

# An assessment of the role of computing in systems biology

S. Burbeck  
K. E. Jordan

*Systems biology is a burgeoning field in which researchers are investigating a flood of new data that is gathered in high-throughput genomics, proteomics, and related analyses. Systems biologists focus on what this data reveals about the functioning of living systems. The large volume of data, and the complexity of living systems, ensures that computing plays a central role in analyzing, modeling, and simulating these systems. In this paper, we discuss some of the key challenges in the field of computational systems biology. We also discuss possible ways in which the field of systems biology may evolve in coming years, along with some of the demands that systems biology research places on computing resources.*

## Introduction

Since the 1980s, computing has changed many aspects of the way in which biological science is conducted. For example, computer-automated techniques have dramatically enhanced scientists' ability to sequence DNA and to quantitate and identify proteins and RNA transcripts present in biological samples. The cost of sequencing DNA has fallen roughly a thousandfold during the period from 1990 to 2006. These kinds of rapid analysis techniques are examples of "high-throughput" biology. Computer databases store, manage, and make accessible the flood of data that results from high-throughput methods. A visitor to a large genome-sequencing center, such as the Sanger Institute, located near Cambridge, UK, encounters large rooms filled with computer-controlled DNA-sequencing machines and other large rooms filled with many racks of high-powered servers and arrays of storage disks. A small amount of biological material enters these sequencing machines and, after a great deal of computer processing, megabytes or gigabytes of new genomic data are deposited in the databases and become available on the Web.

However, the gathering of ever more data is not an end in itself. Data gathering simply opens the way for investigating how the molecular components, which are found by using high-throughput techniques, work together in living systems. Various scientific efforts to address that challenge have been consolidated under the

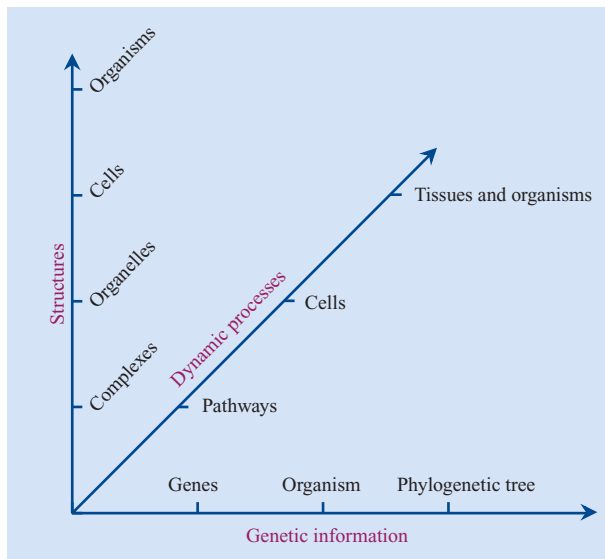
term "systems biology," also sometimes called *in silico* biology or integrative biology. All of these terms are imprecise. In this paper, we use "systems biology" because it is the most frequently used term and it emphasizes the importance of the complex systems under study. Whatever terminology is used, computing that involves modeling and simulation is clearly becoming more closely intertwined with the biological research and the science itself [1].

To explore the question of how systems biology is likely to grow and the roles that computing will play, we conducted a series of in-depth discussions with leaders in the field. We interviewed people at pharmaceutical companies (Merck and GlaxoSmithKline), research institutes (e.g., the Institute for Systems Biology, the European Molecular Biology Laboratory, and The Molecular Sciences Institute), biotechnology firms (e.g., BG Medicine and Genomatica), universities (e.g., University of California, Berkeley; University of California, San Diego; Cambridge University; and Oxford University) and biological standards groups [e.g., the Systems Biology Markup Language (SBML) group]. See the Appendix for a full list of interviewees whose insights made this report possible. However, note that the views expressed here are the responsibility of the authors.

In this paper, we first characterize various kinds of efforts that can be considered to be part of systems biology. We then discuss some of the challenges that must

©Copyright 2006 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/06/\$5.00 © 2006 IBM



**Figure 1**

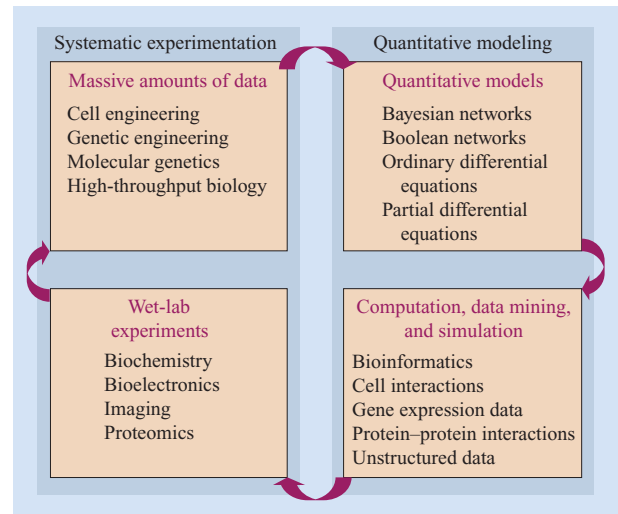
The space of systems biology. Research is required in all areas of this space in order to understand the complexity and interactions among the various structure, information, and time scales.

be overcome for systems biology to achieve its promise. Finally, we provide a brief discussion of how the field may progress over the next few years.

### What is systems biology?

If one thinks of the output of high-throughput biology as a large living systems “parts list,” which includes components such as genes, proteins, RNA transcripts, and various other biomolecules, then researchers in systems biology seek to understand how the various parts fit and function together in the larger systems. Researchers investigate how synergy arises in these systems, the behavior of which cannot be understood by examining the parts in isolation.

The goals of systems biology, which address the biological areas depicted in **Figure 1**, are not new. Systems biology researchers seek to understand the structural, genetic, and dynamic relationships of living systems at scales that range from small biomolecular complexes and biochemical pathways to whole organisms and even interrelated ecologies of organisms. Many biologists, including several of the systems biologists we interviewed, have pointed out that understanding these relationships has always been the goal of biology. However, making sense of the mass of new data calls for computing to play an increasingly important role in meeting the goals of biological research, and this new focus warrants coining a new name. In a sense, computational systems biology is an inevitable



**Figure 2**

Systems biology paradigm: The essential cycle between wet-lab experimentation and computational modeling and analysis.

consequence of the success of high-throughput biology. The term is used in connection with a wide variety of scientific efforts, out of which four major themes and focus areas have emerged:

- *Automated analysis.* Sophisticated computer analysis techniques make inferences or predictions from various kinds of high-throughput data. Examples include predicting structure and function from protein sequence information, and automated DNA-sequence annotation, using various techniques for identifying potential genes and other functional sequences.
- *Modeling.* Descriptive computer models are built that summarize and organize data. For example, pathway or network models are derived from DNA expression data and from protein–protein or protein–DNA interaction data.
- *Simulation.* Predictive dynamic computational models (e.g., ordinary or partial differential equation models) are built that use researchers’ knowledge of the biological parts in order to better understand the implications of their interactions and their roles in larger systems.
- *Integration of computational biology and experimental biology.* Computational results are used to guide new experimentation followed by additional computational analysis and modeling of the new data (**Figure 2**). Most leaders in the field consider this marriage of computation and experimentation to be vital to systems biology. The complexities of

**Table 1** Nomenclature, tools, and challenges at the different levels in the systems biology hierarchy.

<i>Level of organization</i>	<i>Entities</i>	<i>Computational tools and techniques</i>	<i>Challenges</i>
Parts	Genes, proteins, DNA binding sites, splice sites, membrane targeting signals	<ul style="list-style-type: none"> <li>• Sequence matching</li> <li>• Gene prediction</li> <li>• Protein structure prediction</li> <li>• Data mining</li> </ul>	<ul style="list-style-type: none"> <li>• Consistent naming</li> <li>• Referential integrity</li> <li>• Curation and accuracy</li> <li>• Ontologies</li> </ul>
Networks and pathways	Pairwise relationships between parts, such as the entities in row 1	<ul style="list-style-type: none"> <li>• Path tracing in networks, shortest path finding</li> <li>• Cluster analysis</li> <li>• Descriptive models and simulations</li> <li>• Entity-relationship diagram tools (e.g., Gene Ontology tools)</li> <li>• Visualization</li> </ul>	<ul style="list-style-type: none"> <li>• Separating meaningful relationships from artifacts</li> <li>• Understanding temporal relationships</li> <li>• Understanding effects of perturbations</li> </ul>
Assemblies and complexes	3D structures of parts, such as proteins and RNAs	<ul style="list-style-type: none"> <li>• Modeling</li> <li>• Simulation</li> <li>• Visualization</li> </ul>	<ul style="list-style-type: none"> <li>• Description and modeling of location in the cell</li> <li>• How structure determines function</li> <li>• Relationship to pathways</li> <li>• Interactions with other assemblies</li> </ul>
Systems	Structures of substructures, such as organelles, cytoskeleton, cells, biofilms, tissues, organs	<ul style="list-style-type: none"> <li>• Modeling</li> <li>• Simulation</li> <li>• Visualization</li> </ul>	<ul style="list-style-type: none"> <li>• Prediction and explanation across levels of hierarchy</li> <li>• Building multiscale models, both physical and temporal</li> <li>• 3D structural connections</li> <li>• Experimental validation of models and simulations</li> </ul>

living systems can be disentangled only by an iterative process of modeling, simulation, and wet lab experimentation. The cycle of continual refinement, represented by arrows in the figure, requires a multidisciplinary effort.

### The hierarchy of biological systems

Living systems are organized at many different spatial scales, ranging from individual small molecules, such as amino acids and nucleotides, to whole organisms. The elements at a given level (such as genes and proteins) are organized into both logical and physical structures. The sequences of nucleotides that make up DNA and RNA are organized into higher-level logical structures such as regulatory binding sites, splice sites, exons (DNA regions that code for proteins), introns (DNA regions that do not code for proteins), and genes. DNA is also organized into larger physical structures such as heterochromatin and chromosomes. The sequences of amino acids in proteins are logically organized into leader and mature sequences, and the mature proteins are physically organized, or folded, into secondary and tertiary structures that are crucial for proper protein function. At the next larger scale, sets of proteins can be logically organized into pathways—for example, networks in which the nodes are proteins, and each connecting arc represents the potential

binding of one protein to another or a modification of one protein by the other, such as in the process of phosphorylation. Sets of proteins, which may contain small molecules of RNA, are often physically organized into multi-molecular structures called assemblies or complexes. Two examples of many such complexes are the nucleosome and the ribosome [2]. The ways in which these various structures interact determine the workings of a cell. Cells, and the extracellular matrix they produce, in turn form tissues that are organized into organs. Systems biology aims to understand *at each level* how the logical and physical organizing principles of the parts determine the function of the higher-level system [3, 4].

Individual researchers usually work on manageable pieces of the larger systems biology puzzle. Some focus on logical relationships and others on physical relationships. Researchers also investigate different levels of the systems biology hierarchy. For example, some investigate the function of individual genes or proteins. Others investigate the pathways and networks that characterize the interactions between pairs of molecular components or focus on the physical molecular assemblies that perform most biological functions. Still other researchers study higher-level structures such as whole cells, biofilms, tissues, and organs. Each of these endeavors tends to have a different nomenclature, use different tools, and confront different problems, which are summarized in **Table 1**.

In the next few paragraphs, we briefly discuss these hierarchical issues and associated challenges. We then discuss issues facing the field of systems biology as a whole.

Genes or proteins can have multiple names within a single organism and additional names when different organisms are being considered because, over the years, these genes and proteins have been discovered and rediscovered in different contexts and different organisms. This proliferation of names causes confusion and complicates the tasks of automated analysis and manual data curation. Existing high-throughput data may be revised as errors in the data are found and corrected. Revision of the underlying sequence data creates problems of “referential integrity” for the higher-level models, descriptions, and simulations that attempt to refer consistently and accurately to their component parts. Referential integrity is an example of a key systems biology challenge, listed in the rightmost column of Table 1. These difficulties highlight the value of careful human curation and the need for commonly accepted ontologies, which may be thought of as agreed-upon categorizations or naming conventions [5].

Biologists and bioinformaticians analyze genomic and proteomic sequence data with a wide variety of computing tools such as BLAST for DNA or protein sequence matching, ClustalW for multiple DNA or protein sequence alignment, GeneScan for identifying gene features such as exons and splice sites, and HMMER for deducing hidden Markov models underlying amino acid sequences in families of related proteins. Tools such as PREDATOR, NNPREDICT, and Jpred attempt to predict secondary protein structure. More than a hundred research groups are working to develop and improve techniques for protein folding from amino-acid sequence data, techniques that are more precisely known as *in silico* tertiary structure prediction. All of these tools and techniques are compute-intensive. For example, BLAST and its variants consume an increasingly large number of compute cycles at universities and research institutes. When *in silico* protein folding becomes more reliable, it promises to consume even more compute power.

Systems biology researchers are often concerned with the ways in which biomolecules and portions of those biomolecules participate in pairwise interactions with one another. (Some of these biomolecules are listed in the first row and column of Table 1.) For example, two proteins may bind to one another, or one protein may modify the behavior of another protein by attaching a phosphate to it. Genes interact with one another as well. The product of the expression of one gene may bind to DNA to enhance or suppress the expression of the same or

another gene. Biologists, armed with various kinds of pairwise interaction data about genes and proteins, create genetic control-network models and protein-pathway models. These network models can illuminate patterns in the interactions between parts and thereby uncover biological mechanisms [6]. However, it can be difficult to interpret network models, especially large models such as the genetic interaction network of *C. elegans* (a small roundworm), which comprises more than eighteen thousand interactions [7]. Not all of the putative or possible interactions in such a large model are biologically significant in the living cell. The networks are often incomplete, and second- or third-hand effects, such as those that involve one or more intermediary nodes, may be as important for biological function as the direct interaction represented by a single arc in the network. In addition, researchers who study the dynamics of network models typically lack time-series data to guide their efforts. Thus, the temporal properties of these systems are often poorly understood.

Note that proteins do not associate in only a pairwise manner. Some researchers believe that nearly every important function in the cell is carried out by a physical assembly (or complex) of ten or more proteins [8]. For example, the Eukaryotic Polymerase II (Pol-II) pre-initiation complex, the molecular “machine” that attaches to the initiation point of a gene to begin the transcription of DNA into RNA, is made up of at least a dozen different protein components [9]. The spliceosome, a nuclear complex or molecular machine that processes the RNA transcript in order to excise introns and splice the exons together, is a much larger assembly composed of as many as 300 distinct proteins and five RNAs [10]. As research in pathways and assemblies advances in parallel efforts, one challenge will be to determine how highly interconnected clusters of nodes in network models [11] relate to such physical assemblies. We also need to understand precisely how the assemblies function both in isolation and in concert with other assemblies. This will involve detailed dynamic 3D models that require substantial computing resources.

At an even higher level of abstraction, biologists model systems of subsystems (bottom row of Table 1). The MCell simulator program [12], for example, models the microphysiology of neuromuscular synaptic junctions in terms of membranes, synaptic vesicles, neurotransmitter molecules, receptors, and ion channels. Virtual cell models attempt to represent a cell by relating underlying subsystems such as pathways, complexes, organelles, membranes, and sometimes the cytoskeleton. Examples include 1) the erythrocyte (red blood cell) model developed in 1988 by Bernhard Palsson [13]; 2) the E-Cell initiative, an international research project aimed at

developing necessary theoretical supports, technologies, and software platforms to allow precise whole cell simulation [14]; and 3) the CyberCell project [15, 16], which is affiliated with the International E. coli Alliance and is concerned, in part, with detailed mathematical models to simulate all or part of a bacterial cell at nanoscopic ( $10^{-9}$  m), mesoscopic ( $10^{-8}$  m), and microscopic ( $10^{-6}$  m) levels [17]. Several biotechnology companies, such as Genomatica and Gene Network Sciences, have also produced cell models. Virtual tissue and organ models, also in the last column of Table 1, include angiogenesis models [18], and models for diabetes, obesity, and asthma produced by Entelos, Inc. [19]. Research at the level of systems of systems demands complex models and simulations, which can often be challenging to validate experimentally. Accurately understanding system behavior, from known or presumed behavior of parts, requires an iterative research cycle. For example, the initial simulations generate predictions that require experimental validation. One or more unfulfilled predictions will often require revisions to the simulation and another cycle of research. Finally, many of these simulations will eventually have to deal explicitly with the 3D realities of the underlying physical mechanisms. For example, pathways for fatty-acid synthesis (which are targets for both anti-obesity and anti-cancer drugs in humans) are very similar in bacteria, fungi, and mammals. However, the underlying physical structures for these pathways are quite different [20]. In bacteria such as *E. coli*, the individual enzymes, represented by nodes in the pathway model, are separate, freestanding protein molecules that diffuse independently in the cell cytoplasm. In fungi and mammals, the individual enzymes are assembled into large physical complexes that function as efficient molecular assembly lines [21, 22]. Interestingly, the physical architecture of the mammalian complex differs dramatically from that of the fungi complex. The mammalian architecture features parts that function like hinges to help orchestrate the synthesis, whereas the fungi complex does not seem to have any moving parts. These 3D structural differences may produce very different dynamics in fatty acid synthesis.

### General issues facing systems biology

The long-term benefits of systems biology will become manifest in the form of advances in biotechnology, medicine, and pharmaceuticals. Better crops, fermentation processes, methods of drug discovery, and personalized medicine, all of which are based on systems models of cells and organs, are the most obvious possible benefits to come from research in systems biology. However, the benefits will not come easily or quickly. Today's simple models only give hints as to their

potential. Many challenges must be overcome before this potential is fully realized.

Systems biology is not simply an extension of high-throughput biology. Systems biology is a synthetic endeavor rather than a reductionist one, because it relies heavily on capturing the “big picture” with respect to the wealth of data now available, and embodying those ideas in models and simulations. These models eventually must relate to the full range of complexity of living systems. Some models must deal explicitly with detailed spatial and temporal relationships between the elements of a system so that interactions can be faithful to the effects of proximity in space and time. For example, both calcium ion signals and protein kinases are notoriously promiscuous (i.e., nonspecific); their spatial distribution in the cell, rather than their ability to bind to specific receptors, determines their function [23, 24]. Models will sometimes require stochastic components as well in order to realistically capture system behaviors [25].

Numerous systems biology questions come to mind. How do we deal with such issues when we do not know all the details of the structures? Which details are crucial, and which details might be ignored? How do we sift through and organize all the data we have? How do we choose the “right,” or appropriate, model? When multiple researchers such as biologists, mathematicians, and computer scientists are involved, how should these scientists reach a consensus about model choice? How do we balance the needs for model *testability* (which motivates the creation of simple models) and the need to account for the *richness* of the real biological systems (which argues for making models more complex)? Some of these issues are addressed in the following sections, but obviously the answers to these questions vary according to the specific problems being studied.

### Obtaining appropriate data

The apparent glut of genomic and proteomic data may give the impression that much more data exists than is needed. However, interviews with scientists suggest that data generated by experimentation is likely to be the limiting factor in systems biology research. Systems biology modelers we interviewed repeatedly told us that critical kinds of data are missing. The iterative interplay of modeling and simulation with the related experimentation is often most useful when the models make nonintuitive predictions. Verification or rejection of unexpected or nonintuitive predictions may require not only new experimental data, but also new *kinds* of experiments. Carrying out these experiments can take far longer than adding or changing features of a computer model. As a particular challenge, such kinds of data and kinds of new experiments may not be interesting to more

traditional biologists or even publishable, on their own, in more traditional journals. Thus, these kinds of data may have to be gathered as integral parts of interdisciplinary systems biology programs.

### **Managing, organizing, and searching network data**

New DNA sequencing techniques will be much faster than current methods in common use [26], and similar advances are also accelerating the collection of RNA transcript and proteomic data. When such techniques become widespread, much more data will demand attention. For example, personalized medicine may help physicians guide treatment specific to an individual when the genomes of individual people are sequenced.

Systems biology benefits from flexible and rapid query of many disparate and differently organized databases. Database challenges include data integration problems, which will require solutions to the naming, referential integrity, and ontology problems. These problems are the subject of several standards efforts, for example, BioPax [27], a collaborative effort to create a data exchange format for biological pathway data.

Analysis of the raw data generates derived data such as DNA annotations, sequence alignment information, putative participation in various pathways, and structure predictions that will require significant storage. For example, comparative BLAST searches of whole genomes generate large files of results. We are nearing a time when computational power may allow “all-by-all” sequence comparisons (i.e., all genomes are compared to all other genomes). Results of all-by-all searches will be quite large, and so costly to generate that they will have to be stored. Once stored, search results will themselves become the object of data mining and further analysis.

A considerable portion of the flood of new data takes the form of networks or tree structures. For example, metabolic and gene regulation pathways are inherently networks. Annotation data is often in the form of networks. 3D structures can often be described as a hierarchy of containment and attachment; for example, a molecular complex such as a ribosome may be attached to the membrane of a “container” organelle such as the endoplasmic reticulum that in turn may connect to the nuclear membrane. One common perspective of systems biology argues that networks are the fundamental organizing principle of cellular function [6, 28]. Unfortunately, the commercial databases that can handle large volumes of data in a robust and reliable way are relational databases that are not well suited to storing network structures. Of course, one can represent any kind of network model in a relational structure by storing relations between nodes and arcs, from which the network can be recreated. However, this kind of

relational representation does not provide much support for directly searching, navigating, or modifying a network. A separate application must usually be provided to reconstruct the network, in computer memory, from the information in the database, and provide search, traversal, and modification functions. Appropriate changes must then be written back into the database. Therefore, a number of leaders in computational systems biology argue that they need a commercial-quality database that can directly provide the needed search, traversal, and modification functions [29]. Lacking what is really needed, some researchers use homemade (custom) databases or the rudimentary object/relational capabilities of commercial databases. Others use relational databases with front-end servers to rapidly reconstruct networks. None of the robust commercial databases are optimal.

### **Choosing the appropriate kinds of models and simulation techniques**

Modeling complex systems continues to be as much an art as a science [6]. Part of that art is in recognizing that different systems and subsystems may require very different kinds of models. For example, C. Rao and A. Arkin distinguish at least five basic kinds of cellular simulation models [30]:

- *Metabolic models* are characterized by conservation of mass and mass action, whereby the rate of a chemical reaction is proportional to the quantity of the reacting substances. These kinds of models describe networks of basic metabolic biomolecules and the enzymatic reactions that create them and/or break them apart. These reactions and their associated reaction rates are described by differential equations.
- *Bifurcation models* (genetic switch models) represent systems that probabilistically and irretrievably assume one of two possible paths. These models require modeling of positive feedback biological “circuits” that magnify an initial small bias toward one outcome or the other so as to produce a permanent strong bias [31].
- *Multistage growth models* characterize systems that are analogous to an assembly line—for example, a system involved in the production of many copies of the T7 bacteriophage in an infected *E. coli* bacterium [32].
- *Cell cycle models*, such as cell-division (mitosis) or cell-death (apoptosis [33]) models, must be able to mimic the careful cellular control of transitions from one stage of the process to another via biochemical checkpoints that prevent progress to the next stage

until all necessary steps in the prior stage are complete.

- *Signal transduction models* are characterized by networks through which information flows. The information may be communicated by cascades of small, temporary changes to the state of various biomolecules—for example, a change made by attaching a phosphate to a particular amino acid residue in a protein molecule. To the extent that signal transduction models involve very small numbers of molecules, the models may require stochastic treatment.

While one could devise other categorizations of models, it is clear that one type of model does not fit all circumstances. However, most systems biology efforts, including the well-known virtual cell programs, focus their attention on just one or two of these kinds of models because the researchers building a particular model typically seek to explain just one or two sorts of biological function. Building a complete model of a cell, especially one that can predict both the beneficial and the deleterious side effects of a new drug, will require dealing with multiple cellular simulation models at once.

#### ***Computational demands of modeling and simulation***

The computational demands of the many kinds of models and simulations vary widely. Some models run on personal computers, some on workstations, some on Linux\*\* clusters, and some on supercomputers. Models that attempt to simulate physical structures, especially at multiple temporal and spatial scales, tend to require much greater computational resources than models that analyze logical structures such as pathways. For example, protein folding, i.e., tertiary structure prediction, can seemingly consume as much compute power as a researcher can devote to it. Reverse-engineering of pathways [34] and Monte Carlo simulation of stochastic processes are also notorious for consuming great amounts of compute power.

Whatever the current compute demands for simulation and modeling by the researchers interviewed for this report, they shared a belief that their future research will require far more compute power. Most foresaw a need for at least Linux cluster support. Their view of the required size of the cluster appeared to correlate with their current consumption of computational power. Those researchers currently using personal computers envisioned a need for relatively small clusters (perhaps 8 to 16 nodes). Those already doing more ambitious computing tended to see a need for clusters with several hundred nodes. A few groups mentioned 1,000-node clusters, and one group, which used a 96-node cluster for reverse-engineering large

pathway models, suggested that they would like a 10,000-node cluster. To some degree, these differing views of computational requirements reflect the old adage that computing needs expand to fill the available supply. However, in general, understanding the complexity of living systems will require all of the compute power that researchers can obtain for many years to come.

#### ***Simulation frameworks***

Several efforts are underway to build either open-source or proprietary general software frameworks for systems biology simulation. For example, the Japanese E-Cell project, with contributions from the California Institute of Technology and the University of Hertfordshire, UK, provides an open-source simulation framework called “The Systems Biology Workbench” [35], which uses differential-equation models and some simple stochastic models. Another open-source framework effort is Bio/Spice [36], a biological data analysis and modeling workspace that is based loosely on SPICE tools used by electrical engineers for circuit analysis and modeling. The Virtual Cell framework [37], which is a Java\*\* framework from The Center for Cell Analysis and Modeling (CCAM) at the University of Connecticut, is intended to be useful for a wide variety of modeling efforts. Several relatively new companies also provide proprietary simulation frameworks. Physiomics, a UK company based in Oxford Science Park, markets a simulation system called SystemCell\*\* that can model systems such as the EGF (epidermal growth factor) signaling pathway or the Ras control circuit, which is involved with cancer growth. Entelos, an American biosimulation company, markets another proprietary system called PhysioLab\*\*, technology for simulating metabolic disease processes such as diabetes. Other examples of companies involved in the creation of simulation frameworks include the American companies Genomatica and Gene Network Sciences. Genomatica markets SimPheny\*\*, a client-server application that enables the development of predictive computer models of organisms, from bacteria to humans. Gene Network Sciences markets VisualCell\*\*, a data integration platform and drawing software toolkit that enables large-scale cellular modeling.

Simulation frameworks are most useful when they address the general parts of the problem and leave the specializations needed for a given problem to the researcher. However, enough experience may not yet have accumulated in simulating biological processes to judge what a framework should provide. Frameworks created for a single laboratory or research effort are valuable for those researchers within that lab who share assumptions, data structures, and research aims. However, such frameworks may impose constraints on their users that

will not be acceptable to researchers with slightly different aims and assumptions.

### ***Finding people with the appropriate analysis, modeling, and simulation skills***

Computational biology, bioinformatics, and systems biology are all relatively new disciplines. Twenty years ago, most biologists outside the field of biophysics considered computing to be unimportant to biological research. Today, computing plays an important role in almost all areas. Such a shift in focus requires the efforts of multitiered individuals. Some systems biologists are truly interdisciplinary, combining biology expertise with mathematical and/or computational modeling expertise. However, many biologists still lack the math and computer skills needed for the analysis, modeling, and simulation required by systems biology. Conversely, few people with the necessary math and computing skills have a good understanding of cellular and molecular biology or the experimental skills needed to gather the data necessary to support and verify models.

One response to this shortage is the formation of new multidisciplinary university undergraduate and graduate programs, typically associated with new research programs. Examples include MIT's Computational and Systems Biology Initiative (CSBi), Harvard University's Department of Systems Biology, the University of Ottawa's Institute of Systems Biology, Rutgers University's BioMaPS Institute for Quantitative Biology, Stanford University's Bio-X Program, Amsterdam's Institute for Systems Biology, Keio University's Institute for Advanced Biosciences in Tsuruoka, Japan, and the Swiss Federal Institute of Technology's Institute for Molecular Systems Biology in Zurich. Despite these and other excellent university programs, demand for interdisciplinary systems biology researchers and programs nevertheless continues to outstrip supply.

Until the skills shortage abates, systems biology efforts will continue to attract new people from other scientific disciplines who have the desired mathematical, modeling, or simulation skills but lack deep knowledge of biology. These "immigrants," or newcomers to systems biology research, will bring with them preconceived biases toward particular simulation methodologies, software development, hardware vendors, operating systems, programming languages, and tools. The biologists who rely on these newcomers often lack the knowledge needed to evaluate the modeling and simulation approaches recommended to them. As a result, the modeling methods familiar to the newcomers may improperly influence the choice of modeling approaches in the early stages of research.

Immigration to the field of systems biology also affects the character and goals of both research and commercial efforts. Groups dominated by biologists see the world

rather differently from groups dominated by experts in computing, engineering, chemistry, mathematics, or physics. Biologists show more humility in the face of the complexity of living systems. Biologists tend to use modeling and simulation to understand interactions and dynamics in systems that they have studied by *other* means. In contrast, non-biologists tend to use computing to organize and sift through large volumes of data, such as high-throughput data, in the hope of discovering hitherto unrecognized patterns and function. Perhaps a useful analogy is to think of the biologists as tending to prefer a target rifle when focusing on a biological problem, whereas the non-biologists may prefer a very broad shotgun approach. Both of these approaches can provide valuable results.

### ***Facing the limits of Occam's Razor***

Since the 14th century, science and scientists have relied on Occam's Razor to guide theories and their models. When choosing between two alternative models, theorists are admonished to pick the one that requires the fewest assumptions. Occam's Razor may indeed work well in the physical sciences, where complex phenomena result from a relatively few underlying laws. Biology, however, is a science that deals with a complex evolutionary history, in the sense that biology focuses on mechanisms that reflect the consequences of 3.5 billion years of evolution. Each successive "advance" in the function and complexity of living systems must coexist and survive in competition with preexisting mechanisms. Thus, biological systems are a triumph of layer after layer of what software professionals would call "clever hacks." Each new layer exploits the hacks that have come before. To use another programming metaphor, in biology most "bugs" that aren't fatal turn out to be features.

Because of the long evolutionary history underlying biological systems, these systems tend to be complex and "messy" rather than simple and elegant. For example, the outdated "simple and elegant" model of gene function is a model in which one gene codes for one protein which, in turn, has one function. However, life is not so simple. For instance, some viruses use what is called a "-1 programmed frameshift" to produce two proteins in the correct relative proportion from one gene [38, 39]. This biological trick amounts to turning the common "off-by-one" programming error into a powerful biological systems feature—a very clever hack indeed.

Terry Gaasterland, previously of Rockefeller University, asserted in her 2002 keynote talk given at the Intelligent Systems for Molecular Biology (ISMB) conference, "if you can think of [some surprising quirky cellular mechanism], you will find that somewhere the biological machinery does it." For example, the

previously mentioned “two-proteins-for-one-gene” feature of viruses seems almost unremarkable when compared to alternative splicing in eukaryotes (cells with a membrane-enclosed nucleus). As discussed earlier in this paper, eukaryotic genes consist of coding regions called exons interspersed with non-coding regions called introns. In the cell nucleus, after the gene is first transcribed from DNA into RNA, but before the RNA is translated into protein, large RNA–protein complexes called spliceosomes remove the introns and splice the ends of the exons together into one contiguous coding sequence. This splicing is somewhat error-prone or probabilistic [40]. However, nature turns “errors” in splicing into powerful features. Evolution has seized on the fallibility of splicing to provide multiple proteins from a single RNA transcript. By various estimates, 35–60% of human genes generate alternative splice variants [41, 42]. Some human genes produce *hundreds* of alternative splice variants. The functions of the splice variants may be complementary and related, or may even oppose one another. For example, the bullfrog gene for the gonadotropin-releasing hormone receptor (GnRH) encodes a splice variant that acts as a repressor of the receptor itself [43]. The product of one splice variant is a receptor, and that of another variant is a protein that inhibits that same receptor. This “clever hack” in nature provides a control circuit, because changes in those factors that bias the system toward the receptor splice or the repressor splice will control the cell’s sensitivity to the hormone. The opportunities for other surprising complex and unforeseen mechanisms are endless. Such chaotic evolutionary creativity makes a mockery of Occam’s Razor.

From what we have just discussed, it becomes clear that the complexity of biological systems challenges those who want to build biologically relevant models and simulations. Modelers steeped in centuries of scientific tradition, especially newcomers to systems biology from the physical sciences, tend to look first for simple models, not complex ones. However, overly simple models are a liability in biology; many a potential drug target has turned out to be worthless because the pathway in which it participates is more complex and sophisticated than expected [44–46]. This tension between complexity and simplicity suggests at least two important lessons for systems biology modeling:

- The systems biology field would benefit from an *explicit* discussion about principles that should replace or modify Occam’s Razor in biological modeling. Without such explicit discussion, the many newcomers from physical sciences are likely to repeatedly underestimate the complexity needed for useful models. To minimize that risk, the computing

and modeling must always be grounded in the biology.

- Modeling and simulation in the absence of experimental feedback and validation are likely to lead us down blind alleys. This is one reason why the senior researchers in the field unanimously emphasize the need for iterative collaboration between computational and experimental work. Relatively simple models may be used at the start of research, but early apparent success may often mean that a researcher simply has not yet gathered the disconfirming data. Many elaborations are usually needed before the models have much predictive value.

### ***Adding stochastic properties to models***

Not only are biomolecular mechanisms more complex than might have been imagined, they are often more probabilistic as well. Since many important cellular functions are carried out by very small numbers of molecules, the randomness inherent in individual molecular events can become apparent at a cellular level. For example, C. Rao and A. Arkin note that “[genetic] . . . switches are stochastic, underlining the single-molecule nature of the DNA medium in which they are implemented” [30]. Stochastic gene expression has been observed directly in prokaryotic cells [47, 48] and eukaryotic cells [49].

While we may be tempted to assume that such biological randomness is merely a nuisance, evolution exploits it. Stochasticity in gene expression can be essential for many biological processes [49, 50]. For example, researchers writing in *Nature Reviews Genetics* have recently noted that stochasticity “. . . can provide the flexibility needed by cells to adapt to fluctuating environments or respond to sudden stresses, and a mechanism by which population heterogeneity can be established during cellular differentiation and development” [25].

Wherever stochastic properties of systems are manifested, deterministic models may not be able to account for the observed behavior. Instead, simulation techniques such as Monte Carlo methods are needed. Monte Carlo methods can require considerably more compute power because they require many iterations for a given set of parameters. They also require carefully chosen mathematical descriptions of the underlying random processes.

### ***Accounting for 3D structure and location***

Those researchers interested in studying higher-level structures, such as protein assemblies, whole cells, tissues, and organs, generally need to take into account the 3D

structure of the system and its components. For example, heart models generally must explicitly model the anatomy of the heart as well as its electrophysiology. New techniques such as cryoelectron tomography are providing an increasingly detailed look at the internal 3D structure of eukaryotic cells [51]. The case for using 3D models for tissues is even more self-evident, as noted by researchers writing in the *Annals of the New York Academy of Sciences*: "... the complex motions that characterize tissue and organ formation in 3-D space are not found encoded in DNA and cannot be fully recapitulated in [ex vivo] model systems" [52]. The need to model 3D structure adds another level of complexity and computational demand that is not required for studying logical structures.

### **Modular biology**

In the mid-1980s, when high-throughput biology originated, individual genes and proteins were thought to be the primary functional elements in the cell. Genomics and proteomics involved searches for those functional elements. In the past decade, as we have already discussed, researchers have come to realize that most protein machinery is composed of multi-molecule assemblies, also sometimes called complexes or modules, that act as molecular machines to carry out biological function. For example, researchers have identified more than one hundred protein assemblies in baker's yeast that range in size from two to 83 protein molecules. Many of these assemblies have close analogs in humans [8]. As Leland Hartwell and his colleagues at the Fred Hutchinson Cancer Center have noted, we are seeing a shift from molecular biology to modular biology [53].

Assemblies are important in cellular function because they are exquisitely structured to provide many avenues of control and to minimize crosstalk between unrelated pathways and mechanisms. For example, three-dimensional assemblies of scaffold proteins physically organize the proteins in certain signaling pathways. In the yeast mating pathway, "... scaffolds not only direct basic pathway connectivity but can precisely tune quantitative pathway input-output properties" [24]. Each part of an assembly has a precise location in the assembly, and assemblies themselves tend to be located in specific regions of the cell and in specific arrangements. Therefore, many cell models, such as cell-signaling models that ignore location, distance, and three-dimensional structure, will eventually have to be replaced by models that explicitly account for the structure and location of the assemblies.

At this time, cell models that take location into account are rare and simplified. For example, they cannot take into account large-scale cytoskeletal structures. Even so, the models require substantial compute power to run.

As such models become more common and more sophisticated, higher-performance computing hardware will be required to run the resulting simulations.

### **Self-assembly**

Many models will also have to take into account the "life cycle" of the physical structures—that is, how the structures are created and when and how they are destroyed. Biological systems are not constructed on assembly lines by external agents that put together parts found in a giant catalog according to some blueprint. The parts and the systems assemble themselves as if in a perpetual dance, according to choreography that has emerged over long evolutionary trial and error. Thus, the models we synthesize to explain living systems must not only permit but also explain the self-assembly of the many emergent systems. Although all of the important protein complexes presumably result from a process of self-assembly, explicit models of self-assembling structures, such as microtubules [54], are still rare.

Structural models may also be required to take into account the deconstruction or disassembly of structures. Many cellular structures have relatively short lifetimes. They self-assemble when needed and are disassembled by various mechanisms when no longer needed. Any systems biology research that focuses on modeling the assembly process, with no attention given to the disassembly, may miss significant opportunities to understand important biological and medically relevant phenomena.

### **Assessment of future trends in computational systems biology**

Professor Fotis Kafatos, who at the time of our interview was Director General of The European Molecular Biology Laboratory (EMBL), asserted that biology will become the single largest scientific consumer of computing. If he is correct—and many of the other experts interviewed for this paper agree with him—the "center of gravity" of scientific computing activity will shift from the physical sciences to computational biology. As discussed, the amount of raw information in biological systems is immense. Each biological species has unique DNA and proteins, and, in the case of sexually reproducing species, every individual has a unique genome. Manipulating that information has far-ranging potential economic value. The growth of personalized medicine is one opportunity, and others include agricultural and animal husbandry applications, and harnessing the incredible abilities of single-cell organisms to act as chemical micro-factories. To exploit those opportunities, the vast jungle of biological information must be tamed by computational biology.

As we have discussed above, different approaches to systems biology face different challenges, offer different

benefits, and are therefore likely to take very different paths to success.

### **The workhorse—automated annotation**

Biological databases are rapidly being filled with new DNA sequences and microarray expression data. In order to understand this new sequence data, the first step is to deduce as much as possible about the structure or function of noteworthy subsequences. This task is called annotation. Some annotation is done by manual human effort. However, the volume of new data motivates the need for automatic annotation [55, 56]. For example, various bioinformatics techniques are deployed to

- Search the new sequences for similarity with known genes or proteins in other species.
- Predict coding regions and exon/intron boundaries and derive the amino-acid sequences that would be produced if these putative coding regions were to be expressed as proteins.
- Search for potential splice sites.
- Classify the possible functions of derived amino-acid sequences in putatively expressed genes according to predicted secondary structures.
- Search for patterns of amino-acid sequences that suggest a likely function.
- Search for patterns of amino acids that indicate the likely cellular location in which the gene product is used.

Reliable annotation, properly interpreted, can provide valuable insights into biological function, evolutionary relatedness, and other useful relationships. Given the usefulness of annotation, researchers continue to improve existing techniques and to develop new techniques for automated annotation.

### **Pathway models**

Many pathway models have been deduced from decades of experimental work and, more recently, deduced from high-throughput interaction data. One example pathway is the “Wnt signaling pathway” that mediates many cell-development functions in a wide variety of organisms, including vertebrates, and is also implicated in the cancer process [57]. Pathway models will eventually be validated using wet-lab experiments and time-series data, and they will take into account stochastic issues, as does the research of Roger Brent and his team at The Molecular Sciences Institute in simulating the yeast mating pheromone (G-protein receptor) pathway [58]. Over the next few years, pathway models will become more sophisticated and useful as high-throughput biologists provide more and higher-quality data. The previously

mentioned systems biology companies, such as Genomatica, Physiomics, Gene Network Sciences, and Entelos, are working with pharmaceutical companies to investigate the value of quantitative pathway models for drug discovery. Moreover, scientific consortia, such as the BioPathways Consortium [59] and the BioPax standards group [27], help researchers to share techniques for fostering progress in computational pathway models.

Pathway models can help organize large bodies of high-throughput data, such as DNA transcription array data. Such models may also be used to better understand precise functional relationships between parts of relatively small pathways that are deduced from experiments. Although researchers involved in the study of both large and small networks can share data representation standards and modeling techniques, the role of modeling and simulation is quite different in the two approaches. For example, relatively small pathways, containing perhaps a few dozen nodes (e.g., proteins and/or genes), can now be modeled and simulated with some precision [58]. More significantly, small pathways are more accessible to experiment, which means that the iteration of simulation and experiment can progress at a reasonable rate. In contrast, the larger networks with hundreds of nodes, such as those deduced from analysis of high-throughput data, are not as amenable to simulation simply because not enough is known about the dynamics of the interactions between nodes. However, computer visualization tools can help researchers gain insights into the processes represented by these larger networks [60].

These caveats aside, pathway models are expected to result in improved drug discovery, more hearty and resistant crops, and insights into the management of such pests as insects, nematodes, fungus, and weeds. Some of these advances may be expected in the next five years.

### **Virtual cell models**

Current work on virtual cell models is considerably less mature than work on pathway models. Those researchers who would build a virtual cell face daunting problems. Most of the recent efforts focus on the relatively “simple” *E. coli* model. This bacterium has “only” 4,000 open reading frames (i.e., DNA presumed to give rise to proteins); it has essentially no post-translational protein modification, and its genes are unitary (i.e., they are not divided into introns and exons). Moreover, since it lacks most kinds of organelles, most of its metabolism takes place in a less structured environment than is found in eukaryotes. Other systems biology efforts focus on yeast, because yeast is by far the best-studied eukaryote model.

Roughly two dozen projects exist worldwide to build virtual cell models. The European Bioinformatics Institute maintains a database of quantitative kinetic

models of biochemical and cellular systems [61]. Most cell models are in the preliminary stages. Currently, it would be fair to say that virtual cell models do not actually deal with any whole cell. At best they represent two or three aspects of the cell, such as metabolic processing, which is expressed in terms of differential equations together with some stochastic aspects. As cell models expand their coverage, they will be overwhelmed by complexity, especially as they attempt to represent the multiscale 3D structure of protein complexes, cell membranes, cytoskeletal elements, and organelles, and the multiscale temporal properties ranging from protein-protein interactions to mitosis. The effort will be worthwhile because virtual cell models that do take into account physical pathway scaffolds, physical compartments, and temporal structures will be able to more accurately account for cell behavior. No doubt many of those models will not initially be useful, but some will offer genuine advantages and will become commonly used, though this will probably not happen for at least five to ten years.

### ***Virtual tissue or organ models***

At this time, it appears that virtual tissue models do not have to account for much of the cellular complexity within the tissues they model. Thus, the complexity barrier for systems biology modeling may be substantially lower for well-chosen tissue or organ models than for cell models. Nonetheless, multiscale temporal and spatial issues must still be considered, confronted, and solved as needed for such systems. When mature, virtual tissue models, such as tissue angiogenesis models [18, 52, 62] or heart models [63], may produce clinically useful results.

### **Conclusions**

While traditional biology will certainly continue to make dramatic discoveries, the role of computational biology in general, and systems biology in particular, will continue to grow rapidly as it has for the past several years. Systems biology papers are becoming commonplace. *Science* magazine declared recently that “Systems Biology Signals Its Arrival” [64]. Additionally, new systems biology programs proliferate in academe. Despite this rapid growth, however, computational systems biology is still in its infancy. This paper has attempted to outline some of the challenges that must be overcome. The authors are confident that they *will* be overcome in the coming decades.

Today’s efforts to find commercial or clinical applications of systems biology must surmount technical problems and also generate marketable results. This combined challenge is so difficult that many early efforts are likely to suffer the usual fate of most pioneers. For example, commercial groups are building systems biology

programs with the hope that pathway and virtual cell models will quickly lead to more effective drug discovery. In the long term, this promise will no doubt be fulfilled. However, early systems biology models will necessarily be too simple. As discussed, these models will be missing important elements, dynamics, and interactions, and they may include extraneous elements and interactions. Hence, results of early models will tend to be misleading, perhaps expensively so. In short, we can expect the familiar “Hype Curve” from Gartner, Inc. [65] to apply to the technologies of systems biology. We will see too much early optimism largely based on hype, followed by subsequent disappointment (which Gartner calls the “trough of disillusionment”) and finally by more realistic expectations. This progression often occurs with the introduction of new technologies. Once realistic expectations for a technology are set, the early disillusionment will be followed by a slow realization of the technology’s actual value. We believe that the same process will occur with systems biology. Opinions may differ on whether the peak of the initial hype curve is past or is yet to come. Either way, we should be prepared to persevere during the inevitable trough of disillusionment. Those who do so will likely reap substantial benefits.

### **Appendix**

The interviews and discussions that led to this assessment took place between July 2002 and January 2005. We are grateful to the following individuals for their patience and insights:

- BG Medicine (formerly Beyond Genomics)—Dr. Eric Neumann, Vice President Bioinformatics. Dr. Neumann is now at Sanofi-Aventis Pharmaceuticals.
- Cambridge University—Prof. Dennis Bray, a leader in simulating *E. coli* chemotaxis signaling pathways.
- Deutsches Krebsforschungszentrum Heidelberg (DKFZ) (German Cancer Research Institute)—Dr. Roland Eils, Director of Bioinformatics. He is also Professor of Bioinformatics at the University of Heidelberg.
- The European Molecular Biology Laboratory (EMBL), Heidelberg—Prof. Dr. Fotis Kafatos, former Director General of EMBL; Dr. Luis Serrano, in charge of bioinformatics modeling; and Dr. Christian Boulton, group leader, Scientific Core Facilities, and other colleagues.
- Fraunhofer Institute, Stuttgart—Prof. Dr. Herwig Brunner, Director, and two of his staff.
- GBF (German Biotechnology Research Center), Braunschweig—Prof. Dr. Rudi Balling, Director.
- Gene Network Sciences—Colin Hill, CEO; Dr. Iya Khalil, VP of Research and Development; Dr. Robert

- Miller, and other colleagues. GNS develops virtual cell network models for organisms such as *E. coli*, and tools for pathway/network modeling.
- Genomatica (a systems biology company based near the University of California, San Diego)—Dr. Bernhard Palsson, Co-Founder, Chairman, and Professor of Bioengineering at the University of California, San Diego, and Dr. Christophe Schilling, Chief Technology Officer and co-founder. Bernhard Palsson was one of the creators of the first “virtual cell” model of a human erythrocyte in 1988.
  - Geospiza (a company providing data management for pharmaceutical companies)—Dr. Todd Smith, President.
  - GlaxoSmithKline (GSK)—Dr. Igor Goryanin, then Head of Cell Simulations and Pathway Modeling, GSK Medicines Research Centre, UK. Dr. Goryanin works on *E. coli* models and simulations. He is now Professor and Chair, Computational Systems Biology and Director, Edinburgh Center for Bioinformatics, The University of Edinburgh, Scotland.
  - Indiana University—Dr. Craig Stewart, Director of Research and Academic Computing.
  - Institute for Systems Biology, Seattle—Dr. Leroy Hood and colleagues.
  - Massachusetts Institute of Technology, Department of Biology—Dr. Peter Sorger and colleagues.
  - Merck Research Laboratories—Dr. Jeff Saltzman, Senior Director of Applied Computer Sciences and Mathematics; Dr. Jeff Sachs, Senior Research Fellow, Applied Computer Science and Mathematics; Dr. John Thompson, expert on molecular profiling; Dr. Alex Elbrecht, expert on bioinformatics, and colleagues.
  - The Molecular Sciences Institute—Dr. Roger Brent, Director and President, and colleagues.
  - Oxford University—Professor David Fell, a leader in modeling metabolic pathways.
  - Systems Biology Markup Language (SBML) community leaders—Dr. Michael Hucka, California Institute of Technology, and Dr. Andrew Finney, University of Hertfordshire, UK.
  - San Diego Supercomputing Center—Dr. Shankar Subramanian, head of the bioinformatics effort for the Alliance for Cell Signaling.
  - University of Alberta, Canada—Prof. Michael Ellison, Department of Biochemistry and Executive Director of the Institute for Biomolecular Design (Project CyberCell).
  - University of Auckland—Dr. Peter Hunter, Academic/Institute Director, Bioengineering Institute.

- University of California, Berkeley, Department of Bioengineering, and Lawrence Berkeley Laboratories—Dr. Teresa Head-Gordon, a leading protein folding researcher; Dr. Adam Arkin, an influential expert in modeling and simulation; and Dr. Denise Wolf, a Senior Research Associate in the Arkin laboratory.
- University of California, San Diego—Dr. Philip Bourne, Professor of Pharmacology, Director, Integrative Bioscience, San Diego Supercomputer Center, President, International Society for Computational Biology.
- University of North Carolina, Chapel Hill—Prof. Robert Bourret, a leader in the biophysics of the *E. coli* chemotaxis signaling pathway.
- University of Virginia, Department of Bioengineering—Dr. Tom Skalak, Chairman of the department, and a leader in tissues modeling and angiogenesis.

\*\*Trademark, service mark, or registered trademark of Linus Torvalds, Sun Microsystems, Inc., Physiomics, Genomatica, Inc., or Gene Network Science in the United States, other countries, or both.

## References

1. J. Wooley and H. Lin, Eds., *Catalyzing Inquiry at the Interface of Computation and Biology*, A National Research Council Publication, The National Academic Press, Washington, DC, 2005.
2. B. Alberts, “The Cell as a Collection of Protein Machines: Preparing the Next Generation of Molecular Biologists,” *Cell* **92**, No. 3, 291–294 (1998).
3. T. Ideker, T. Galitski, and L. Hood, “A New Approach to Decoding Life: Systems Biology,” *Annu. Rev. Genomics Hum. Genet.* **2**, 343–372 (2001).
4. H. Kitano, “Computational Systems Biology,” *Nature* **420**, No. 6912, 206–210 (2002).
5. P. Karp, S. Paley, C. Krieger, and P. Zhang, “An Evidence Ontology for Use in Pathway/Genome Databases,” *Proceedings of the Pacific Symposium on Biocomputing*, 2004, pp. 190–201.
6. S. Bornholdt, “Less Is More in Modeling Large Genetic Networks,” *Science* **310**, No. 5747, 449–451 (2005).
7. W. Zhong and P. Sternberg, “Genome-Wide Prediction of *C. elegans* Genetic Interaction,” *Science* **311**, No. 5766, 1481–1484 (2006).
8. A. Gavin, M. Bosche, R. Krause, P. Grandi, M. Marzioch, A. Bauer, J. Schultz, J. Rick, A. Michon, C. Cruciat, M. Remor, C. Hofert, M. Schelder, M. Brajenovic, H. Ruffner, A. Merino, K. Klein, K. M. Hudak, D. Dickson, T. Rudi, V. Gnau, A. Bauch, S. Bastuck, B. Huhse, C. Leutwein, M. Heurtier, R. Copley, A. Edelmann, E. Querfurth, V. Rybin, G. Drewes, M. Raida, T. Bouwmeester, P. Bork, B. Seraphin, B. Kuster, G. Neubauer, and G. Superti-Furga, “Functional Organization of the Yeast Proteome,” *Nature* **415**, No. 6868, 141–147 (2002).
9. D. B. Nikolov and S. K. Burley, “RNA Polymerase II Transcription Initiation: A Structural View,” *Proc. Natl. Acad. Sci. USA* **94**, 15–22 (1997).
10. T. W. Nilsen, “The Spliceosome: The Most Complex Macromolecular Machine in the Cell?,” *BioEssays* **25**, 1147–1149 (2003).

11. E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, and A.-L. Barabasi, "Hierarchical Organization of Modularity in Metabolic Networks," *Science* **297**, No. 5586, 1551–1555 (2002).
12. J. Coggan, T. Bartol, E. Esquenazi, J. Stiles, S. Lamont, M. Martone, D. Berg, M. Ellisman, and T. Sejnowski, "Evidence for Ectopic Neurotransmission at a Neuronal Synapse," *Science* **309**, No. 5733, 446–451 (2005).
13. A. Joshi and B. O. Palsson, "Metabolic Dynamics in the Human Red Cell. Part I—A Comprehensive Kinetic Model," *J. Theor. Biol.* **141**, No. 4, 515–528 (1989).
14. M. Tomita, K. Hashimoto, K. Takahashi, T. Shimizu, Y. Matsuzaki, F. Miyoshi, K. Saito, S. Tanida, K. Yugi, J. C. Venter, and C. Hutchison, "E-CELL: Software Environment for Whole Cell Simulation," *Bioinformatics* **15**, No. 1, 72–84 (1999).
15. G. Broderick, C. E. R'uaiani, and M. J. Ellison, "A Life-Like Virtual Cell Membrane Using Discrete Automata," *In Silico Biol.* **5**, No. 2, 163–178 (2005).
16. S. Sundararaj, A. Guo, B. Habibi-Nazhad, M. Rouani, P. Stothard, M. Ellison, and D. S. Wishart, "The CyberCell Database (CCDB): A Comprehensive, Self-Updating, Relational Database to Coordinate and Facilitate In Silico Modeling of *Escherichia coli*," *Nucl. Acids Res.* **32** (Database issue), D293–D295 (2004).
17. C. Holden, "Alliance Launched to Model *E. coli*," *Science* **297**, No. 5586, 1459–1460 (2002).
18. A. Anderson and M. Chaplain, "Continuous and Discrete Mathematical Models of Tumor-Induced Angiogenesis," *Math. Biol.* **60**, No. 5, 857–899 (1998).
19. G. S. Mack, "Can Complexity Be Commercialized?" *Nature Biotechnol.* **22**, No. 10, 1223–1229 (2004).
20. S. Smith, "Architectural Options for a Fatty Acid Synthase," *Science* **311**, No. 5765, 1251–1252 (2006).
21. T. Maier, S. Jenni, and N. Ban, "Architecture of Mammalian Fatty Acid Synthase at 4.5 Å Resolution," *Science* **311**, No. 5765, 1258–1262 (2006).
22. S. Jenni, M. Leibundgut, T. Maier, and N. Ban, "Architecture of a Fungal Fatty Acid Synthase at 5 Å Resolution," *Science* **311**, No. 5765, 1263–1267 (2006).
23. C. R. Raymond and S. J. Redman, "Spatial Segregation of Neuronal Calcium Signals Encodes Different Forms of LTP in Rat Hippocampus," *J. Physiol.* **570**, No. 1, 97–111 (2005).
24. R. P. Bhattacharyya, A. Remenyi, M. C. Good, C. J. Bashor, A. M. Falick, and W. A. Lim, "The Ste5 Scaffold Allosterically Modulates Signaling Output of the Yeast Mating Pathway," *Science* **311**, No. 5762, 822–826 (2006).
25. M. Kaern, T. Elston, W. Blake, and J. Collins, "Stochasticity in Gene Expression: From Theories to Phenotypes," *Nature Rev. Genet.* **6**, No. 6, 451–464 (2005).
26. J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, and G. M. Church, "Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome," *Science* **309**, No. 5741, 1728–1732 (2005).
27. "BioPAX: Biological Pathways Exchange"; see [www.biopax.org](http://www.biopax.org).
28. Z. Oltvai and A. Barabasi, "Life's Complexity Pyramid," *Science* **298**, No. 5594, 763–764 (2002).
29. H. V. Jagadish and F. Olken, "Database Management for Life Science Research: Summary Report of the Workshop on Data Management for Molecular and Cell Biology at the National Library of Medicine, Bethesda, Maryland, February 2–3, 2003," *OMICS* **7**, No. 1, 131–137 (2003).
30. C. V. Rao and A. P. Arkin, "Control Motifs for Intracellular Regulatory Networks," *Annu. Rev. Biomed. Eng.* **3**, 391–419 (2001).
31. T. Gardner, C. R. Cantor, and J. J. Collins, "Construction of a Genetic Toggle Switch in *Escherichia coli*," *Nature* **403**, No. 6767, 339–342 (2000).
32. D. Endy, D. Kong, and J. Yin, "Intracellular Kinetics of a Growing Virus: A Genetically Structured Simulation for Bacteriophage T7," *Biotechnol. Bioeng.* **55**, No. 2, 375–389 (1997).
33. K. A. Janes, J. G. Albeck, S. Gaudet, P. K. Sorger, D. A. Lauffenburger, and M. B. Yaffe, "A Systems Model of Signaling Identifies a Molecular Basis Set for Cytokine-Induced Apoptosis," *Science* **310**, No. 5754, 1646–1653 (2005).
34. J. J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* **2**, No. 2, 70–77 (2004).
35. H. Sauro, M. Hucka, A. Finney, C. Wellock, H. Bolouri, J. Doyle, and H. Kitano, "Next Generation Simulation Tools: The Systems Biology Workbench and BioSPICE Integration," *OMICS* **7**, No. 4, 355–372 (2003).
36. S. P. Kumar and J. C. Feidler, "BioSPICE: A Computational Infrastructure for Integrative Biology," *OMICS* **7**, No. 3, 225–226 (2003).
37. B. M. Schaff, B. Slepchenko, and L. M. Loew, "Physiological Modeling with Virtual Cell Framework," *Methods Enzymol.* **321**, No. 1, 1–23 (2000).
38. J. Dinman, T. Icho, and R. Wickner, "A -1 Ribosomal Frameshift in a Double-Stranded RNA Virus of Yeast Forms a Gag-Pol Fusion Protein," *Proc. Natl. Acad. Sci. USA* **88**, No. 1, 174–178 (1991).
39. J. Lopinski, J. Dinman, and J. Bruenn, "Kinetics of Ribosomal Pausing During Programmed -1 Translational Frameshifting," *Molec. Cell. Biol.* **20**, No. 4, 1095–1103 (2000).
40. T. A. Thanaraj and F. Clark, "Human GC–AG Alternative Intron Isoforms with Weak Donor Sites Show Enhanced Consensus at Acceptor Exon Positions," *Nucl. Acids Res.* **29**, No. 12, 2581–2593 (2001).
41. E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, and W. FitzHugh, "Initial Sequencing and Analysis of the Human Genome," *Nature* **409**, No. 6822, 860–921 (2001).
42. B. R. Graveley, "Alternative Splicing: Increasing Diversity in the Proteomic World," *Trends Genet.* **17**, No. 2, 100–107 (2001).
43. L. Wang, D. Y. Oh, J. Bogerd, H. S. Choi, R. S. Ahn, J. Y. Seong, and H. B. Kwon, "Inhibitory Activity of Alternative Splice Variants of the Bullfrog GnRH Receptor-3 on Wild-Type Receptor Signaling," *Endocrinology* **142**, No. 9, 4015–4025 (2001).
44. G. Smith, R. Knowles, C. Pogson, M. Salter, M. Hanlon, and R. Mullin, "Flux Control Coefficients of Glycinamide Ribonucleotide Transformylase for de novo Purine Biosynthesis," *Control of Metabolic Processes*, A. Cornish-Bowden and M. Cardenas, Eds., Plenum Press, New York, 1990, pp. 385–387.
45. M. Costi and S. Ferrari, "Update on Antifolate Drug Targets," *Current Drug Targets* **2**, No. 2, 135–166 (2001).
46. M. Gelb and W. Hol, "Drugs to Combat Tropical Protozoan Parasites," *Science* **297**, No. 5580, 343–344 (2002).
47. H. McAdams and A. Arkin, "Simulation of Prokaryotic Genetic Circuits," *Annu. Rev. Biophys. Biomol. Struct.* **27**, 199–224 (1998).
48. H. McAdams and A. Arkin, "It's a Noisy Business! Genetic Regulation at the Nanomolar Scale," *Trends Genet.* **15**, No. 2, 65–69 (1999).
49. W. J. Blake, M. Kaern, C. R. Cantor, and J. J. Collins, "Noise in Eukaryotic Gene Expression," *Nature* **422**, 633–637 (2003).
50. M. Elowitz, A. Levine, E. Siggia, and P. Swain, "Stochastic Gene Expression in a Single Cell," *Science* **297**, No. 5584, 1183–1186 (2002).
51. O. Medalia, I. Weber, A. Frangakis, D. Nicastro, G. Gerisch, and W. Baumeister, "Macromolecular Architecture in Eukaryotic Cells Visualized by Cryoelectron Tomography," *Science* **298**, No. 5596, 1209–1213 (2002).
52. K. Hirschi, T. Skalak, S. Peirce, and C. Little, "Vascular Assembly in Natural and Engineered Tissues," *Ann. NY Acad. Sci.* **961**, 223–242 (2002).

53. L. Hartwell, J. Hopfield, S. Leibler, and A. Murray, "From Molecular to Modular Cell Biology," *Nature* **402**, No. 6761, C47–C52 (1999).
54. T. Surrey, F. Nedelec, S. Leibler, and E. Karsenti, "Physical Properties Determining Self-Organization of Motors and Microtubules," *Science* **292**, No. 5519, 1167–1171 (2001).
55. T. Kasukawa, M. Furuno, I. Nikaido, H. Bono, D. A. Hume, C. Bult, D. P. Hill, R. Baldarelli, J. Gough, A. Kanapin, H. Matsuda, L. M. Schriml, Y. Hayashizaki, and Y. Okazaki, and J. Quackenbush, "Development and Evaluation of an Automated Annotation Pipeline and cDNA Annotation System," *Genome Res.* **13**, No. 6B, 1542–1551 (2003).
56. J. R. Wortman, B. J. Haas, L. I. Hannick, R. K. Smith, Jr., R. Maiti, C. M. Ronning, A. P. Chan, C. Yu, M. Ayele, C. A. Whitelaw, O. R. White, and C. D. Town, "Annotation of the Arabidopsis Genome," *Plant Physiol.* **132**, 461–468 (2003).
57. R. Nusse, "Making Head or Tail of Dickkopf," *Nature* **411**, 255–256 (2001).
58. A. Colman-Lerner, A. Gordon, E. Serra, T. Chin, O. Resnekov, D. Endy, C. G. Pesce, and R. Brent, "Regulated Cell-to-Cell Variation in a Cell-Fate Decision System," *Nature* **437**, No. 7059, 699–706 (2005).
59. "BioPathways Consortium"; see [www.biopathways.org](http://www.biopathways.org).
60. P. Saraiya, C. North, and K. Duca, "Visualizing Biological Pathways: Requirements Analysis, Systems Evaluation and Research Agenda," *Info. Visualiz.* **4**, No. 3, 191–205 (2005).
61. N. Le Novère, B. Bornstein, A. Broicher, M. Courtot, M. Donizelli, H. Dharuri, L. Li, H. Sauro, M. Schilstra, B. Shapiro, J. L. Snoep, and M. Hucka, "BioModels Database: A Free, Centralized Database of Curated, Published, Quantitative Kinetic Models of Biochemical and Cellular Systems," *Nucl. Acids Res.* **34** (Database issue), D689–D691 (2006).
62. S. M. Peirce, E. J. Van Gieson, and T. C. Skalak, "Multicellular Simulation Predicts Microvascular Patterning and In Silico Tissue Assembly," *FASEB J.* **18**, No. 6, 731–733 (2004).
63. R. Winslow, J. Rice, and M. Jafri, "Modeling the Cellular Basis of Altered Excitation–Contraction Coupling in Heart Failure," *Prog. Biophys. Mol. Biol.* **69**, No. 2–3, 497–514 (1998).
64. The News Staff, "Breakthrough of the Year: The Runners-Up," *Science* **310**, No. 5756, 1884 (2005).
65. Gartner, Inc., "Hype Cycle"; see [www.gartner.com/pages/story.php?id.8795.s.8.jsp](http://www.gartner.com/pages/story.php?id.8795.s.8.jsp).

*Received November 11, 2005; accepted for publication February 3, 2006; Internet publication July 27, 2006*

**Steve Burbeck** ([sburbeck@mindspring.com](mailto:sburbeck@mindspring.com)). Dr. Burbeck recently retired from IBM; he is currently an independent consultant focusing primarily on scientific computing. He received a B.A. degree in mathematics from California State University and a Ph.D. degree in mathematical and cognitive psychology from the University of California at Irvine in 1979. He was Director of Data Processing and Statistics for eight years at the Linus Pauling Institute of Science and Medicine, where he participated in both computational genomic and proteomic research. He worked in the computing industry at two startup companies, at Apple Computer, and finally at IBM from 1995 to 2005. At IBM Dr. Burbeck worked on object-oriented technologies, spent a year in IBM Research helping pioneer the ideas that are now known as autonomic computing, and thereafter worked in various aspects of emerging technologies, including web services, open-source software, peer-to-peer technologies, and, for the last three years, the implications of computational systems biology. He was elected to the IBM Academy of Technology within two years of joining the company. Dr. Burbeck is an author of some two dozen refereed publications plus a similar number of internal IBM confidential publications. His URL is [www.runningempty.org/Steve/HistoryAndInfo.html](http://www.runningempty.org/Steve/HistoryAndInfo.html).

**Kirk E. Jordan** *IBM Deep Computing, One Rogers Street, Cambridge, Massachusetts 02142* ([kjordan@us.ibm.com](mailto:kjordan@us.ibm.com)). Dr. Jordan is Emerging Solutions Executive in the IBM Deep Computing organization within the Systems and Technology Group. He is responsible for overseeing the development of applications for IBM advanced computing architectures; investigating and developing concepts for new areas of growth for IBM, especially in healthcare and life sciences involving high-performance computing; and providing leadership in high-end computing and simulation in such areas as systems biology and high-end visualization. Dr. Jordan received a B.S. degree in mathematics from Hobart College, and M.S. and Ph.D. degrees in applied mathematics from the University of Delaware in 1980. Prior to joining IBM, he held positions at Exxon Research and Engineering, Thinking Machines, and Kendall Square Research. He is currently Vice President for Industry in the Society for Industrial and Applied Mathematics (SIAM). Dr. Jordan's main research interests are in the efficient use of advanced architecture computers for modeling and simulation. He is an author of more than 35 papers on subjects that include interactive visualization using parallel computers, parallel domain decomposition for reservoir/groundwater simulation, turbulent convection flows, parallel spectral methods, multigrid techniques, multi-resolution wavelets, and wave propagation.