are structured in certain ways or that come from institutions that demonstrate commitments to change. For example, priority or preference could be given to proposals that
  —Involve co-principal investigators from different disciplines;
  —Originate in institutions that offer grant awardees tenure-track faculty appointments with minimal teaching responsibilities (as illustrated by the Burroughs-Welcome Career Awards (Section 10.2.2.5.2));
  —Have significant and active educational efforts or programs at the BioComp interface; and
  —Make data available to the larger biological community in standard forms that facilitate reuse and common interpretation.[7] (This action is predicated on the existence of such standards, and agencies should continue to support efforts to develop these common data standards.)
- *Use publication venues to promote institutional change.* Funding agencies could require as a condition of publication that authors deposit the data associated with a given publication into appropriate community databases in accordance with relevant curation standards. They could also insist that published work describing computational models be accompanied by assurances that detailed code inspection of models is possible under an appropriate nondisclosure agreement.

[7] The committee notes without comment that the desire on the part of science agencies to promote wider data sharing and interoperability may conflict with requirements emanating from other parts of the federal government with regard to information management in biomedical research. While science agencies are urging data sharing, other parts of the government can impose restrictions on sharing biomedical data associated with individual human beings in the name of privacy, and these restrictions can have significant impact on the architecture of biomedical information systems. In some cases, these regulatory compliance issues have such impact that biomedical scientists have strong incentives to introduce a *paper* step into their data management processes in order to escape some of the more onerous consequences of these regulations for their information systems.

- *Support cyberinfrastructure for biological research.* Though the National Science Foundation has taken a lead in this area, the issue of supporting cyberinfrastructure for biological research transcends any single agency. Chapter 7 discussed the importance of data repositories and digital libraries in cyberinfrastructure, and it is in these areas that other agencies have important roles to play. Across the board, agencies engaged in supporting biological research will need to support mechanisms for long-term data storage and for continuous curation and annotation of the information resources gathered in publicly supported research for 21st century biology to reach its full potential as a global distributed intellectual enterprise.
- *Recognize quality publicly.* Given the role of peer recognition in the value sets of most scientists (especially in their earlier years), public recognition of innovative work can be a strong motivator. Public recognition can take many forms—though by definition the number of people that can be recognized is necessarily limited. For example, outstanding researchers can be invited to give keynote addresses at important conferences or profiled in reports to Congress or other important public documents.
- *Recognize the costs of providing access to computing and information resources.* Especially at the BioComp interface, collaboration between peers as compared to an investigation conducted by an individual researcher almost always requires larger grants. Researchers need more support for computing and information technology as well as the expertise needed to exploit those capabilities and, in instances that push the computing state of the art, support for high-level expertise as well.
- *Define specific challenge problems that stretch the existing state of the art but are nevertheless amenable to progress in a reasonable time frame.* An agency could pose challenge problems drawn from the problem domains described in Chapter 9. Any number of such challenge problems would be arbitrary, but a selected few goals of broad impact would influence more complete participation by the community and make further funding opportunities by other agencies more likely. Note that when common test sets or other common criteria can be provided or used, clearer metrics for success can be established. A corollary is that agencies should obtain community buy-in with respect to the specifics of such problems. (As one example, the DOE Office of Biological and Environmental Research specified what microbes to tackle for complete genome sequencing through a series of "which bug" workshops to obtain community input on the projects that would be best.)
- *Work with other agencies.* Different agencies bring to the table different types of expertise, and for work at the interface, multiple kinds of expertise are always necessary. Thus, agency partnerships (such as the current collaboration between NIH's National Institute of General Medical Sciences (NIGMS) and NSF's Mathematical and Physical Sciences Directorate) may allow proposals at the interface to be evaluated more fairly and ongoing projects to be overseen more effectively.[8]
- *Provide the funding necessary to capitalize on the intellectual potential of 21st century biology.* Chapters 2-7 of this report have sought to demonstrate the broad impact of computing and information technology on biology. However, a necessary condition to realize this impact is a funding stream that is adequate in magnitude and sustained over long enough periods. As noted in Section 10.3.5.2, a benchmark for comparison is that spending in information-intensive fields such as finance is on the order of 5 to 10 percent of overall budgets. A second necessary condition is the use of a peer review process that is broadly sensitive to the perspectives of researchers in the new field and is willing to take chances on new ideas and approaches. As always, the public sector should focus on growing the seed corn for both people and ideas on which the future depends. Finally, although the committee would gladly endorse an increased flow of funding to the furtherance of a truly integrated 21st century biology, it does understand the realities of a budget-constrained environment.

[8] This partnership between the NIGMS and the NSF seeks to award 20 grants in mathematical biology and anticipates more than $24 million in awards over 5 years. NIGMS supports research and training in the basic biomedical sciences. NSF funds mathematical and other quantitative sciences such as physics, computer science, and engineering. See http://www.nigms.nih.gov/news/releases/biomath.html.

The following sections are addressed to specific funding agencies.

11.4.2 National Institutes of Health

As the largest funder of life sciences research, the NIH has a special responsibility to support and facilitate the building of bridges between biology and other disciplines, especially including computing. The NIH has already taken a number of commendable steps in seeking to collaborate with other agencies, including the formal NIGMS partnership with NSF for mathematical biology mentioned in the previous section and other less formal partnerships with NSF and DOE for structural biology and the National Center for Research Resources (NCRR)-NSF collaborations in instrumentation. As noted in Chapter 10, a National Research Council report in 2003 called for NIH to increase investment in high-risk, high-potential-payoff life sciences research that would be supported outside the usual NIH peer review system.

Such steps, and others like them, are to be encouraged. At the same time, NIH must address obstacles in a number of other areas that impede the building of bridges between biology and computing. One important issue is that cooperation across organizational boundaries within NIH leaves much to be desired. Translational medicine will not arise from funding mechanisms that isolate narrow slices of human biology, and yet the NIH structure is oriented toward specific diseases and body functions.[9] No component of a human works separately, in isolation. Most diseases are not single-gene defects, most proteins act in macromolecular assemblies, organ systems interact by chemical messengers, the immune system and the circulatory system not only work together but impact all organs of the body, and so on. The NIH structure has been successful for many years, but the fact remains that its organizational structure tends to place similar restraints on cross-institute support for collaborative research (Box 11.2).

A consequence of organization of research fields in biology by subfield (e.g., by disease, or by body function) is that efforts that can benefit the entire community may suffer, even though specialization is necessary to achieve depth of knowledge. The true value of the large-scale deployment of cyberinfrastructure—and especially its data components—is that cyberinfrastructure spans disciplines to integrate findings in one subfield with findings in another subfield—to connect information from one subfield to another subfield, perhaps even via a third subfield. In the absence of explicit direction and coordination, cyberinfrastructure in one subfield is likely to be incompatible in important ways with cyberinfrastructure designed and deployed in another. Achieving coordination is likely to require a level of cooperation across agencies that is substantially greater than has historically been true. It will also require a level of planning and agency involvement in the actual design of the cyberinfrastructure that does not typically happen in the funding of research, in which the role of program officers is primarily to ensure a fair assessment of the science by peer reviewers. In supporting cyberinfrastructure, program officers must act as procurement officers on behalf of the overall scientific community, and not just as impartial brokers of an independent discipline-focused review process.

[9] There are 20 institutes within the National Institutes of Health, including the National Cancer Institute (NCI); the National Eye Institute (NEI); the National Heart, Lung, and Blood Institute (NHLBI); the National Human Genome Research Institute (NHGRI); the National Institute on Aging (NIA); the National Institute on Alcohol Abuse and Alcoholism (NIAAA); the National Institute of Allergy and Infectious Diseases (NIAID); the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS); the National Institute of Biomedical Imaging and Bioengineering (NIBIB); the National Institute of Child Health and Human Development (NICHD); the National Institute on Deafness and Other Communication Disorders (NIDCD); the National Institute of Dental and Craniofacial Research (NIDCR); the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK); the National Institute on Drug Abuse (NIDA); the National Institute of Environmental Health Sciences (NIEHS); the National Institute of General Medical Sciences (NIGMS); the National Institute of Mental Health (NIMH); the National Institute of Neurological Disorders and Stroke (NINDS); the National Institute of Nursing Research (NINR); and the National Library of Medicine (NLM).

Box 11.2
The Shapiro Report on the Structure and Organization of the National Institutes of Health

A 2003 National Research Council report on the structure and organization of NIH came to conclusions and made recommendations that are consistent with the view of NIH described in this report. Specifically, the earlier report noted:

> [T]here is a high payoff potential for carefully selected large- and small-scale strategic projects that require the participation of numerous organizations working in partnership. . . . Well-planned, broad-based, trans-NIH programs will be necessary to meet most effectively scientific or public health needs. . . . Furthermore, there is no formal mandate for NIH to identify, plan, and implement such crosscutting strategic initiatives. [Such crosscutting initiatives are necessary because] scientific mechanisms, risk factors, and social and behavioral influences on health and disease cut across traditional disease categories. Many patients have multiple chronic conditions, so a patient-centered approach to health care and health promotion will sometimes require integration and synergy across [Institutes and Centers]. [Such issues] lend themselves to a strategic coordinated trans-NIH response in which multiple institutes could collaborate on a research plan that cuts across administrative structures in terms of planning, funding, and sharing and disseminating results. . . . Proteomics . . . is [an] example [of such an issue]. . . . [C]oncerted trans-NIH work on the assessment of existing and emerging technology platforms and database formats utilizing reference specimens, could help to advance the whole field and guide NIH-supported studies.

The report went on to recommend that initially 5 percent of the NIH budget and eventually 10 percent should be allocated to the support of such trans-NIH initiatives.

SOURCE: National Research Council, *Enhancing the Vitality of the National Institutes of Health: Organizational Change to Meet New Challenges,* The National Academies Press, Washington, DC, 2003, pp. 84-86.

NIH also supports some of the most scientifically sophisticated research environments in the world. As noted in the Botstein-Smarr report,[10] it is in these environments that it makes the most sense to train the leaders of the new generation of biologists with computing expertise. These environments are generally mature enough to support the conduct of interdisciplinary research at the interface, and a widespread geographical diffusion of young scientists with such expertise will help to generate the broad impact sought by NIH.

Perhaps the most important barrier of all is the philosophy that governs much of the current study group approach to proposal review. For historical reasons, the most important and prominent supporters of life sciences research—such as NIH—have focused almost exclusively on hypothesis-testing research—research that investigates well-isolated biological phenomena that can be controlled or manipulated and hypotheses that can be tested in straightforward ways with existing methods. This focus is at the center of reductionist biology and has undeniably been central to much of biology's success in the past several decades.

At the same time, the nearly exclusive focus on hypothesis testing has some important negative consequences. For example, experiments that require breakthrough approaches are unlikely to be directly supported. Just as importantly, advancing technology that could facilitate research is almost always done as a sideline. Thus, investigators must often disguise an attempt to undertake the development of tools or models of great generality by applying them to some (any!) biological system. Subsequent citations of such papers are almost always for the part that explains the new tool or model rather than the phenomenon to which the tool or model was applied.

[10] NIH Working Group on Biomedical Computing, *The Biomedical Information Science and Technology Initiative*, June 1999. Available at http://www.nih.gov/about/director/060399.htm.

This has had a considerable chilling effect in general on what could have been, but the impact is particularly severe for implementation of computational technologies within the biological sciences. That is, in effect as a cultural aspect of modern biological research, technology development to facilitate research is not considered real research and is not considered a legitimate focus of a standard grant. Thus, even computing research that would have a major impact on the advancement of biological science is simply not done.

The committee believes that 21st century biology will be based on a synergistic mix of reductionist and systems biologies. For systems biology researchers, the committee emphasizes that hypothesis-testing research will continue to be central in providing experimental verification of putative discoveries—and indeed, relevant as much to studies of how components interact as to studies of components themselves. Thus, disparaging rhetoric about the inadequacies and failures of reductionist biology and overheated zeal in promoting systems biology should be avoided. For researchers more oriented toward experimental or empirical work, the committee emphasizes that systems biology will be central in formulating novel, interesting, and in some cases, counterintuitive hypotheses to test. The point suggests that agencies that have traditionally supported hypothesis-testing research would do well to cast a wide "discovery" net that supports the development of alternative hypotheses as well as research that supports traditional hypothesis testing.

11.4.3 National Science Foundation

The primary large-scale initiative of NSF relevant to 21st century biology is its cyberinfrastructure effort. Efforts in this area, including major community databases, collaborative research networks, and interdisciplinary modeling efforts, will require grants that are larger than the Directorate for Biological Sciences (BIO) of NSF has traditionally made, as well as greater continuity and stability. In particular, cyberinfrastructure entails personnel costs (e.g., for programmers, systems administrators, and staff scientists with the necessary computing expertise) that are not associated with the usual BIO-supported grant. As for continuity, windows for support must be consistent with the practical considerations to achieve success. Five-year awards and initial review at that point against specific milestones and deliverables to the community are essential, and only at longer intervals should there be open calls for proposals and competitive processes, save in the case of a resource failing to live up to community expectations.[11]

The professional biological community at large has at least two important roles to play with respect to cyberinfrastructure. First, it must articulate its needs and explicate how it can best exploit the resources that cyberinfrastructure will make available. Second, it must develop a consensus on the expectations that cyberinfrastructure facilities must meet if they are to be continued. Society events (e.g., annual meetings) provide a forum for such discussions to take place.

11.4.4 Department of Energy

The DOE's Office of Science supports a number of programs in genomic studies and structural biology (as described in Chapter 10). This office has the capacity to provide sufficient funds and a stable environment, but doing so has been a challenge in its overall institutional setting. The committee believes that the payoffs for DOE missions will be extraordinary from the biology supported by the Office of Science, but success requires that priority be given to stable, long-term programs.

[11] In principle, review provisions could be analogous to the sunset considerations for NSF-supported Science and Technology Centers and Engineering Research Centers.

11.4.5 Defense Advanced Research Projects Agency

Of all the federal agencies, DARPA appears to be the most heavily involved in exploring the potential of biology for computing. Chapter 8 describes a variety of potential influences of biology on computing (the term "applications of biology for computing" would be promising too much), but in truth, the ultimate value of biology for changing computing paradigms in deep and fundamental ways is as yet unproven. Nevertheless, various biological attributes—robustness, adaptation, damage recovery, and so on—are so desirable from a computing point of view that any intellectual inquiry is valuable if it can contribute to artificial humanly purposive systems with these attributes.

In other words, investigations that consider the impact of biology on computing are—in the vernacular—high-risk, high-payoff studies. They are high risk because biology is not prescriptive in its contributions and success is far from ensured. They are high payoff because computers that possess attributes associated with biological systems would be enormously valuable. It is for this reason that they do logically fall into programs supported by DARPA, which has a long tradition of supporting high-risk, high-payoff work as part of its research portfolio. (As noted in Chapter 10, NSF also sponsors a Small Grants Exploratory Research Program that supports high-risk research on a small scale.)

From the committee's perspective, the high-level goals articulated by DARPA and other agencies that support work related to biology's potential contribution to computing seem generally sensible. This is not to say that every proposal supported under the auspices of these agencies' programs would necessarily have garnered the support of the committee—but that would be true of any research portfolio associated with any program.

One important consequence of supporting high-risk research is that it is unlikely to be successful in the short term. Research—particularly of the high-risk variety—is often more "messy" and takes longer to succeed than managers would like. Managers understandably wish to terminate unproductive lines of inquiry, especially when budgets are constrained. However, short-term success cannot be the only metric of the value of research, and when it is, funding managers invite hyperbole and exaggeration on the part of proposal submitters, and unrealistic expectations begin to characterize the field. Those believing the hyperbole (and those contributing to it as well) thus overstate the importance and centrality of the research to the broader goal of improving computing. When unrealistic expectations are not met (and they will not be met, almost by definition), disillusionment sets in, and the field becomes disfavored from both a funding and an intellectual standpoint.

From this perspective, it is easy to see why support for fields can rise rapidly only to drop precipitously a few years later. Wild budget fluctuations and an unpredictable funding environment that changes goals rapidly can damage the long-term prospects of a field to produce useful and substantive knowledge. Funding levels do matter, but programs that provide steady funding in the context of broadly stated but consistent intellectual goals are more likely to yield useful results than those that do not.

Thus, the committee believes that in the area of biologically inspired computing, funding agencies should have realistic expectations, and these expectations should be relatively modest in the near term. Intellectually, their programs should continue to take a broad view of what "biological inspiration" means. Funding levels in these areas ought to be established on a "level-of-effort" basis (i.e., what DARPA believes is a reasonable level of effort to be expended in this area), taking into account the number of researchers doing and likely to do good work in this area and the potential availability of other avenues to improved computing. Also, programmatic continuity should be the rule, with playing rules and priorities remaining more or less constant in the absence of profound scientific discovery or technology advances in the area.

11.5 CONCLUSIONS REGARDING INDUSTRY

Over the past decade, the commercial sector has provided important validation for the proposition that information technology (IT) can have a profound impact on the life sciences. As noted in Chapter

10, there are a host of firms, ranging in size from small start-ups to established multibillion-dollar companies that have significant investments in research efforts and products, that make substantial use of IT in support of medical and pharmaceutical business.

Nevertheless, the committee is aware that some large life science companies (e.g., large pharmaceutical companies) have not found their investments in information technology living up to their expectations. Some such companies have reported investing a great deal of money and time in bioinformatics software and are now looking for and failing to find economic justification for further investment.

The hype of the genome era was as intoxicating to many drug companies, it seems, as the Internet was to mainstream investors, with just as much a comedown. There is a growing realization that the availability of genomic information is not, by itself, sufficient to lead directly to immediately profitable drug breakthroughs, regardless of the IT available to help manage and analyze that information. Indeed, many bottlenecks in drug discovery remain that result from the lack of fundamental biological knowledge about specific expression and pathways. Whereas the initial expectation was that the genome could be mined for likely drug targets, today's approach involves a greater tendency to start with the biology that is known to select likely targets, and then to look to the genome to find genes that interact with those targets.

The committee believes that bioinformatics—and broader uses of information technology—are likely to have a positive effect on drug discovery in the long run, but that those enterprises looking to investments in IT for short-term gain are likely to continue to be disappointed. Commercial advantages to the use of IT will accrue from its integration into the entire process, from gene discovery to clinical trials, benefiting both the entire process and the local situation to which information technology is applied. Also, because of rapidly increasing biological knowledge, the promise of discovering appropriate drug targets in the genome remains, although it is likely to be realized primarily in the long term. Bioinformatics will also enable a more precise genome-based identification of individuals susceptible to a given drug's side effects, possibly providing a basis for excluding them from clinical trials and pharmaceutical applications involving that drug.

11.6 CLOSING THOUGHTS

The impact of computing on biology could fairly be considered a paradigm change as biology enters the 21st century. Twenty-five years ago, biology saw the integration of multiple disciplines from the physical and biological sciences and the application of new approaches to understand the mechanisms by which simple bacteria and viruses function. The impact of the early efforts was so significant that a new discipline, molecular biology, emerged, and many biologists, including those working at the level of tissues or systems and whole organisms, came to adopt the approaches and often even the techniques. Molecular biology has had such success that it is no longer a discipline but simply part of bioscience research itself.

Today, the revolution lies in the application of a new set of interdisciplinary tools: computational approaches will provide the underpinning for the integration of broad disciplines in developing a quantitative systems approach, an integrative or synthetic approach to understanding the interplay of biological complexes as biological research moves up in scale. Bioinformatics provides the glue for systems biology, and computational biology provides new insights into key experimental approaches and how to tackle the challenges of nature. In short, computing and information technology applied to biological problems is likely to play a role for 21st century biology that is in many ways analogous to the role that molecular biology has played in biological research across all fields for the last quarter century—and computing and information technology will likely become embedded with biological research itself.

Appendixes

A

The Secrets of Life: A Mathematician's Introduction to Molecular Biology

NOTE: This appendix is a reprint of Chapter 1 of the National Research Council report *Calculating the Secrets of Life: Contributions of the Mathematical Sciences to Molecular Biology* (National Academy Press, Washington, DC, 1995), copyright 1995 by the National Academy of Sciences.

Chapter 1 The Secrets of Life: A Mathematician's Introduction to Molecular Biology

Eric S. Lander
Whitehead Institute for Biomedical Research
and Massachusetts Institute of Technology

Michael S. Waterman
University of Southern California

Molecular biology has emerged from the synthesis of two complementary approaches to the study of life—biochemistry and genetics—to become one of the most exciting and vibrant scientific fields at the end of the twentieth century. This introductory chapter provides a brief history of the intellectual foundations of modern molecular biology and defines key terms and concepts that recur throughout the subsequent chapters.

The concepts of molecular biology have become household words. DNA, RNA, and enzymes are routinely discussed in newspaper stories, prime-time television shows, and business weeklies. The passage into popular culture is complete only 40 years after the discovery of the structure of deoxyribonucleic acid (DNA) by James Watson and Francis Crick and only 20 years after the first steps toward genetic engineering. With breathtaking speed, these basic scientific discoveries have led to astonishing scientific and practical implications: the fundamental biochemical processes of life have been laid bare. The evolutionary record of life can be read from DNA sequences. Genes for proteins such as human insulin can be inserted into bacteria, which then can inexpensively produce large and pure amounts of the protein. Farm animals and crops can be engineered to produce healthier and more desirable products. Sensitive and reliable diagnostics can be developed for viral diseases such as AIDS, and treatments can be developed for some hereditary diseases, such as cystic fibrosis.

Molecular biology is certain to continue its exciting growth well into the next century. As its frontiers expand, the character of the field is changing. With ever growing databases of DNA and protein sequences and increasingly powerful techniques for investigating structure and function, molecular biology is becoming not just an experimental science, but a *theoretical* science as well. The role of theory in molecular biology is not likely to resemble the role of theory in physics, in which mathematicians can offer grand unifying theories. In biology, key insights emerge less often from first principles than from interpreting the crazy quilt of solutions that evolution has devised. Interpretation depends on having theoretical tools and frameworks. Sometimes, these constructs are nonmathematical. Increasingly, however, the mathematical sciences—mathematics, statistics, and computational science—are playing an important role.

This book emerged from the recognition of the need to cultivate the interface between molecular biology and the mathematical sciences. In the following chapters, various mathematicians working in molecular biology provide glimpses of that interface. The essays are not intended to be comprehensive up-to-date reviews, but rather vignettes that describe just enough to tempt the reader to learn more about fertile areas for research in molecular biology.

This introductory chapter briefly outlines the intellectual foundations of molecular biology, introduces some key terms and concepts that recur throughout the book, and previews the chapters to follow.

BIOCHEMISTRY

Historically, molecular biology grew out of two complementary experimental approaches to studying biological function: biochemistry and genetics (Figure 1.1). Biochemistry involves fractionating (breaking up) the molecules in a living organism, with the goal of purifying and characterizing the chemical components responsible for carrying out a particular function. To do this, a biochemist devises an assay for

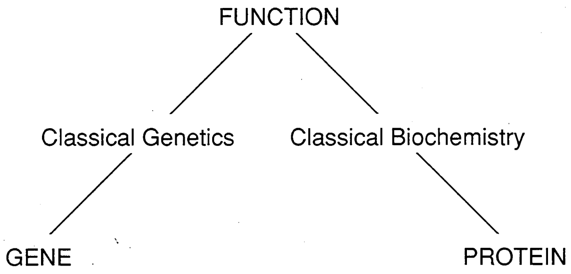


FIGURE 1.1 Genetics and biochemistry began as independent ways to study biological function.

measuring an "activity" and then tries successive fractionation procedures to isolate a pure fraction having the activity. For example, a biochemist might study an organism's ability to metabolize sugar by purifying a component that could break down sugar in a test tube.

In vitro (literally, in glass) assays were accomplished back in the days when biologists were still grappling with the notion of vitalism. Originally, it was thought that life and biochemical reactions did not obey the known laws of chemistry and physics. Such vitalism held sway until about 1900, when it was shown that material from dead yeast cells could ferment sugar into ethanol, proving that important processes of living organisms were "just chemistry." The catalysts promoting these transformations were called enzymes.

Living organisms are composed principally of carbon, hydrogen, oxygen, and nitrogen; they also contain small amounts of other key elements (such as sodium, potassium, magnesium, sulfur, manganese, and selenium). These elements are combined in a vast array of complex macromolecules that can be classified into a number of major types: proteins, nucleic acids, lipids (fats), and carbohydrates (starches and sugars). Of all the macromolecules, the proteins have the most diverse range of functions. The human body makes about 100,000 distinct proteins, including:

- enzymes, which catalyze chemical reactions, such as digestion of food;
- structural molecules, which make up hair, skin, and cell walls;

- transporters of substances, such as hemoglobin, which carries oxygen in blood; and
- transporters of information, such as receptors in the surface of cells and insulin and other hormones.

In short, proteins do the work of the cell. From a structural standpoint, a protein is an ordered linear chain made of building blocks known as amino acids (Figures 1.2 and 1.3). There are 20 distinct amino acids, each with its own chemical properties (including size, charge, polarity, and hydrophobicity, or the tendency to avoid packing with water). Each protein is defined by its unique sequence of amino acids; there are typically 50 to 500 amino acids in a protein.

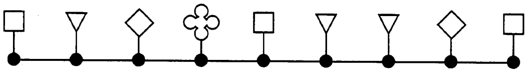


FIGURE 1.2 Proteins are a linear polymer, assembled from 20 building blocks called amino acids that differ in their side chains. The diagram shows a highly stylized view of this linear structure.

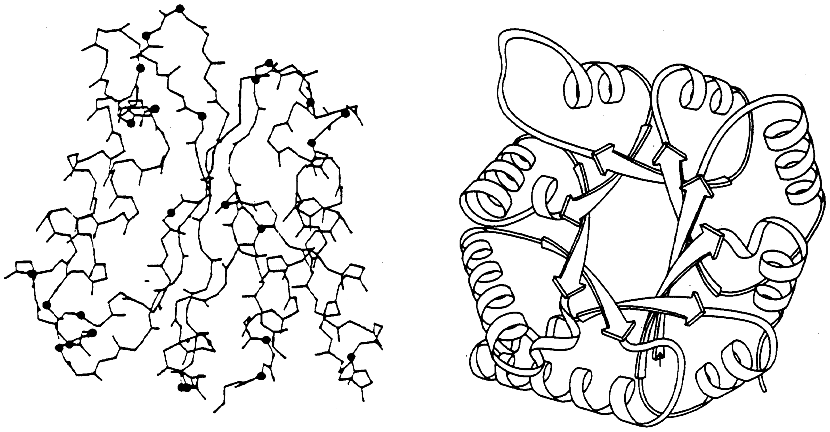


FIGURE 1.3 Examples of different representations of protein structures focusing on (left) chemical bonds and (right) secondary structural features such as helices and sheet-like elements. Reprinted, by permission, from Richardson and Richardson (1989). Copyright © 1989 by the Plenum Publishing Corporation.

The amino acid sequence of a protein causes it to fold into the particular three-dimensional shape having the lowest energy. This gives the protein its specific biochemical properties, that is, its function. Typically, the shape of a protein is quite robust. If the protein is heated, it will be denatured (that is, lose its three-dimensional structure), but it will often reassume that structure (refold) when cooled. Predicting the folded structure of a protein from the amino acid sequence remains an extremely challenging problem in mathematical optimization. The challenge is created by the combinatorial explosion of plausible shapes, each of which represents a local minimum of a complicated nonconvex function of which the global minimum is sought.

CLASSICAL GENETICS

The second major approach to studying biological function has been genetics. Whereas biochemists try to study one single component purified away from the organism, geneticists study mutant organisms that are intact except for a single component. Thus a biochemist might study an organism's ability to metabolize sugar by finding mutants that have lost the ability to grow using sugar as a food source.

Genetics can be traced back to the pioneering experiments of Gregor Mendel in 1865. These key experiments elegantly illustrate the role of theory and abstraction in biology. For his experiments, Mendel started with **pure breeding** strains of peas—that is, ones for which all offspring, generation after generation, consistently show a trait of interest. This choice was key to interpreting the data.

One of the traits that he studied was whether the pea made round or wrinkled seeds. Starting with pure breeding round and wrinkled strains, Mendel made a controlled cross to produce an F_1 generation. (The *i*th generation of the cross is denoted F_i.) Mendel noted that all of the F_1 generation consisted of round peas; the wrinkled trait had completely vanished. However, when Mendel crossed these F_1 peas back to the pure breeding wrinkled parent, the wrinkled trait reappeared: of the second generation, approximately half were round and half were wrinkled. Moreover, when Mendel crossed the F_1 peas to themselves, he found that

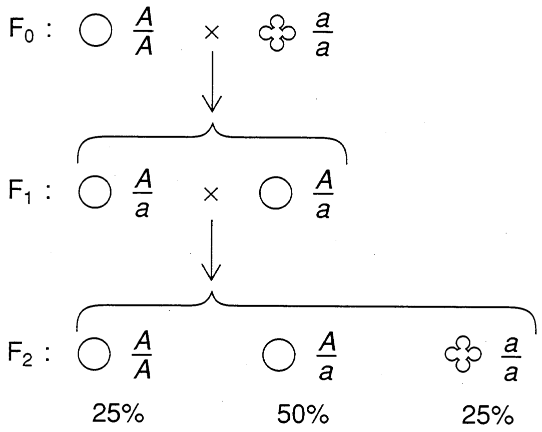


FIGURE 1.4 Mendel's crosses between pure breeding peas with round and wrinkled seeds revealed the telltale binomial ratio 1:2:1 in the second generation that led Mendel to infer the existence of discrete particles of inheritance.

the second generation showed 75 percent round and 25 percent wrinkled (Figure 1.4).

On the basis of these and other experiments, Mendel hypothesized that traits such as roundness are affected by discrete factors—which today we call genes. In particular, Mendel suggested the following:

- Each organism inherits two copies of a gene, one from each parent. Each parent passes on one of the two copies, chosen at random, to each offspring. (These important postulates are called Mendel's First Law of Inheritance.)
- Genes can occur in alternative forms, called **alleles**. For example, the gene affecting seed shape occurs in one form (allele *A*) causing roundness and one form (allele *a*) causing wrinkledness.
- The pure breeding round and wrinkled plants carried two copies of the same allele, *AA* and *aa*, respectively. Individuals carrying two copies of the same gene are called **homozygotes**. The F_1 generation consists of individuals with genotype *Aa*, with the round trait dominant over the wrinkled trait. Such individuals are called **heterozygotes**.

- In the cross of the F_1 generation (*Aa*) to the pure breeding wrinkled strain (*aa*), the offspring were a 1:1 mixture of *Aa:aa* according to which allele was inherited from the F_1 parent. In the cross between two F_1 parents (*Aa*), the offspring were a 1:2:1 mixture of *AA:Aa:aa* according to the binomial selection of alleles from the two parents.

It is striking to realize that the existence of genes was deduced in this abstract mathematical way. Probability and statistics were an intrinsic part of early genetics, and they have remained so. Of course, Mendel did not have formal statistical analysis at his disposal, but he managed to grasp the key concepts intuitively. Incidentally, the famous geneticist and statistician R.A. Fisher analyzed Mendel's data many years later and concluded that they fit statistical expectation a bit too well. Mendel probably discarded some outliers as likely experimental errors.

It was almost 35 years before biologists had an inkling of where these hypothetical genes resided in the cell (in the chromosomes) and almost 100 years before they understood their biochemical nature.

MOLECULAR BIOLOGY

As suggested in Figure 1.1, the biochemical and the genetic approaches were virtually disjoint: the biochemist primarily studied proteins, whereas the geneticist primarily studied genes. Much like the great unifications in mathematics, molecular biology emerged from the recognition that the two apparently unrelated fields were, in fact, complementary perspectives on the same subject.

The first clues emerged from the study of mutant microorganisms in which gene defects rendered them unable to synthesize certain key macromolecules. Biochemical study of these genetic mutants showed that each lacked a specific enzyme. From these experiments the hypothesis became clear that genes somehow must "encode" enzymes. This (Nobel-Prize-winning) notion was dubbed the "one gene-one enzyme" hypothesis, although today it has been modified to "one gene-one protein." Of course, the mystery remained: How do genes encode proteins?

The answer depended on finding the biochemical nature of the gene itself, thereby uniting the fields. To purify the gene as a biochemical entity, one needed a test tube assay for heredity—something that might seem impossible. Fortunately, scientific serendipity provided a solution. In a famous series of bacteriological studies, Griffith showed 50 years ago that certain properties (such as pathogenicity) could be transferred from dead bacteria to live bacteria. Avery et al. (1944) were able to successively fractionate the dead bacteria so as to purify the elusive "transforming principle," the material that could confer new heredity on bacteria. The surprising conclusion was that the gene appeared to be made of DNA.

The notion of DNA as the material of heredity came as a surprise to most biochemists. DNA was known to be a linear polymer of four building blocks called nucleotides (referred to as adenine, thymine, cytosine, and guanine, and abbreviated as A, T, C, and G) joined by a sugar-phosphate backbone. However, most knowledgeable scientists reckoned that the polymer was a boring, repetitive structural molecule that functioned as some sort of scaffold for more important components. In the days before computers, it was not apparent how a linear polymer might encode information. If DNA contained the genes, the structure of DNA became a key issue.

In their legendary work in 1953, Watson and Crick correctly inferred the structure of most DNA and, in so doing, explained the main secret of heredity. While some viruses have single-stranded DNA, the DNA of humans and of most other forms of life consists of two antiparallel chains (strands) in the form of a double helix in which the bases (nucleotides) pair up to form **base pairs** in a certain way (Figure 1.5) so that the sequence of one chain completely specifies the sequence of the other: an A on one chain always corresponds to a T on the other, and a G to a C. The sequences are complementary. The fact that the information is redundant explains the basis for the replication of living organisms: the two strands of the double helix unwind, and each serves as a template for the synthesis of a complete double helix that is passed on to a daughter cell. This process of replication is carried out by enzymes called **DNA polymerases**. Mutations are changes in the nucleotide sequence in DNA. Mutations can be induced by external

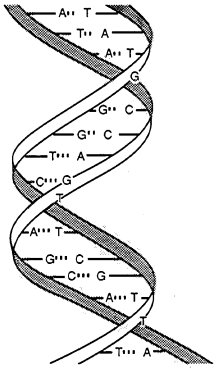
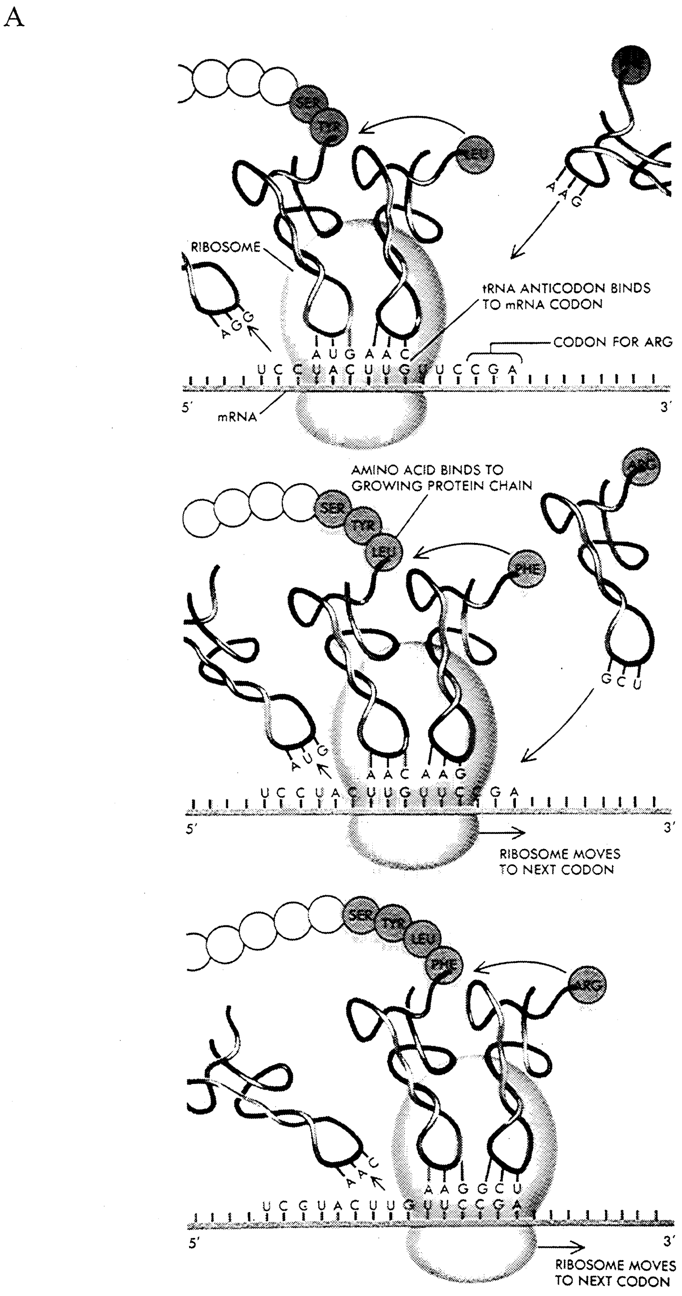


FIGURE 1.5 The DNA double helix consists of anti-parallel helical strands, with complementary bases (G-C and A-T).

forces such as sunlight and chemical agents or can occur as random copying errors during replication.

There remained the question of how the 4-letter alphabet of DNA could "encode" the instructions for the 20-letter alphabet of protein sequences. Biochemical studies over the next decade showed that genes correspond to specific stretches of DNA along a chromosome (much like individual files on a hard disk). These stretches of DNA can be expressed at particular times or under particular circumstances. Typically, gene expression begins with **transcription** of the DNA sequence into a messenger molecule made of ribonucleic acid (RNA) (Figure 1.6A). This transcription process is carried out by enzymes called **RNA polymerases**. RNA is structurally similar to DNA and consists of four building blocks, the nucleotides denoted A, U, C, and G, with U (uracil) playing the role of T. The **messenger RNA** (mRNA) is copied from the DNA of a gene according to the usual base pairing rules (a U in RNA corresponds to an A in DNA, an A corresponds to a T, a G to a C, and a C to a G). The messenger RNA copied from a gene is single-stranded and is just an unstable intermediate used for transmitting information from the cell nucleus (where the DNA resides) to the cytoplasm (where protein synthesis occurs). The mRNA is then translated into a protein by a remarkable molecular machine called the **ribosome.**


A
SER
TYR
LEU
A A G
RIBOSOME
A G G
tRNA ANTICODON BINDS TO mRNA CODON
A U G A A C
CODON FOR ARG
U C C U A C U U G U U C C G A
5′
mRNA
3′
AMINO ACID BINDS TO GROWING PROTEIN CHAIN
SER
TYR
LEU
PHE
ARG
G C U
A U G
A A C A A G
U C C U A C U U G U U C C G A
5′
3′
RIBOSOME MOVES TO NEXT CODON
SER
TYR
LEU
PHE
ARG
A A C
A A G G C U
U C G U A C U U G U U C C G A
5′
3′
RIBOSOME MOVES TO NEXT CODON

B

FIRST POSITION (5′ END)	SECOND POSITION				THIRD POSITION (3′ END)
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Note: Given the position of the bases in a codon, it is possible to find the corresponding amino acid. For example, the codon (5′) AUG (3′) on mRNA specifies methionine, whereas CAU specifies histidine. UAA, UAG, and UGA are termination signals. AUG is part of the initiation signal, and it codes for internal methionines as well.

FIGURE 1.6 After messenger RNA is transcribed from the DNA sequence of a gene, it is translated into protein by a remarkable molecular device called the ribosome. (A) Ribosomes read the RNA bases and write a corresponding amino acid sequence. The correct amino acid is brought into juxtaposition with the correct nucleotide triplet through the mediation of an adapter molecule known as transfer RNA. (B) The table showing the correspondence between triplets of bases and amino acids is called the genetic code. Reprinted from *Recombinant DNA: A Short Course* by Watson, Tooze, and Kurtz (1994). Copyright © 1994 James D. Watson, John Tooze, and David T. Kurtz. Used with permission of W.H. Freeman and Company.

The ribosome "reads" the linear sequence of the mRNA and "writes" (i.e., creates) a corresponding linear sequence of amino acids of the encoded protein. Translation is carried out according to a three-letter code: a group of three letters is a **codon** that specifies a particular amino acid according to a look-up table called the genetic code (Figure 1.6B). There are 4^3 different codons. The codons are read in contiguous, nonoverlapping fashion from a defined starting point, called the translational start site. Finally, the newly synthesized amino acid chain spontaneously folds into its three-dimensional structure. (For a recent discussion of protein folding, see Sali et al., 1994.)

The details of the genetic code were solved by elegant biochemical tricks, which were necessary because chemists had only the ability to synthesize random collections of RNA having defined proportions of different bases. With some combinatorial reasoning, this proved to be sufficient. For example, if the ribosome is given an mRNA with the sequence UUUUU. . ., then it makes a protein chain consisting of only the amino acid phenylalanine (Phe). Thus UUU must encode phenylalanine. By examining more complex mixtures, researchers soon worked out the entire genetic code.

Molecular biology provides the third leg of the triangle, relating genetics and biochemistry (Figure 1.7).

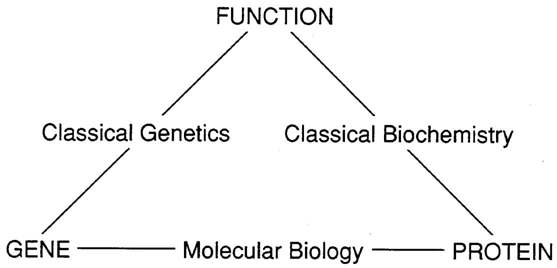


FIGURE 1.7 Molecular biology connected the disciplines of genetics and biochemistry by showing how genes encoded proteins.

THE RECOMBINANT DNA REVOLUTION

By 1965, molecular biology had laid bare the basic secrets of life. Without the ability to manipulate genes, however, the understanding was more theoretical than operational. In the 1970s, this situation was transformed by the recombinant DNA revolution.

Biochemists discovered a variety of enzymes made by bacteria that allowed one to manipulate DNA at will. Bacteria made **restriction enzymes**, which cut DNA at specific sequences and served as a defense against invading viruses, and **ligases**, which join DNA fragments. With these and other tools (which are now all readily available from commercial suppliers), it became possible to cut and paste DNA fragments at will and to introduce them into living cells (Figure 1.8). Such cloning experiments allow scientists to reproduce unlimited quantities of specific DNA molecules and have led to detailed understanding of individual genes. Moreover, producing recombinant DNA molecules that contain bacterial DNA instructions for making a particular human protein (such as insulin) gave birth to the biotechnology industry.

A key development was the invention of **DNA sequencing**, the process of determining the precise nucleotide sequence of a cloned DNA molecule. With DNA sequencing, it became possible to read the sequence of any gene in stretches of 300 to 500 nucleotides at a time. DNA sequencing has revealed striking similarities among living creatures as diverse as humans and yeast, with far-reaching consequences for our understanding of molecular structure and evolution. DNA sequencing has also led to an information explosion in biology, with public databases still expanding at a rapid exponential rate. In early 1993, there were over 100 million bases of DNA in the public databases. For reference, the entire genome of the intestinal bacteria *Escherichia coli* (*E. coli*) consists of about 4.6 million bases, and the human genome sequence has roughly 3 billion bases.

In recent years a powerful new technique called the **polymerase chain reaction** (PCR) has been added to the molecular biologist's tool kit (Figure 1.9). PCR allows one to directly amplify a specific DNA sequence without resort to cloning. To perform PCR, one uses short DNA molecules called **primers** (typically about 20 bases long) that are complementary to the sequences flanking the region of interest. Each

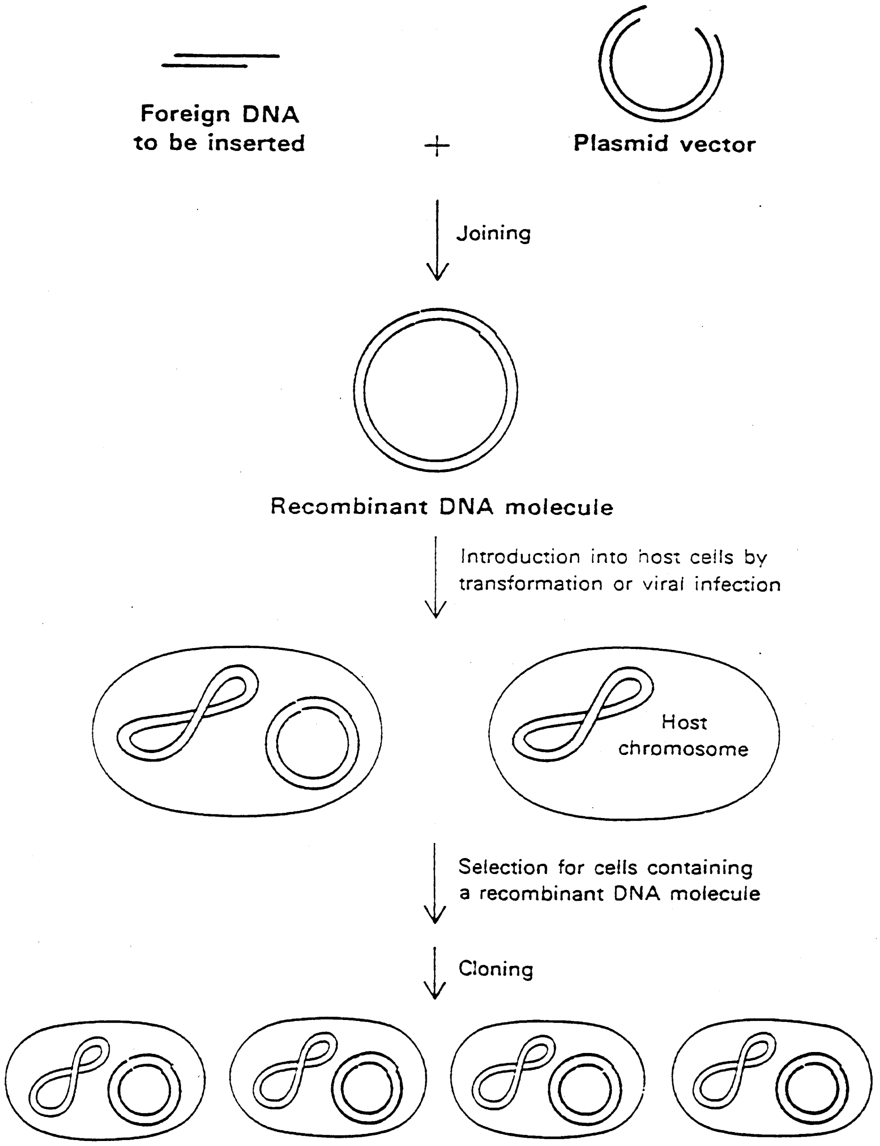

Foreign DNA
to be inserted
+
Plasmid vector
Joining
Recombinant DNA molecule
Introduction into host cells by
transformation or viral infection
Host
chromosome
Selection for cells containing
a recombinant DNA molecule
Cloning

FIGURE 1.8 By cloning a foreign DNA molecule in a plasmid vector, it is possible to propagate the DNA in a bacterial or other host cell.

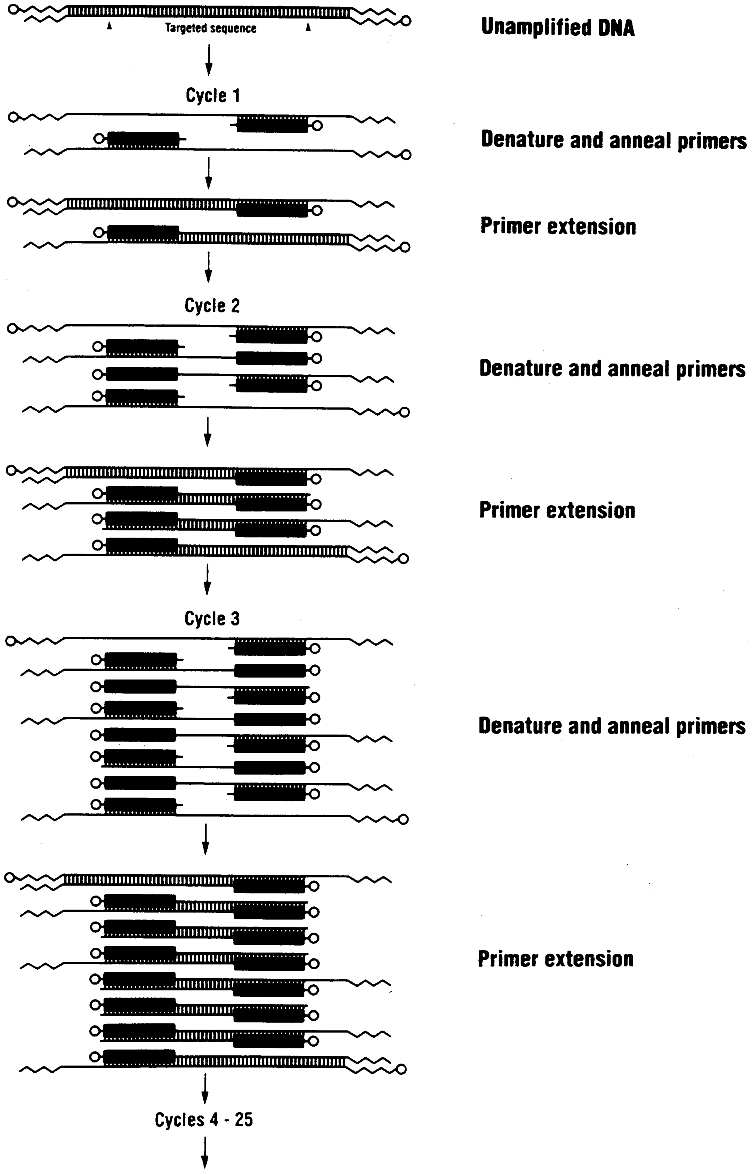


FIGURE 1.9 The polymerase chain reaction (PCR) allows exponential amplification of DNA. The method involves successive rounds of copying (using the enzyme DNA polymerase) between two synthetic primers corresponding to nearby DNA sequences. Each round doubles the number of copies. Courtesy of the Perkin-Elmer Corporation. Reprinted from the National Research Council (1992).

primer is allowed to pair with a base in the complementary region and is then extended to contain the full sequence from the region by using the enzyme DNA polymerase. In this fashion a single copy of the region gives rise to two copies. By iterating this step *n* times, one might make 2^n copies of the region. In practice, one can start with a small drop of blood or saliva and obtain a millionfold amplification of a region. Not surprisingly, PCR has found myriad applications, especially in genetic diagnostics.

MOLECULAR GENETICS IN THE 1990s

With the tools of recombinant DNA, the triangle of knowledge (see Figure 1.7) has been transformed, to use a mathematical metaphor, into a commutative diagram (Figure 1.10). It is possible to traverse the diagram in any direction—for example, to find the genes and proteins underlying a biological function or to find the protein and function associated with a given gene.

A good illustration of the power of the techniques is provided by recent studies of the inherited disease cystic fibrosis (CF). CF is a recessive disease, the genetics of which is formally identical to wrinkledness in peas as studied by Mendel: if two non-affected carriers of the recessive CF gene *a* (that is, heterozygotes with genotype *Aa*) marry, one fourth of their offspring will be affected (that is, will have genotype *aa*). The frequency of the disease-causing allele is about 1/42 in the Caucasian population, and so about 1/21 of all Caucasians are carriers. Since a marriage between two carriers produces 1/4 affected children, the disease frequency in the population is about $1/2000\,(\approx 1/4\times 1/21\times 1/21)$.

Although CF was recognized relatively early in the century, the molecular basis for the disease remained a mystery until 1989. The first breakthrough was the **genetic mapping** of CF to human chromosome 7 in 1985 (Figure 1.11). Genetic mapping involved showing that the inheritance pattern of the disease in families is closely correlated with the inheritance pattern of a particular **DNA polymorphism** (that is, a common spelling variation in the DNA), in this case on chromosome 7.

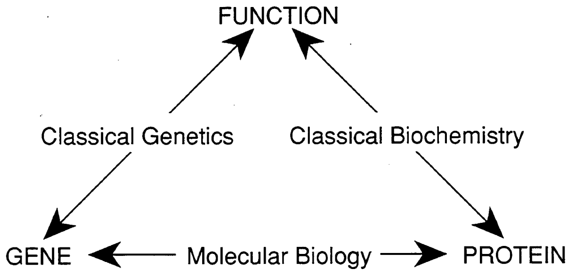


FIGURE 1.10 Recombinant DNA provided the ability to move freely in any direction among gene, protein, and function, thereby converting the triangle of Figure 1.7 into a commutative diagram.

The correlation does not imply that the polymorphism causes the disease, but rather that the polymorphism must be located near the site of the disease gene. Of course, "near" is a relative term. In this case, "near" meant that the CF gene must be within 1 million to 2 million bases of DNA along the chromosome. The next step was the physical mapping and the DNA sequencing of the CF gene itself, which took four more years to accomplish. This involved starting from the nearby polymorphism and sequentially isolating adjacent fragments in a tedious process called **chromosomal walking** until the disease gene was reached. Once the disease gene was found, its complete DNA sequence was determined. (A description of how one knows that one has found the disease gene is beyond the scope of this introduction.)

From the DNA sequence, it became clear that the CF gene encoded a protein of 1,480 amino acids and that the most common misspelling in the population (accounting for about 70 percent of all CF alleles) was a three-letter deletion that removed a single codon specifying an amino acid, a phenylalanine at position 508 of the protein. On the basis of this finding, it became possible to perform DNA diagnostics on individuals to see if they carried the common CF mutation.

Even more intriguingly, the sequence gave immediate clues to the structure and function of the gene product. When the protein sequence was compared with the public databases of previous sequences, it was found to show strong similarities to a class of proteins that were membrane-bound transporters—molecules that reside in the cell membrane, bind adenosinetriphosphate (ATP), and transport substances

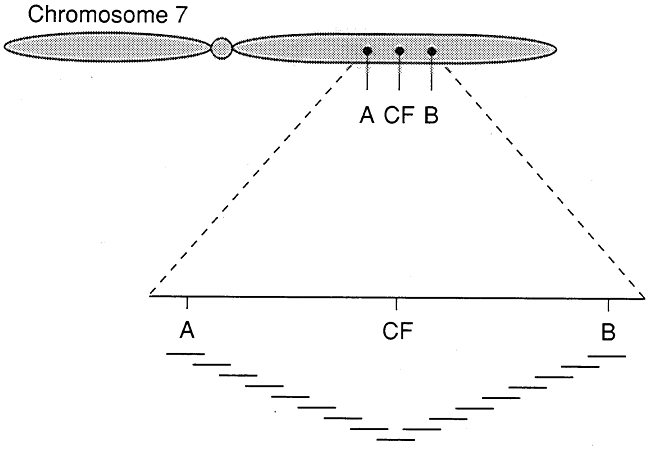


FIGURE 1.11 Chromosomal walking from flanking genetic markers to the gene responsible for cystic fibrosis. The distance covered totaled more than 1 million DNA bases.

into and out of the cell (Figure 1.12A). By analogy, it was even possible to infer a likely three-dimensional shape for the CF protein (Figure 1.12B). In this way, computer-based sequence analysis shed substantial light on the structure and function of this important disease gene.

With the recent advent of gene therapy—the ability to use a virus as a shuttle to deliver a working copy of a gene into cells carrying a defective version—clinical trials have been started to try to cure the disease in the lung cells of CF patients. The path from the initial discovery of the gene to potential therapies has been stunningly short in this case.

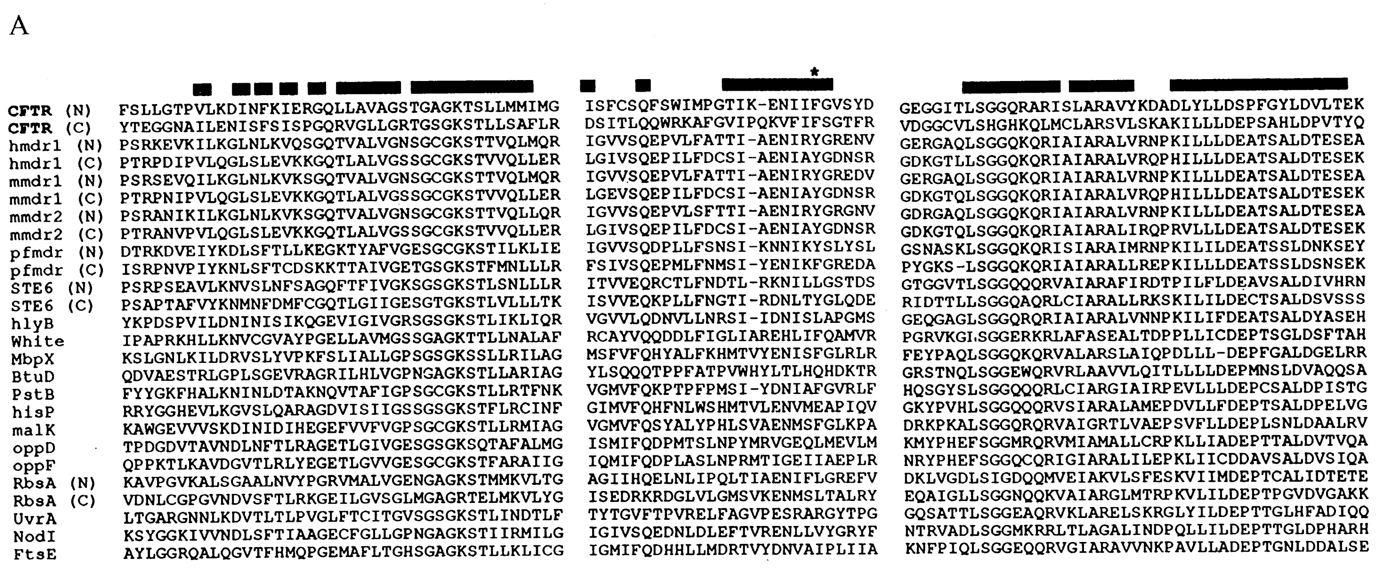
THE HUMAN GENOME PROJECT

With the identification of the CF gene as well as a number of other successes, it has become clear that molecular genetics has developed a powerful general paradigm that can be applied to many inherited diseases and will have a profound impact on our understanding of human health. Unfortunately, the paradigm involves many tedious laboratory steps: genetic mapping (finding a polymorphism closely linked to the disease gene), physical mapping (isolating the consecutive fragments of DNA along the chromosome), and DNA sequencing (typically performed in pieces of only 300 to 500 letters at a time). It would be inefficient to repeat these steps for each of the more than 4,000 genetic traits and diseases already known. To accelerate progress, molecular geneticists have seen the value of building infrastructure—a common set of maps, tools, and information—that can be applied to all genetic problems. This recognition led to the creation of the Human Genome Project (National Research Council, 1988), an international effort to analyze the structure of the human genome (as well as the genomes of certain key experimental model systems, such as *E. coli*, yeast, nematodes, fruit flies, and mice).

Because most molecular biological methods are applicable only to small fragments of DNA, it is not practical to sequence the human genome by simply starting at one end and proceeding sequentially. Moreover, because the current cost of sequencing is about $1 per base, it would be expensive to sequence the 3×10^9 bases of the human chromosomes by conventional methods. Instead, it is more sensible to construct maps of increasing resolution and to develop more efficient sequencing technology. The current goals of the Human Genome Project include development of the following tools:

- *Genetic maps*. The goal is to produce a genetic map showing the location of 5,000 polymorphisms that can be used to trace inheritance of diseases in families. As of this writing, the goal is nearly complete.
- *Physical maps*. The goal is to produce a collection of overlapping pieces of DNA that cover all the human chromosomes. This goal is not completed yet but should be by 1996.
- *DNA sequence*. The ultimate goal is to sequence the entire genome, but the intermediate steps include sequencing particular regions, generating more efficient and automated technology, and developing better analytical methods for handling DNA information.

With the vast quantities of information being generated, the Human Genome Project is one of the driving forces behind the expanding role



A

```
                                                                                     *
CFTR (N)   FSLLGTPVLKDINFKIERGQLLAVAGSTGAGKTSLLMMIMG   ISFCSQFSWIMPGTIK-ENIIFGVSYD   GEGGITLSGGQRARISLARAVYKDADLYLLDSPFGYLDVLTEK
CFTR (C)   YTEGGNAILENISFSISPGQRVGLLGRTGSGKSTLLSAFLR   DSITLQQWRKAFGVIPQKVFIFSGTFR   VDGGCVLSHGHKQLMCLARSVLSKAKILLLDEPSAHLDPVTYQ
hmdr1 (N)  PSRKEVKILKGLNLKVQSGQTVALVGNSGCGKSTTVQLMQR   IGVVSQEPVLFATTI-AENIRYGRENV   GERGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
hmdr1 (C)  PTRPDIPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER   LGIVSQEPILFDCSI-AENIAYGDNSR   GDKGTLLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEK
mmdr1 (N)  PSRSEVQILKGLNLKVKSGQTVALVGNSGCGKSTTVQLMQR   IGVVSQEPVLFATTI-AENIRYGREDV   GERGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
mmdr1 (C)  PTRPNIPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER   LGEVSQEPILFDCSI-AENIAYGDNSR   GDKGTQLSGGQKQRIAIARALVRQPHILLLDEATSALDTESEK
mmdr2 (N)  PSRANIKILKGLNLKVKSGQTVALVGNSGCGKSTTVQLLQR   IGVVSQEPVLSFTTI-AENIRYGRGNV   GDRGAQLSGGQKQRIAIARALVRNPKILLLDEATSALDTESEA
mmdr2 (C)  PTRANVPVLQGLSLEVKKGQTLALVGSSGCGKSTVVQLLER   LGIVSQEPILFDCSI-AENIAYGDNSR   GDKGTQLSGGQKQRIAIARALIRQPRVLLLDEATSALDTESEK
pfmdr (N)  DTRKDVEIYKDLSFTLLKEGKTYAFVGESGCGKSTILKLIE   IGVVSQDPLLFSNSI-KNNIKYSLYSL   GSNASKLSGGQKQRISIARAIMRNPKILILDEATSSLDNKSEY
pfmdr (C)  ISRPNVPIYKNLSFTCDSKKTTAIVGETGSGKSTFMNLLLR   FSIVSQEPMLFNMSI-YENIKFGREDA   PYGKS-LSGGQKQRIAIARALLREPKILLLDEATSSLDSNSEK
STE6 (N)   PSRPSEAVLKNVSLNFSAGQFTFIVGKSGSGKSTLSNLLLR   ITVVEQRCTLFNDTL-RKNILLGSTDS   GTGGVTLSGGQQQRVAIARAFIRDTPILFLDEAVSALDIVHRN
STE6 (C)   PSAPTAFVYKNMNFDMFCGQTLGIIGESGTGKSTLVLLLTK   ISVVEQKPLLFNGTI-RDNLTYGLQDE   RIDTTLLSGGQAQRLCIARALLRKSKILILDECTSALDSVSSS
hlyB       YKPDSPVILDNINISIKQGEVIGIVGRSGSGKSTLIKLIQR   VGVVLQDNVLLNRSI-IDNISLAPGMS   GEQGAGLSGGQRQRIAIARALVNNPKILIFDEATSALDYASEH
White      IPAPRKHLLKNVCGVAYPGELLAVMGSSGAGKTTLLNALAF   RCAYVQQDDLFIGLIAREHLIFQAMVR   PGRVKGI.SGGERKRLAFASEALTDPPLLICDEPTSGLDSFTAH
MbpX       KSLGNLKILDRVSLYVPKFSLIALLGPSGSGKSSLLRILAG   MSFVFQHYALFKHMTVYENISFGLRLR   FEYPAQLSGGQKQRVALARSLAIQPDLLL-DEPFGALDGELRR
BtuD       QDVAESTRLGPLSGEVRAGRILHLVGPNGAGKSTLLARIAG   YLSQQQTPPFATPVWHYLTLHQHDKTR   GRSTNQLSGGEWQRVRLAAVVLQITLLLLDEPMNSLDVAQQSA
PstB       FYYGKFHALKNINLDTAKNQVTAFIGPSGCGKSTLLRTFNK   VGMVFQKPTPFPMSI-YDNIAFGVRLF   HQSGYSLSGGQQQRLCIARGIAIRPEVLLLDEPCSALDPISTG
hisP       RRYGGHEVLKGVSLQARAGDVISIIGSSGSGKSTFLRCINF   GIMVFQHFNLWSHMTVLENVMEAPIQV   GKYPVHLSGGQQQRVSIARALAMEPDVLLFDEPTSALDPELVG
malK       KAWGEVVVSKDINIDIHEGEFVVFVGPSGCGKSTLLRMIAG   VGMVFQSYALYPHLSVAENMSFGLKPA   DRKPKALSGGQRQRVAIGRTLVAEPSVFLLDEPLSNLDAALRV
oppD       TPDGDVTAVNDLNFTLRAGETLGIVGESGSGKSQTAFALMG   ISMIFQDPMTSLNPYMRVGEQLMEVLM   KMYPHEFSGGMRQRVMIAMALLCRPKLLIADEPTTALDVTVQA
oppF       QPPKTLKAVDGVTLRLYEGETLGVVGESGCGKSTFARAIIG   IQMIFQDPLASLNPRMTIGEIIAEPLR   NRYPHEFSGGQCQRIGIARALILEPKLIICDDAVSALDVSIQA
RbsA (N)   KAVPGVKALSGAALNVYPGRVMALVGENGAGKSTMMKVLTG   AGIIHQELNLIPQLTIAENIFLGREFV   DKLVGDLSIGDQQMVEIAKVLSFESKVIIMDEPTCALIDTETE
RbsA (C)   VDNLCGPGVNDVSFTLRKGEILGVSGLMGAGRTELMKVLYG   ISEDRKRDGLVLGMSVKENMSLTALRY   EQAIGLLSGGNQQKVAIARGLMTRPKVLILDEPTPGVDVGAKK
UvrA       LTGARGNNLKDVTLTLPVGLFTCITGVSGSGKSTLINDTLF   TYTGVFTPVRELFAGVPESRARGYTPG   GQSATTLSGGEAQRVKLARELSKRGLYILDEPTTGLHFADIQQ
NodI       KSYGGKIVVNDLSFTIAAGECFGLLGPNGAGKSTIIRMILG   IGIVSQEDNLDLEFTVRENLLVYGRYF   NTRVADLSGGMKRRLTLAGALINDPQLLILDEPTTGLDPHARH
FtsE       AYLGGRQALQGVTFHMQPGEMAFLTGHSGAGKSTLLKLICG   IGMIFQDHHLLMDRTVYDNVAIPLIIA   KNFPIQLSGGEQQRVGIARAVVNKPAVLLADEPTGNLDDALSE
```

B

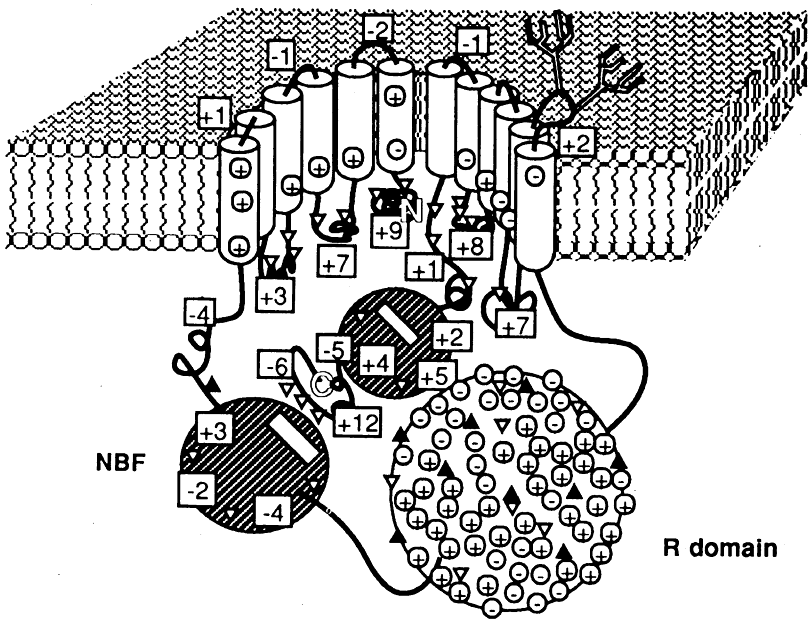


FIGURE 1.12 (A) The protein sequence of the cystic fibrosis gene showed striking similarities to a variety of proteins known to transport molecules across cell membranes. (B) Based on these similarities, it was possible to construct a basic molecular model of the architecture of the CF protein. Reprinted, by permission, from Riordan et al. (1989). Copyright © 1989 by the American Association for the Advancement of Science.

for mathematics, statistics, and computer science in modern molecular biology.

COMING ATTRACTIONS

The chapters of this book describe important applications of mathematical, statistical, and computational methods to molecular biology. These methods are developing rapidly, and, mainly because of this situation, the presentations in this book are intended to be introductory sketches rather than scholarly reviews. Without claiming to be a complete survey, this book should convey to readers some of the exciting uses of mathematics, statistics, and computing in molecular biology. Other introductions to various aspects of molecular biology can be found in Watson et al. (1994), Streyer (1988), U.S. Department of Energy (1992), Watson et al. (1987), Lewin (1990), and Alberts et al. (1989).

Chapter 2 ("Mapping Heredity") describes how statistical models can be used to map the approximate location of genes on chromosomes. Gene mapping was mentioned above for the case of the cystic fibrosis gene. The problem becomes especially challenging—and mathematics plays a bigger role—when the disease does not follow simple Mendelian inheritance patterns—for example, when it is caused by multiple genes or when the trait is quantitative rather than qualitative in nature. This is an important subject for the Human Genome Project and its applications in modern medical genetics.

The next three chapters focus on the analysis of DNA and protein sequences. As new genes are sequenced, they are routinely compared with public databases to look for similarities that might indicate common evolutionary origin, structure, or function. As databases expand at ever-increasing rates, the computational efficiency of such comparisons is crucial. Chapter 3 ("Seeing Conserved Signals") describes combinatorial algorithms for this problem. Because coincidences abound in such comparisons, careful statistical analysis is needed. Chapter 4 ("Hearing Distant Echoes") discusses the application of extremal statistics to sequence similarity. For closely related sequences, sequence comparison also sheds light on the process of evolution. Chapter 5 ("Calibrating the Clock") discusses the applications of stochastic processes to such evolutionary analysis. The discovery and reading of genetic sequences have breathed new life into the study of the stochastic processes of evolution. The chapter focuses on one of the most exciting new tools, the use of the coalescent to estimate times to the most recent common ancestor.

Geometric methods applied to DNA structure and function are the focus of the next three chapters. Watson and Crick's famous DNA double helix can be thought of as local geometrical structure. There is also much interesting geometry in the more global structure of DNA molecules. Chapter 6 ("Winding the Double Helix") uses methods from geometry to describe the coiling and packing of chromosomes. The chapter describes the supercoiling of the double helix, in terms of key geometric quantities—link, twist, and writhe—that are related by a fundamental theorem. Chapter 7 ("Unwinding the Double Helix") employs differential mechanics to study how stresses on a DNA molecule cause it to unwind in certain areas, thereby allowing access by key enzymes needed for gene expression. Chapter 8 ("Lifting the Curtain") uses topology to infer the mechanism of enzymes that recombine DNA strands, providing a glimpse of details that cannot be seen via experiment.

Finally, Chapter 9 ("Folding the Sheets") discusses one of the hardest open questions in computational biology: the protein-folding problem, which concerns predicting the three-dimensional structure of a protein on the basis of the sequence of its amino acids. Probably no simple solution will ever be given for this central problem, but many useful and interesting approximate approaches have been developed. The concluding chapter surveys various computational approaches for structure prediction.

Together, these chapters provide glimpses of the roles of mathematics, statistics, and computing in some of the most exciting and dynamic areas of molecular biology. If this book tempts some mathematicians, statisticians, and computational scientists to learn more about and to contribute to molecular biology, it will have accomplished one of its goals. Its two other goals are to encourage molecular biologists to be more cognizant of the importance of the mathematical and computational sciences in molecular biology and to encourage scientifically literate people to be aware of the increasing impact of both molecular biology and mathematical and computational sciences on their lives. If this book makes progress toward these three goals, it shall have been well worth the effort.

REFERENCES

Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts, and J.D. Watson, 1989, *Molecular Biology of the Cell*, 2nd ed., New York: Garland.

Avery, O.T., C.M. McLeod, and M. McCarty, 1944, "Studies on the chemical nature of the substance inducing transformation of pneumococcal types," *J. Exp. Med.* **79**, 137-158.

Lewin, B., 1990, *Genes IV,* Oxford: Oxford University Press.

National Research Council, 1988, *Mapping and Sequencing the Human Genome,* Washington, D.C.: National Academy Press.

National Research Council, 1992, *DNA Technology in Forensic Science*, Washington, D.C.: National Academy Press.

Richardson, J. S., and D. C. Richardson, 1989, "Principles and patterns of protein conformation," pp. 1-98 in *Prediction of Protein Structure and the Principles of Protein Conformation,* Gerald D. Fasman (ed.), New York: Plenum Publishing Corporation.

Riordan, J.R., J.M. Rommens, B. Kreme, N. Alon, R. Rozmahel, Z. Grzelczak, J. Zielenski, S. Lok, N. Plavsic, J-L. Chou, M.L. Drumm, M.C. Innuzzi, F.S. Collins, and L-C. Tsui, 1989, "Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA," *Science* **245** (September 8), 1066-1073.

Sali, A., E. Shakhnovich, and M. Karplus, 1994. "How does a protein fold?" *Nature* **369** (19 May), 248-251.

Streyer, Lubert, 1988, *Biochemistry,* San Francisco, Calif.: W.H. Freeman.

U.S. Department of Energy, Human Genome Program, 1992, *Primer on Molecular Genetics,* Office of Energy Research, Office of Health and Environmental Research, Washington, D.C.: U.S. Government Printing Office.

Watson, J.D., N. Hopkins, J. Roberts, J.A. Steitz, and A. Weiner, 1987, *Molecular Biology of the Gene*, Menlo Park, Calif.: Benjamin-Cummings.

Watson, J.D., J. Tooze, and D.T. Kurtz, 1994. *Recombinant DNA: A Short Course,* 2nd ed., New York: W.H. Freeman and Co.

B

Challenge Problems in Bioinformatics and Computational Biology from Other Reports

B.1 GRAND CHALLENGES IN COMPUTATIONAL BIOLOGY (David Searls)[1]

1. Protein structure prediction
2. Homology searches
3. Multiple alignment and phylogeny construction
4. Genomic sequence analysis and gene-finding

B.2 OPPORTUNITIES IN MOLECULAR BIOMEDICINE IN THE ERA OF TERAFLOP COMPUTING (Klaus Schulten et al.)[2]

1. Study protein-protein and protein-nucleic acid recognition and assembly
2. Investigate integral functional units (dynamic form and function of large macromolecular and supramolecular complexes)
3. Bridge the gap between computationally feasible and functionally relevant time scales
4. Improve multiresolution structure prediction
5. Combine classical molecular dynamics simulations with quantum chemical forces
6. Sample larger sets of dynamical events and chemical species
7. Realize interactive modeling
8. Foster the development of biomolecular modeling and bioinformatics
9. Train computational biologists in teraflop technologies, numerical algorithms, and physical concepts
10. Bring experimental and computational groups in molecular biomedicine closer together.

[1] D. Searls, "Grand Challenges in Computational Biology," *Computational Methods in Molecular Biology*, S. Salzberg, D. Searls, and Simon Kasif, eds., Elsevier Science, 1998.

[2] K. Schulten, G. Budescu, F. Molnar, *Opportunities in Molecular Biomedicine in the Era of Teraflop Computing*, NIH Resource for Macromolecular Modeling and Bioinformatics, March 3-4, 1999, Rockville, MD; see http://whitepapers.zdnet.co.uk/0,39025945,60014729p-39000617q,00.htm.

B.3 WORKSHOP ON MODELING OF BIOLOGICAL SYSTEMS (Peter Kollman and Simon Levin)[3]

Challenging Issues That Span All Areas of Modeling Systems

A. Integrating data and developing models of complex systems across multiple spatial and temporal scales

- Scale relations and coupling
- Temporal complexity and coding
- Parameter estimation and treatment of uncertainty
- Statistical analysis and data mining
- Simulation modeling and prediction

B. Structure-function relationships

- Large and small nucleic acids
- Proteins
- Membrane systems
- General macromolecular assemblies
- CeIlular, tissue, organismal systems
- Ecological and evolutionary systems

C. Image analysis and visualization

- Image interpretation and data fusion
- Inverse problems
- Two-, three- and higher-dimensional visualization and virtual reality

D. Basic mathematical issues

- Formalisms for spatial and temporal encoding
- Complex geometry
- Relationships between network architecture and dynamics
- Combinatorial complexity
- Theory for systems that combine stochastic and nonlinear effects often in partially distributed systems

E. Data management

- Data modeling and data structure design
- Query algorithms, especially across heterogeneous data types
- Data server communication, especially peer-to-peer replication
- Distributed memory management and process management

B.4 WORKSHOP ON NEXT-GENERATION BIOLOGY: THE ROLE OF NEXT-GENERATION COMPUTING (Shankar Subramaniam and John Wooley)[4]

Exemplar Challenges for Bioinformatics and Computational Biology

1. Full genome-genome comparisons
2. Rapid assessment of polymorphic genetic variations

[3] "Modeling of Biological Systems," P. Kollman and S. Levin (chairs), a workshop at the National Science Foundation, March 14 and 15, 1996, available at http://www.resnet.wm.edu/~jxshix/math490/Modeling%20of%20Biological%20Systems.htm.

[4] S. Subramaniam and J. Wooley, DOE-NSF-NIH 1998 Workshop on Next-Generation Biology: The Role of Next Generation Computing, available at http://cbcg.lbl.gov/ssi-csb/nextGenBioWS.html.

3. Complete construction of orthologous and paralogous groups of genes
4. Structure determination of large macromolecular assemblies/complexes
5. Dynamical simulation of realistic oligomeric systems
6. Rapid structural/topological clustering of proteins
7. Prediction of unknown molecular structures; protein folding
8. Computer simulation of membrane structure and dynamic function
9. Simulation of genetic networks and the sensitivity of these pathways to component stoichiometry and kinetics
10. Integration of observations across scales of vastly different dimensions and organization to yield realistic environmental models for basic biology and societal needs

B.5 TECHNOLOGIES FOR BIOLOGICAL COMPUTER-AIDED DESIGN (Masaru Tomita)[5]

1. Enzyme engineering: to refine enzymes and to analyze kinetic parameters in vitro
2. Metabolic engineering: to analyze flux rates in vivo
3. Analytical chemistry: to determine and analyze the quantity of metabolites efficiently
4. Genetic engineering: to cut and paste genes on demand, for modifying metabolic pathways
5. Simulation science: to efficiently and accurately simulate a large number of reactions
6. Knowledge engineering: to construct, edit and maintain large metabolic knowledge bases
7. Mathematical engineering: to estimate and tune unknown parameters

B.6 TOP BIOINFORMATICS CHALLENGES (Chris Burge et al.)[6]

1. Precise, predictive model of transcription initiation and termination: ability to predict where and when transcription will occur in a genome
2. Precise, predictive model of RNA splicing/alternative splicing: ability to predict the splicing pattern of any primary transcript
3. Precise, quantitative models of signal transduction pathways:ability to predict cellular response to external stimuli
4. Determining effective protein-DNA, protein-RNA and protein-protein recognition codes
5. Accurate ab initio structure prediction
6. Rational design of small molecule inhibitors of proteins
7. Mechanistic understanding of protein evolution: understanding exactly how new protein functions evolve
8. Mechanistic understanding of speciation: molecular details of how speciation occurs
9. Continued development of effective gene ontologies-systematic ways to describe the functions of any gene or protein
10. (Infrastructure and education challenge)
11. Education: development of appropriate bioinformatics curricula for secondary, undergraduate, and graduate education

B.7 EMERGING FIELDS IN BIOINFORMATICS (Patricia Babbitt)[7]

1. Data storage and retrieval, database structures, annotation
2. Analysis of genomic/proteomic/other high-throughput information

[5] M. Tomita, "Towards Computer Aided Design (CAD) of Useful Microorganisms," *Bioinformatics* 17(12):1091-1092, 2001.

[6] C. Burge, "Bioinformaticists Will Be Busy Bees," *Genome Technology*, No. 17, January, 2002. Available (by free subscription) at http://www.genome-technology.com/articles/view-article.asp?Article=20021023161457.

[7] P. Babbitt et al., "A Very Very Very Short Introduction to Protein Bioinformatics," August 22-23, 2002, University of California, San Francisco, available at http://baygenomics.ucsf.edu/education/workshop1/lectures/w1.print2.pdf.

3. Evolutionary model building and phylogenic analysis
4. Architecture and content of genomes
5. Complex systems analysis/genetic circuits
6. Information content in DNA, RNA, protein sequences and structure
7. Metabolic computing
8. Data mining using machine learning tools, neural nets, artificial intelligence
9. Nucleic acid and protein sequence analyses

B.8 TEN GRAND CHALLENGES (Sylvia Spengler)[8]

1. The origin, structure, and fate of the universe
2. The fundamental structure of matter
3. Earth's physical systems
4. The diversity of life on Earth
5. The tree of life
6. The language of life
7. The web of life
8. Human ecology
9. The brain and artificial thinking machines
10. Integrating Earth and human systems
11. A knowledge server for planetary management

Research Across Domains: Data

- Information management—human evolution continued
- Exponential increase in data and information across domains
- Access to information across domains—as or more important than the information itself
- Integration of data across knowledge domains
- Apply analytical tools across knowledge domains
- Modeling of complex systems
- Simulation of phenomena—descriptive science becomes predictive science

Research Across Domains: People

- Share data across disciplines
- Build and use analytical and modeling tools across disciplines
- Work in collaborative, cross-domain groups

Research Across Domains: Time

- Real-time data access, integration, and analysis
- Real-time modeling and effects prediction
- Real-time dissemination of research results
- Real-time testing by research community
- Real-time policy discussions
- Real-time policy decisions

[8] S. Spengler, Lawrence Berkeley National Laboratory, personal communication to John Wooley, January 3, 2005.

B.9 GRAND CHALLENGES IN BIOMEDICAL COMPUTING (John A. Board, Jr.)[9]

Biomedical Applications from Coupling Imaging and Modeling

- Real-time noninvasive three-dimensional imaging of many body systems
- Real-time generation of three-dimensional patient-specific models
- Multiple-technology (multimodal) imaging and modeling
- Whole-organ modeling
- Multiple-organ system modeling
- Patient-specific modeling of organ anomalies
- Model support for (partial) restoration of hearing, coarse vision, and locomotion (via both paralyzed and artificial limbs)

All of these applications make use of:

- Three-dimensional models
- Increasingly refined grids and increasing levels of tissue discrimination
- Anatomically realistic models
- Special-purpose hardware for visualization
- Distributed computing techniques.

B.10 ACCELERATING MATHEMATICAL-BIOLOGICAL LINKAGES: REPORT OF A JOINT NSF-NIH WORKSHOP (Margaret Palmer et al.)[10]

List of Top Ten Problems at the Mathematical Biology Interface

1. Model multilevel systems: from the cells in people, to human communities in physical, chemical, and biotic ecologies.
2. Model networks of complex metabolic pathways, cell signaling, and species interactions.
3. Integrate probabilistic theories: understand uncertainty and risk.
4. Understand computation: gaining insight and proving theorems from numerical computation and agent-based models.
5. Provide tools for data mining and inference.
6. Address linguistic and graph theoretical approaches.
7. Model brain function.
8. Build computational tools for problems with multiple temporal and spatial scales.
9. Provide ecological forecasts.
10. Understand effects of erroneous data on biological understanding.

B.11 GRAND CHALLENGES OF MULTIMODAL BIOMEDICAL SYSTEMS (J. Chen et al.)[11]

Science Challenges

1. Allow early detection of where and when an infectious disease outbreak occurs, whether it is naturally occurring or man-made, in real time.

[9] J.A. Board, Jr., "Grand Challenges in Biomedical Computing, *High-Performance Computing in Biomedical Research*, T.C. Pilkington, B. Loftis, J.F. Thompson, S.L.Y. Woo, T.C. Palmer, and T.F. Budinger, eds., CRC Press, Boca Raton, FL, 1993.

[10] M. Palmer et al., "Accelerating Mathematical-Biological Linkages: Report of a Joint NSF-NIH Workshop," February 2003, available at www.maa.org/mtc/NIH-feb03-report.pdf.

[11] J. Chen et al., "Grand Challenges of Multimodal Bio-Medical Systems," *IEEE Circuits and Systems Magazine*, pp. 46-52, 2nd Quarter 2005, available at http://gsp.tamu.edu/Publications/PDFpapers/pap_CASmag_MBM.pdf.

2. Develop multidimensional drug profiling databases to facilitate drug discovery and to identify biomarkers for diagnosis and monitoring the progress of individual disease treatments.
3. Connect activities and events derived from cellular processes to high-level cognitions.
4. Support personalized medical care and clinical decision for patients

Technology Challenges and Enabling Technologies

1. Formalization of biological knowledge into predictive models for systems biology and system-based analysis
2. Interdisciplinary training
3. Development of open source, multiscale modality informatics toolkits

B.12 THE DEPARTMENT OF ENERGY'S GENOMES TO LIFE PROGRAM[12]

21st Century Biology Requiring "Biocomp" Tools

1. Population models, symbiosis, and stability
2. Discrete growth models
3. Reaction kinetics
4. Biological oscillators and switches
5. Coupled oscillators
6. Reaction-diffusion, chemotaxis, and nonlocality
7. Oscillator-generated wave phenomena and patterns
8. Spatial pattern formation with population interactions
9. Mechanical models for generating pattern and form in development
10. Evolution and morphogenesis

A Mathematica for Molecular, Cellular, and Systems Biology

1. Core data models and structures [database management]
2. Optimized functions [core libraries]
3. Scripting environment [e.g., Python, PERL, ruby, etc.]
4. Database accessors and built-in schemas
5. Simulation interfaces
6. Parallel and accelerated kernels
7. Visualization interfaces (for information visualization and scientific visualization)
8. Collaborative workflow and group use interfaces

Hierarchical Biological Modeling Environment

1. Genetic sequences
2. Molecular machines
3. Molecular complexes and modules
4. Networks + pathways [metabolic, signaling, regulation]
5. Structural components [ultrastructures]
6. Cell structure and morphology
7. Extracellular environment
8. Populations and consortia

[12] R. Stevens, "GTL Software Infrastructure: A Computer Science Perspective," undated presentation, Argonne National Laboratory, available at www.doegenomics.org/compbio/mtg_1_22_02/RickStevens.ppt.

Modeling and Simulation Challenges for 21st Century Biology

1. Modeling activity of single genes
2. Probabilistic models of prokaryotic genes and regulation
3. Logical models of regulatory control in eukaryotic systems
4. Gene regulation networks and genetic network inference in computational models and applications to large-scale gene expression data
5. Atomistic-level simulation of biomolecules
6. Diffusion phenomena in cytoplasm and extracellular environment
7. Kinetic models of excitable membranes and synaptic interactions
8. Stochastic simulation of cell signaling pathways
9. Complex dynamics of cell cycle regulation
10. Model simplification

B.13 HIGH-PERFORMANCE COMPUTING, COMMUNICATION, AND INFORMATION TECHNOLOGY GRAND CHALLENGES (LATE 1980s, EARLY 1990s)[13]

Computing Applications to Map and Sequence Human Genome

1. Understanding protein folding
2. Predicting structure of native protein
3. Exhaustive discovery and analysis of cancer genes
4. Molecular recognition and dynamics
5. Drug discovery

[13] Committee on Physical, Mathematical, and Engineering Sciences of the Federal Coordinating Council for Science, Engineering, and Technology, U.S. Office of Science and Technology Policy, FY1992 Blue Book: *Grand Challenges: High Performance Computing and Communications—The FY 1992 U.S. Research and Development Program.*

C

Biographies of Committee Members and Staff

C.1 COMMITTEE MEMBERS

JOHN C. WOOLEY (*Chair*) is the associate vice chancellor for research, University of California at San Diego (UCSD), an adjunct professor in pharmacology, and in chemistry and biochemistry, and a strategic advisor and senior fellow of the San Diego Supercomputer Center. He received his Ph.D. degree in 1975 at the University of Chicago, working with Al Crewe and Robert Uretz in biological physics. Dr. Wooley created the first programs within the U.S. federal government for funding research in bioinformatics and in computational biology and has been involved in strengthening the interface between computing and biology for more than a decade. For the new UCSD California Institute for Telecommunication and Information Technology (Cal-(IT)2), Dr. Wooley directs the biology and biomedical layer or applications component, termed Digitally-enabled Genomic Medicine (DeGeM), a step in delivering personalized medicine in a wireless clinical setting. His current research involves bioinformatics and structural genomics, while his principal objectives at UCSD are to stimulate new research initiatives for large-scale, multidisciplinary challenges. He also collaborates in developing scientific applications of information technology and high-performance computing; creating industry-university collaborations; expanding applied life science opportunities, notably around drug discovery; and establishing a biotechnology and pharmacology science park on UCSD's health sciences campus zone.

ADAM P. ARKIN is a faculty scientist in computational and theoretical biology at Lawrence Berkeley National Laboratory, an assistant professor of bioengineering and chemistry at the University of California, Berkeley, and an investigator of the Howard Hughes Medical Institute. His focus is on detailed modeling of genetic and biochemical networks with emphasis on developmental systems. The Arkin laboratory applies theoretical and computational analyses from dynamical systems, stochastic processes, chemical kinetics, and statistical mechanics and methods from molecular biology to determine the principles of cellular signal processing and to aid in design of custom cellular circuitry that may, for example, act as sensitive biosensors.

ERIC BRILL is a researcher in the Machine Learning and Applied Statistics Group at Microsoft Research. His research interests include natural language processing (primarily empirical natural language processing), speech recognition and spoken language systems, machine learning, and artificial intelligence. Some specific research topics include lexical disambiguation, parsing, classifier combination, spelling correction, and language modeling. Before joining Microsoft, he was an assistant professor of computer science at Johns Hopkins University. He has served on the editorial board of *Computational Linguistics* and the *Journal for Artificial Intelligence Research*. Dr. Brill received his Ph.D. in computer science from the University of Pennsylvania in 1993.

ROBERT M. CORN is a professor in the Department of Chemistry at the University of California, Irvine. Dr. Corn is a leader in the development and application of surface-sensitive spectroscopic techniques such as surface plasmon resonance (SPR) imaging, optical second harmonic generation (SHG), and polarization modulation Fourier transform infrared (PM-FTIR) spectroscopy. His primary research interests include the study of biopolymer (e.g., DNA, protein) adsorption onto surfaces and the chemical modification of surfaces for the creation of ultrathin films and adsorption-based biosensors. Professor Corn also has ongoing research projects in the implementation of DNA computing algorithms at surfaces and the study of ion transfer processes at liquid-liquid interfaces. He received a B.A. in chemistry summa cum laude in 1978 from the University of California, San Diego, and earned a Ph.D. in 1983 from the University of California, Berkeley, under the direction of Professor Herbert L. Strauss in the application of FTIR to the study of motion in molecular solids. From 1983 to1984, Professor Corn was a visiting scientist at the IBM Research Laboratory in San Jose, California, where he applied the techniques of surface plasmon-enhanced Raman scattering and optical SHE to electrochemical surfaces. In 1985, Professor Corn moved to Wisconsin where he was a member of the Analytical Sciences Division of the Department of Chemistry and the Water Chemistry Program until 2004. In July of 2004, he moved to the University of California, Irvine, where he joined the Department of Chemistry. Professor Corn is a co-founder of two companies: GWC Technologies, Inc., maker of SPR instrumentation and other surface spectroscopic equipment, and GenTel BioSurfaces, Inc.

CHRIS DIORIO is an associate professor of computer science and engineering at the University of Washington. His research focuses on building electronic systems that employ the computational and organizational principles used in the nervous systems of living organisms. This work on neurally inspired computing includes studies of computing with action potentials, silicon learning systems, and implantable computers. He also works on high-speed circuit design. Dr. Diorio teaches courses in both digital electronics and integrated-circuit (IC) design, and is developing new course material in two areas: (1) alternative computing paradigms, including neural, quantum, and DNA computers, and (2) digital IC design at microwave clock frequencies. He received a National Science Foundation (NSF) Presidential Early Career Award in 1999. Dr. Diorio was awarded a 5 year Packard Foundation Fellowship in science and engineering in 1998 and also an NSF Career Award that same year. In 1996, he was awarded the Electron Devices Society's (EDS's) Paul Rappaport Award for the best paper in an Institute of Electrical and Electronics Engineers EDS publication. He completed his doctoral research in electrical engineering at the Physics of Computation Laboratory, California Institute of Technology, in 1997. Dr. Diorio has also served as a senior staff engineer for TRW, Inc., and as a senior staff scientist for American Systems Corporation. He received his B.A. in physics from Occidental College in 1983 and his M.S. in electrical engineering in 1984 from The California Institute of Technology.

LEAH EDELSTEIN-KESHET is a professor of mathematics at the University of British Columbia. She received her Ph.D. in 1982 from the Weizmann Institute of Science in Rehovot, Israel, specializing in applied mathematics and working with Professor Lee A. Segel. She is a member of the Mathematics Department and the Institute of Applied Mathematics at the University of British Colombia. She is also a former president of the Society for Mathematical Biology. Although her main area of interest is mathematical biology, Dr. Edelstein-Keshet works in several areas, including the molecular biology of the cytoskeleton, the dynamics of swarming and social organisms and, more recently, models for neuroinflammation in Alzheimer's disease and pathogenesis of type 1 (autoimmune) diabetes.

MARK H. ELLISMAN is professor in the Department of Neurosciences at the School of Medicine and the Department of Bioengineering, director of the National Center for Microscopy and Imaging Research at UCSD, and chair of the San Diego Supercomputer Center (SDSC) Executive Committee. Dr. Ellisman's research focuses on cellular neurobiology and the dynamic interplay between structure and function in the nervous system, with a focus on excitable membrane properties and enabling remote access to large-scale scientific instrumentation. At UCSD, Dr. Ellisman is director of the Center for Research in Biological Structure and director of the Neurosciences Laboratory for Neurocytology. Since 1997, he has been the neuroscience thrust leader and cross-disciplinary coordinator for the National Partnership for Advanced Computational Infrastructure. Dr. Ellisman is a member of the American Association for the Advancement of Science, Society for Neurosciences, and American Institute for Medical and Biological Engineering. He has served on numerous editorial boards and has been associate editor of the *Journal of Neurocytology* since 1980. Dr. Ellisman is a also grant reviewer for organizations such as the National Institutes of Health and the National Science Foundation, and a consultant for associations such as the Association for Advanced Technology in the Biomedical Sciences and Pfizer. He has published numerous journal and conference articles and technical reports. He holds a Ph.D. degree in biology and an M.A. degree in neurophysiology both from the University of Colorado, Boulder, and an A.B. degree with honors from the University of California, Berkeley.

MARCUS W. FELDMAN is a professor of biological sciences at Stanford University. He uses applied mathematics and computer modeling to simulate and analyze the process of evolution. Specific areas of research include the evolution of complex genetic systems that can undergo both natural selection and recombination and the evolution of learning as one interface between modern methods in artificial intelligence and models of biological processes, including communication. He also studies the evolution of modern humans using models for the dynamics of molecular polymorphisms, especially DNA variants. He is managing editor of *Theoretical Population Biology* and associate editor of *Genetics* and of *Complexity*. Dr. Feldman is a member of the American Society of Naturalists, and the American Society of Human Genetics, and a fellow of the American Academy of Arts and Sciences. He received his B.Sc. in 1964 from the University of Western Australia, his M.Sc. in 1966 from Monash University, Australia, and his Ph.D. in biomathematics from Stanford in 1969.

DAVID K. GIFFORD is a professor of electrical engineering and computer science at the Massachusetts Institute of Technology. He is working on the analysis of RNA expression data using graphical models. Professor Gifford has also developed programmed mutagenesis, a technique for programmatically rewriting DNA sequences by incorporating sequence-specific oligonucleotides into newly manufactured strands of DNA. Dr. Gifford serves as group leader for the Programming Systems Research Group at the MIT Laboratory for Computer Science. This group is dedicated to finding new ways of programming existing systems and developing new programmable systems. The group's efforts concentrate on combining existing technologies and inventing new ones to deliver new ways of computing in selected areas: programming language development; information discovery, retrieval, and distribution; algebraic and computational video; and most recently, computation using biological substrates. Dr. Gifford earned his S.B. in 1976 from MIT and his M.S. and Ph.D. in electrical engineering from Stanford University in 1978 and 1981, respectively. He is a tenured member of the MIT faculty, which he joined in 1982. He was appointed to the Karl Van Tassel Career Development Chair at MIT in 1990.

TAKEO KANADE received his Ph.D. in electrical engineering from Kyoto University, Japan, in 1974. After being on the faculty in the Department of Information Science, Kyoto University, he joined the Computer Science Department and Robotics Institute in 1980. He became associate professor in 1982, a full professor in 1985, the U.A. and Helen Whitaker Professor in 1993, and a University Professor in 1998. He has been the Director of the Robotics Institute since 1992. He served as the founding chairman (1989-1993) of the robotics Ph.D. program at Carnegie Mellon University, probably the first of its kind in the world. Dr. Kanade has worked in multiple areas of robotics, ranging from manipulator, sensor, computer vision, and multimedia applications to autonomous robots, with more than 200 papers on these topics. He is the founding editor of the *International Journal of Computer Vision*. Dr. Kanade's professional honors include election to the National Academy of Engineering, a fellow of the IEEE, a fellow of the ACM, and a fellow of the American Association of Artificial Intelligence, and several awards including the Joseph Engelberger Award, Yokogawa Prize, JARA Award, Otto Franc Award, and Marr Prize Award.

STEPHEN S. LADERMAN is the manager of the Molecular Diagnostics Department, dedicated to molecular biology, biochemistry, computational biology, and engineering for the development of genetic, genomic, and proteomic analysis systems for biomedical research and molecular diagnostics. He earned his B.A. in physics, magna cum laude, from Wesleyan University in 1976 and his Ph.D. in materials science and engineering from Stanford University in 1983. Dr. Laderman was a postdoctoral Scholar from 1982 to 1984 at Stanford University and Exxon Research Corporation. Before joining Agilent Labs, he worked in a variety of positions at Hewlett-Packard Laboratories. Dr. Laderman was a member of the Basic Energy Sciences Advisory Committee Panel on Novel, Coherent Light Sources and chair of the selection committee for the George E. Pake Prize of the American Physical Society. He is currently a member of the International Society for Computational Biology, American Society of Human Genetics, American Physical Society, American Chemical Society, American Association for the Advancement of Science, and a senior member of the IEEE.

JAMES S. SCHWABER is associate professor of pathology, anatomy and cell biology at Thomas Jefferson University Medical College (TJU) and is Director of the Daniel Baugh Institute for Functional Genomics and Computational Biology at TJU. Prior to joining TJU in 2000, he was technical leader and research fellow of the Computational Biology Program in the Core Genomics Group at DuPont. His interest is in neuron and neuronal network modeling (e.g., of cardiorespiratory control functions) and, in particular, how alterations in neuron properties will be dependent on input activity over time, by linking the molecular processes activated by synaptic inputs to cell physiology. His research group focuses on computational analysis of genomic datasets from functionally identified neurons as a cornerstone to support modeling of the adaptive intracellular response to synaptic inputs. Currently the work is related to systems analysis of gene regulatory circuits, the modeling of neuronal inputs into these circuits as modular patterns of transcription factor activation, and the central issue of discovering principles that relate gene output to functional phenotype (electrophysiology; models of ion fluxes) at the systems level.

C.2 STAFF MEMBERS

Herbert S. Lin is senior scientist and senior staff officer at the Computer Science and Telecommunications Board (CSTB), National Research Council (NRC) of the National Academies, where he has been the study director for major projects on public policy and information technology. These studies include a 1996 study on national cryptography policy (*Cryptography's Role in Securing the Information Society*), a 1991 study on the future of computer science (*Computing the Future*), a 1999 study of Defense Department systems for command, control, communications, computing, and intelligence (*Realizing the Potential of C4I: Fundamental Challenges*), and a 2000 study on workforce issues in high technology (*Building a Workforce for the Information Economy*). Prior to his NRC service, he was a professional staff member and staff scientist for the House Armed Services Committee (1986 to 1990), where his portfolio included defense policy and arms control issues. He also has significant expertise in math and science education. He received his Ph.D. in physics from MIT in 1979. Avocationally, he is a long-time folk and swing dancer, and a poor magician. In addition to his CSTB work, he is published in cognitive science, science education, biophysics, and arms control and defense policy.

Robin Schoen is the director of the Board on Agriculture and Natural Resources (BANR) of the National Academies. Prior to joining BANR in March 2005, she was a senior program officer for the Academies' Board on Life Sciences, where she directed several studies, including *Discovery of Antivirals Against Smallpox; Stem Cells and the Promise of Regenerative Medicine*; *The National Plant Genome Initiative: Objectives for 2003-2005*; *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences;* and a BANR study titled *Predicting Invasions of Nonindigenous Plants and Plant Pests*. Robin received a B.S. in biology and chemistry from Frostburg State College, Maryland, and an M.A. in science and technology policy from George Washington University.

C.3 REPORT COORDINATOR

Russ Biagio Altman is a professor of genetics, bioengineering and medicine (and of computer science by courtesy) at Stanford University. His primary research interests are in the application of computing technology to basic molecular biological problems of relevance to medicine. He is currently developing techniques for collaborative scientific computation over the Internet, including novel user interfaces to biological data, particularly for pharmacogenomics. Other work focuses on the analysis of functional microenvironments within macromolecules and the application of nonlinear optimization algorithms for determining the structure and function of biological macromolecules, particularly the bacterial ribosome. Dr. Altman holds an M.D. from Stanford Medical School, a Ph.D. in medical information sciences from Stanford, and an A.B. from Harvard College. He has been the recipient of the U.S. Presidential Early Career Award for Scientists and Engineers, an NSF Career Award, and the Western Society of Clinical Investigation Annual Young Investigator Award. He is a fellow of the American College of Physicians and the American College of Medical Informatics. He is a past-president and founding board member of the International Society for Computational Biology, an organizer of the annual Pacific Symposium on Biocomputing, and an associate editor of the journal *Bioinformatics*. He currently directs the Stanford Center for Biomedical Computation and the training program in Biomedical Informatics, and he won the Stanford Medical School graduate teaching award in 2000.

D

Workshop Participants

To assist in its information-gathering efforts, the Committee on Frontiers at the Interface of Computing and Biology held three workshops on various topics at the interface of computing and biology. The participants in these workshops are listed below.

WORKSHOP ON BIO-INSPIRED COMPUTING AND ENABLING TECHNOLOGIES (JANUARY 2001)

Rick Adrion, National Science Foundation
Roger Brent, Molecular Sciences Institute
Anne Condon, University of British Columbia
Mita Desai, National Science Foundation
Stephanie Forrest, University of New Mexico
Bob Full, University of California, Berkeley
James J. Hickman, National Science Foundation
Ken Johnson, Gen-Tel, Inc.
Tom Knight, Massachusetts Institute of Technology
Christof Koch, California Institute of Technology
Patricia Mead, National Academy of Engineering
Allen Northrup, Cepheid
Shankar Sastry, Defense Advanced Research Projects Agency
Shihab Shamma, University of Maryland
Sylvia Spengler, National Science Foundation
Gary Strong, National Science Foundation
Erik Winfree, California Institute of Technology

WORKSHOP ON CHALLENGES AND OPPORTUNITIES IN DATA MANAGEMENT (MARCH 2001)

Helen Berman, Rutgers University
Pat Brown, Stanford University
Barb Bryant, Millennium Predictive Medicine
Mike Colvin, Lawrence Livermore National Laboratory
Stephen Dahms, California State University Program for Education and Research in Biotechnology
Dan Davison, Bristol Myers Squibb
Joe Deken, Southern Illinois University
Skip Garner, Southwestern Medical Center
Jim Gray, Microsoft
David Haussler, University of California, Santa Cruz
Dick Karp, University of California, Berkeley
David Kingsbury, Discovery Biosciences Corporation
Michael Marron, National Center for Research Resources
Dan Masys, University of California, San Diego
Richard Morris, National Institute of Allergy and Infectious Diseases
Bernhard Palsson, University of California, San Diego
Larry Smarr, University of California, San Diego
Paul Spellman, Stanford University
Sylvia Spengler, National Science Foundation
Gary Strong, National Science Foundation
Art Toga, University of California, Los Angeles
Chris Wood, Los Alamos National Laboratory

WORKSHOP ON MODELING OF BIOLOGICAL SYSTEMS (MAY 2001)

Rick Adrion, National Science Foundation
Ruzena Bajcsy, National Science Foundation
Eugene Bruce, National Science Foundation
Marvin Cassman, National Institutes of Health
Su Chung, geneticXchange
Jim Collins, Boston University
Joe Decken, University of California, San Diego
Mita Desai, National Science Foundation
Drew Endy, Molecular Sciences Institute
Warren Ewens, University of Pennsylvania
Joe Felsenstein, University of Washington
Teresa Head-Gordon, Lawrence Berkeley National Laboratory
James Hickman, National Science Foundation
Sri Kumar, Defense Advanced Research Projects Agency
Simon Levin, Princeton University
Michael Marron, National Center for Research Resources
Andrew McCulloch, University of California, San Diego
Garrett Odell, University of Washington
Dave Polidori, Entelos, Inc.
Terry Sejnowski, Salk Institute
Sylvia J. Spengler, National Science Foundation
Gary Strong, National Science Foundation
John J. Tyson, Virginia Polytechnic Institute and State University

What Is CSTB?

As a part of the National Research Council, the Computer Science and Telecommunications Board (CSTB) was established in 1986 to provide independent advice to the federal government on technical and public policy issues relating to computing and communications. Composed of leaders from industry and academia, CSTB conducts studies of critical national issues and makes recommendations to government, industry, and academic researchers. CSTB also provides a neutral meeting ground for consideration of complex issues where resolution and action may be premature. It convenes invitational discussions that bring together principals from the public and private sectors, ensuring consideration of all perspectives. The majority of CSTB's work is requested by federal agencies and Congress, consistent with its National Academies context.

A pioneer in framing and analyzing Internet policy issues, CSTB is unique in its comprehensive scope and effective, interdisciplinary appraisal of technical, economic, social, and policy issues. From its early work in computer and communications security, cyber-assurance and information systems trustworthiness have been cross-cutting themes in CSTB's work. CSTB has produced several reports regarded as classics in the field, and it continues to address these topics as they grow in importance.

To do its work, CSTB draws on some of the best minds in the country, inviting experts to participate in its projects as a public service. Studies are conducted by balanced committees without direct financial interests in the topics they are addressing. Those committees meet, confer electronically, and build analyses through their deliberations. Additional expertise from around the country is tapped in a rigorous process of review and critique, further enhancing the quality of CSTB reports. By engaging groups of principals, CSTB obtains the facts and insights critical to assessing key issues.

The mission of CSTB is to

Respond to requests from the government, nonprofit organizations, and private industry for advice on computer and telecommunications issues and from the government for advice on computer and telecommunications systems planning, utilization, and modernization;

Monitor and promote the health of the fields of computer science and telecommunications, with attention to issues of human resources, information infrastructure, and societal impacts;

Initiate and conduct studies involving computer science, computer technology, and telecommunications as critical resources; and

Foster interaction among the disciplines underlying computing and telecommunications technologies and other fields, at large and within the National Academies.

More information about CSTB can be obtained online at http://www.cstb.org.