Notes on the Visions for Computational and Systems Biology Workshop for the Genomes to Life Program[1]

U.S. Department of Energy
Washington, D.C.
September 6–7, 2001

Introduction

On September 6–7, 2001, the U.S. Department of Energy's (DOE) Genomes to Life program (DOEGenomesToLife.org) brought together 120 biologists and computational scientists in Washington, D.C., for a workshop entitled "Visions for Computational and Systems Biology." The clear conclusion of participants was that we are on the threshold of an exciting new era in which the biological and information sciences will combine forces to solve critical problems facing the environment, energy production, and human health. From the foundation of whole-genome sequences, the aspiration of the new biology is to build a new, comprehensive, and profound understanding of complex living systems. A central goal of Genomes to Life is to establish a national infrastructure to transform the tremendous outpouring of data and concepts into a new computationally based biology. Success in this quest will require powerful new biological, mathematical, computational, engineering, and physical concepts, approaches, and technologies such as modeling, as well as the capabilities of other federal agencies. With this meeting, Genomes to Life took a first step by starting to create a common language and set of goals across the many scientific disciplines and agencies that must work together to achieve the vision. This objective can be achieved only by joining revolutionary technologies for systems-level and computational biology.

The workshop's central theme was that the current paradigm in biology—variously described as "single gene," "reductionist," or "linear"—is not likely to be successful on its own in providing the necessary data and understanding to permit quantitative predictions or de novo design of biological systems. Instead, the existing research approaches will be augmented by a "systems" approach in which comprehensive data sets will be collected and assembled into predictive computational models. The new paradigm grows out of the rapid advances in instrumentation for the biosciences, the vast improvements in computing speeds and modeling capabilities, the growing interest from physical and information scientists in biological problems, and the recognition that new approaches are needed for biology to achieve its full promise in improving human well-being. This report summarizes the key findings from this workshop. It describes long-term goals and major scientific drivers behind computational and systems biology, as well as discussions related to overcoming the existing barriers in biosciences research.

[1]These are the best available notes and do not represent a verbatim or consensus document of the workshop. Remarks made by individuals are believed to be correct but have not been verified.

Scientific Drivers for Computational and Systems Biology

The ultimate goal of every science is to achieve such a complete understanding of a phenomenon that a set of mathematical laws or models can be developed to accurately predict all relevant properties of the phenomenon. Such a model can then form the foundation for understanding more complex systems and can be applied to useful ends, such as developing more energy efficient cars, reducing pollution, detecting biowarfare agents, or developing new therapeutic drugs. Although such predictive capabilities now exist for certain areas of physics, chemistry, and engineering, virtually no biological systems are understood at this level of quantitative accuracy. Nevertheless, a major conclusion from this workshop is that the biosciences are poised for very rapid progress towards becoming a quantitative and predictive science. The proximity of revolutionary breakthroughs was made clear by workshop speaker Dr. Bruce Alberts, president of the U.S. National Academy of Sciences, who presented the following list of six major challenges that he expects to be addressed during the careers of students currently training to become cell biologists:

1. Graduate from cartoons to a real understanding of each protein machine.
2. Completely understand one type of cell, such as mycoplasmas (i.e., being able to predict what will happen when one of the components is changed).
3. Understand how cells make decisions in complex environments, such as in a multicellular organism (he called this "cell thinking").
4. Understand how cells organize, and reorganize, their internal space.
5. Decipher the pathways by which cells and other organisms evolved on the earth.
6. Use our increasingly profound understanding of biology to design intelligent strategies to understand diseases.

A key challenge to achieving these and other goals for biology will be the development of quantitative experimental methods to identify and characterize comprehensively all of the biological components and their interactions. The following experimental data sets were listed as necessary to achieve a global view of biological processes:

1. A complete, fully annotated genome sequence.
2. An accurate "parts list" of all the proteins and mRNAs in the cell: annotation.
3. A graph of all the interactions taking place between these agents: pathways.
4. A quantitative description of each interaction.
5. A map describing the subcellular localization of each interaction.

For these data sets to be used effectively in predictive models of high-level cellular function, they will need to satisfy many criteria. They must be as complete as possible, include reliable error estimates, and ultimately be able to be assembled into databases from which this data can be extracted and integrated into models. This "systems-level" strategy is a new paradigm for biological research that will be strongly synergistic with the traditional

"hypothesis-driven" approach. As described in the next sections, systems-level biology will require the development of a large informational and computational infrastructure to collect, archive, annotate, integrate, and understand the data from these new experimental tools.

The Nature of Quantitative Biology

The presentations and discussions at the workshop made clear that computational modeling will be at the heart of future biological research. It was noted by several speakers and panelists that theoretical and computational biology are not entirely new fields, but that so far these fields have had relatively little impact in biology. A number of reasons for this were debated, including previous limitations in computer capabilities, but the clear consensus was that these earlier efforts were limited by a lack of experimental data and the means to verify the models quantitatively. There also was agreement on the key requirements necessary to create a successful new biology. The methods and results of quantitative and predictive biology must:

1. Be guided by the important biological questions of the day;
2. Tightly integrate computational analysis and experimental characterization of biological systems;
3. Draw on multiple types of experimental information and computational analyses;
4. Be made accessible to those not extensively trained in computational simulation; and
5. Ultimately use computation and modeling to drive hypothesis formulation, experiment design, and data collection.

Key also will be the need for scientists trained to be part of such a multidisciplinary research program—ideally this new generation of scientists will be equally "intellectually comfortable" in both biology and computation.

Creating the Scientific Environment for Computational and Systems Biology

The challenges to creating a successful environment for this new form of biology were discussed extensively at the workshop. Central to all of the challenges was research funding and the related issue of how credit is awarded for multidisciplinary scientific advances. (One speaker described the quandary of being considered too abstract to be respected by biologists but not sufficiently rigorous to be respected by computational scientists.) On both issues, the current research environment is strongly biased towards the traditional model of an individual researcher guiding a small number of graduate students and post-docs using well-established methods to make incremental progress towards addressing a specific biological hypothesis.

Although this approach has been very successful in bringing biology to its current level of success, there are a number of adverse consequences of this model. It provides very few opportunities for developing and maintaining an information infrastructure, including networks, computers, databases, and "production-grade" software. A point made repeatedly in

the workshop was that the creation and maintenance of robust databases and simulation tools require the sustained efforts of trained professionals and that the development of the necessary mathematics and algorithms will require research investment in these areas. Nor are these tools likely to be provided by private companies. Currently, much of the investment in such information infrastructure is in private companies; consequently, the products can be very expensive to outside users (if available at all) and often are narrowly focused on the individual company's needs. An even greater drawback of leaving the development of computational biology tools to the commercial sector is that they usually are protected by complex intellectual property rules that greatly limit the ability of researchers to evaluate and build upon these methods.

More broadly, the challenge of fostering innovation in biology was discussed, in particular the issue of changing the current tendency for funding agencies to create inadvertently research and training programs that are narrow and overly conservative. (Several workshop speakers cited the lack of funding mechanisms for public-sector multidisciplinary research as one reason that so much talent in computational and systems biology has moved from universities and government labs to private industry.) Not only are successful researchers implicitly discouraged from venturing into new scientific areas, but their former post-docs and graduate students typically must continue to be involved in their advisor's area of research in order to have the best chances for securing their own funding. However, recent experiments to promote expertise from multiple disciplines in applications for research grants have not been as successful as originally hoped. The reasons are complex, but they indicate at least that simply constructing solicitations that encourage multiple disciplines among the principal investigators may not be enough. A clear conclusion from this workshop was that computational and systems biology will need funding models different from those currently available.

Training the Next Generation of Life Scientists

Another issue that was widely discussed was the issue of training life science researchers to have the necessary knowledge to exploit a computational approach to biological research. Bruce Alberts pointed out that life sciences students are receiving less and less breadth in their educations, and, specifically, that biology students receive very little mathematical, physical, or computer science training. Peter Karp noted further that the situation is even worse in the more specialized topics such as databases. Without any individuals with expertise crossing the discipline boundaries, participants believed that there is little prospect that the necessary collaborations can be fostered.

Several models for creating multidisciplinary researchers were discussed. Prospects seem very good for attracting mathematical, computational, and physical scientists to biology—indeed, many of the workshop speakers and attendees were originally trained in fields other than biology. However, there was clear agreement that having scientists from other disciplines simply "parachute" into biology would not make much of a contribution, especially if they try to apply directly the tools of their original disciplines. Instead, prospects are much better if

they take inspiration from the original field but develop new tools and methodology for biological research—for example, applying the concept of model-driven research from solid-state physics to understanding signaling pathways in cells.

The Critical Linkage between Modeling and Experiment

Another common theme at the workshop was the importance of a close linkage between modeling and experiment. In many areas of physical science, this linkage is fairly distant, such as in chemistry and physics, where theoreticians and computational scientists publish in separate journals and attend separate conferences from experimentalists, and train graduate students and post-docs who have no direct experience with experimental methods. Nevertheless, the results of theory and simulation play an important role in the physical sciences, and experimental research groups increasingly perform routine simulations using commercial software. The overwhelming opinion from workshop attendees was that such a model would not be effective for making computational biology fulfill its full promise. This is due to many factors, including the vast complexity of biological systems and the consequent lack of a fundamental theoretical basis for explaining biological phenomena. Additionally, unlike the physical sciences, biology does not have a long history of experimentation driven by quantitative predictions from theory, and hence biologists do not look to the theoretical biology literature for guidance. Theory-driven biology will arise only as breakthroughs in scientific understanding are achieved through collaborations between theorists, computational modelers, and experimental biologists.

Organization and Management of Systems Biology Research

The scientific goals of systems biology will require research management structures that are different from most current biological research projects. During the workshop a number of different organizational strategies were discussed, ranging from large engineering projects, such as those employed in the development of aircraft and satellites, to the large DNA sequencing efforts in the Human Genome Project. Many systems biology projects will involve long-term technology developments and highly multidisciplinary teams of senior scientists. There are many challenges to performing this type of research in the academic model. The new organizational schemes will have to balance many factors:

1. Maintaining innovation and creativity over a long-term project;
2. Avoiding the “not invented here” syndrome;
3. Allowing career advancement for participating researchers;
4. Effective mentoring of student and post-doc team members;
5. Maintaining funding flexibility for different parts of the project;
6. Needing to devote more time to communication between team members; and
7. Providing sufficient management and administrative support for large projects.

Strategies to Design Federal Research Programs in Computational and Systems Biology

Biology is widely noted as the next scientific frontier and as the next "killer application" for high-end computational science. It also will eventually drive both computer science research and the design and investment in high-performance computers and networks. However, funding agencies are still working to refine effective strategies to develop research programs in computational and systems biology. In part, this is because computational biology is still a relatively small subfield of biology and therefore doesn't yet have a large constituency—somewhat like the early days of the genome sequencing programs. As computational biology begins to have more scientific impact on the field and the tools become more widely used, this difficulty will be reduced.

The second challenge is the heterogeneity of computational biology applications. Other scientific communities, such as climate modeling or combustion, typically have a single major computational application that has an unambiguous need for very high performance computing, so that it usually is easy to estimate the improvements that will be achieved by specific investments in software or hardware. In this case, as was clear from the diversity of talks at the workshop, there is a huge variety of computational biology applications, including databases, sequence annotation, protein structure prediction, biochemical simulations, metabolic network modeling, and many others. Each involves different types of computer science and different barriers to progress, typically not just the need for faster computers and more efficient numerical algorithms.

A number of strategies to develop programs in computational and systems biology were discussed at this workshop. One is to link more clearly the results of quantitative biosciences to national needs. For example, DOE is developing new computational and systems biology programs to support its missions in the roles of microorganisms in climate change and energy production, bioremediation of energy and nuclear materials waste, the health risks of low dose radiation exposure, and the basic bioscience needed for effectively defending against biological attack. Another key strategy is to form partnerships between agencies and offices funding biology and other relevant disciplines. For example, a new partnership has been developed between the DOE Offices of Biological and Environmental Research and the Office of Advanced Scientific Computing Research in developing computational and experimental biosciences programs, including joint grant solicitations and multidisciplinary review teams.

Conclusions

More than anything else, this workshop made clear that these are exciting times for biology. We are at the threshold of elucidating the mechanisms for many of the fundamental processes of life, and these results offer vast promise in solving problems in human health, environmental cleanup, energy management, and protection from emerging national security threats. This progress depends on the emergence of a new quantitative, predictive, and, ultimately, systems-level paradigm for the life sciences. There are many challenges to the full realization of this new biology. Many new experimental methods must be developed to provide comprehensive,

highly accurate data sets and the necessary computational infrastructure, software, and algorithms must be developed to use these data sets effectively. A new generation of life scientists must be trained who are facile with the methods of both experimental biology and computational science. Finally, new models for organizing, managing, and funding the biosciences must be developed that will enable large-scale, multidisciplinary research projects in biology. Despite these challenges, the promise that this new biology holds for nearly all aspects of human endeavor, combined with the enthusiasm of scientists from the physical, natural, and informational sciences, means that there are excellent prospects for rapid progress. This workshop constituted a first step towards this goal, by beginning to establish a common language and set of goals across the many scientific disciplines and constituencies involved. The remaining steps will involve the coordinated efforts of many governmental agencies, research and educational institutions, industries, and researchers from many scientific disciplines.

Appendices

Appendix A: Workshop Attendees, September 2001

Appendix B: Agenda

Appendix A: Workshop Attendees, September 2001

Bruce Alberts	National Academy of Sciences
Carl Anderson	Brookhaven National Laboratory
Steve Ashby	Lawrence Livermore National Laboratory
Ray Bair	Pacific Northwest National Laboratory
Michael Banda	Lawrence Berkeley National Laboratory
Yaneer Bar-Yam	New England Complex Systems Institute
Mina Bissell	Lawrence Berkeley National Laboratory
Elbert Branscomb	Joint Genome Institute
Michelle Broido	University of Pittsburgh
Eugene Bruce	National Science Foundation
Carol Bult	The Jackson Laboratory
William Camp	Sandia National Laboratories
Denise Casey	Oak Ridge National Laboratory
Marvin Cassman	NIH National Institute of General Medical Sciences
Su Chung	geneticXchange
Dean Cole	U.S. Department of Energy
Michael Colvin	Lawrence Livermore National Laboratory
David Deerfield	Pittsburgh Supercomputing Center
Charles DeLisi	Boston University
Greg Dilworth	U.S. Department of Energy
David Dixon	Pacific Northwest National Laboratory
Daniel Drell	U.S. Department of Energy
Inna Dubchak	Lawrence Berkeley National Laboratory
Robert Eades	International Business Machines Corporation
Michael Eisen	Lawrence Berkeley National Laboratory
Leland Ellis	U.S. Department of Agriculture
Michael Elowitz	The Rockefeller University
Brendlyn Faison	U.S. Department of Energy
Dan Fraenkel	Harvard Medical School
Marvin Frazier	U.S. Department of Energy
Nir Friedman	Hebrew University
Dave Galas	Keck Graduate Institute
Angel Garcia	Los Alamos National Laboratory
Al Geist	Oak Ridge National Laboratory
Julie Gephart	Pacific Northwest National Laboratory

Paul Gilna ..... Los Alamos National Laboratory
Peter Good ..... NIH National Human Genome Research Institute
Frank Greene ..... National Science Foundation
John Guckenheimer ..... Cornell University
Frank Harris ..... Oak Ridge National Laboratory
Maryanna Henkart ..... National Science Foundation
Daniel Hitchcock ..... U.S. Department of Energy
John Houghton ..... U.S. Department of Energy
Tim Hubbard ..... The Sanger Centre
Tom Hunt ..... Conkling Fiskum & McCormick Inc.
Fred Johnson ..... U.S. Department of Energy
Gary Johnson ..... U.S. Department of Energy
Peter Karp ..... SRI International
Arthur Katz ..... U.S. Department of Energy
Mary Kennedy ..... California Institute of Technology
Michael Knotek ..... DOE Consultant
Daphne Koller ..... Stanford University
Norm Kreisman ..... U.S. Department of Energy
Eric Lander ..... Whitehead Institute/MIT Center for Genome Research
Alan Lapedes ..... Los Alamos National Laboratory
Douglas Lauffenburger ..... Massachusetts Institute of Technology
William Lester, Jr. ..... University of California, Berkeley
Andre Levchenko ..... Johns Hopkins University
Michael Levitt ..... Stanford University School of Medicine
Rob Lipshutz ..... Affymetrix
Phil LoCascio ..... Oak Ridge National Laboratory
William Lorenson ..... General Electric
Peter Lyster ..... National Institutes of Health
Lee Makowski ..... Argonne National Laboratory
Natalia Maltsev ..... Argonne National Laboratory
Reinhold Mann ..... Oak Ridge National Laboratory
Betty Mansfield ..... Oak Ridge National Laboratory
Harley McAdams ..... Stanford University School of Medicine
Carl Melius ..... Sandia National Laboratories
Jill Mesirov ..... Whitehead Institute for Genome Research
Juan Meza ..... Sandia National Laboratories
Saira Mian ..... Lawrence Berkeley National Laboratory

George Michaels ..... Monsanto Company
Edward Monachino ..... NIH National Cancer Institute
Gary Montry ..... Southwest Parallel Software
John Moult ..... University of Maryland Biotechnology Institute
Gene Myers ..... Celera Genomics
Thomas Ndousse-Fetter ..... U.S. Department of Energy
Magnus Nordborg ..... University of Southern California
Edward Oliver ..... U.S. Department of Energy
Bernhard Palsson ..... University of California, San Diego
Aristides (Ari) Patrinos ..... U.S. Department of Energy
Alan Perelson ..... Los Alamos National Laboratory
Walter Polansky ..... U. S. Department of Energy
Kimberly Rasar ..... U.S. Department of Energy
John Rice ..... International Business Machines Corporation
Victoria Roberts ..... The Scripps Research Institute
Daniel Rokhsar ..... Lawrence Berkeley National Laboratory/Joint Genome Institute
Charles Romine ..... U.S. Department of Energy
Joh Von Rosendale ..... U.S. Department of Energy
Kenneth Rudd ..... University of Miami School of Medicine
David Schneider ..... Cornell University
Mary Anne Scott ..... U.S. Department of Energy
Arend Sidow ..... Stanford University
Richard (Dick) Smith ..... Pacific Northwest National Laboratory
Temple Smith ..... Boston University
Sylvia Spengler ..... National Science Foundation
Rick Stevens ..... Argonne National Laboratory
Walter Stevens ..... U.S. Department of Energy
Gary Strong ..... National Science Foundation
Lisa Stubbs ..... Lawrence Livermore National Laboratory
Damir Sudar ..... Lawrence Berkeley National Laboratory
David Thomassen ..... U.S. Department of Energy
Masaru Tomita ..... Keio University
Claire Tomlin ..... Stanford University
Jill Trewhella ..... Los Alamos National Laboratory
Edward Uberbacher ..... Oak Ridge National Laboratory
Mike Viola ..... U.S. Department of Energy
Eberhard Voit ..... Medical University of South Carolina

Scott Weidman ..... National Research Council

Andy White ..... Los Alamos National Laboratory

Owen White ..... The Institute for Genomic Research

Steven Wiley ..... Pacific Northwest National Laboratory

Barbara Wold ..... California Institute of Technology

John Wooley ..... University of California, San Diego

Margaret Wright ..... Bell Labs

Judy Wyrick ..... Oak Ridge National Laboratory

Adong Yu ..... Marshfield Medical Research Foundation

Thomas Zacharia ..... Oak Ridge National Laboratory

Appendix B

Agenda
Visions for Computational and Systems Biology Workshop for the Genomes to Life Program

Thursday, September 6, 2001

Keynote Talks on Visions for Computations and Biology

9:00–10:00	Arrival and Coffee
10:00–10:15	Introductory Remarks: Eric Lander
10:15–10:30	DOE Visions in Computations and Biology: Ari Patrinos, Edward Oliver
10:30–11:00	Visions for the Future of Cell Biology: Bruce Alberts
11:00–11:30	Gene Myers, Celera Genomics
11:30–12:00	Michael Eisen, Lawrence Berkeley National Laboratory
12:00–1:00	Lunch
1:00–1:30	Harley McAdams, Stanford University School of Medicine
1:30–2:00	Claire Tomlin, Stanford University
2:00–2:30	Bernhard Palsson, University of California, San Diego
2:30–3:00	Doug Lauffenburger, Massachusetts Institute of Technology
3:00–3:30	Break (Refreshments served)
3:30–4:00	Peter Karp, SRI International
4:00–4:30	Michael Levitt, Stanford University School of Medicine
4:30–5:00	Summary and Observations: Eric Lander
5:00–7:00	Reception (Latham Hotel)

Friday, September 7, 2001

8:00–8:30	Arrival and Coffee
8:30–9:30	Panel Discussion 1: Barbara Wold, Mary Kennedy, Andre Levchenko, Michael Elowitz—Interaction of Biological Experiments and Modeling
9:30–10:30	Panel Discussion 2: Eric Lander, John Wooley, Gene Myers, Bernard Palsson, Masaru Tomita, Michael Eisen, Owen White—From Functional Annotation to Cell Models
10:30–11:00	Break
11:00–12:00	Panel Discussion 3: Rick Stevens, Steven Ashby, Peter Karp, Bill Lorensen, John Guckenheimer, Dan Reed—Advances in Computer Science and Their Promise for Biology
12:00–1:00	Lunch
1:00–2:00	Panel Discussion 4: David Gifford, Harley McAdams, Doug Lauffenburger, Nir Friedman—High Level vs Low Models
2:00–2:30	Concluding Address: Charles DeLisi