New York University
A private university in the public service

Margaret H. Wright
Silver Professor and Chair

Computer Science Department
Courant Institute of Mathematical Sciences
251 Mercer Street
New York, New York 10012
Telephone: 212-998-3056
Fax: 212-995-4124
Email:
mhw@cs.nyu.edu

December 16, 2002

Dr. Raymond L. Orbach
Director, Office of Science
U.S. Department of Energy, SC-1
19901 Germantown Road
Germantown, Maryland 20874

Dear Ray,

This letter accompanies the final report of the Biotechnology Subcommittee of the Advanced Scientific Computing Advisory Committee (ASCAC), Office of Science, Department of Energy. The ASCAC Biotechnology Subcommittee was created in response to a letter dated April 16, 2001, from Dr. James Decker, then the Acting Director of the Office of Science, in which he asked ASCAC to provide advice on “the computational side of biotechnology”, particularly

- Areas on which the Advanced Scientific Computing Research (ASCR) program should target investments to have maximum impact on the underlying science, and
- How to most effectively couple research supported by ASCR with discipline-specific research carried out by biologists.

Dr. Decker's letter requested that we involve “experts from outside of ASCAC membership as necessary”. The combination of my letter and the accompanying subcommittee report is hereby submitted to you as the official ASCAC response to Dr. Decker's charge.

The Biotechnology Subcommittee, chaired by Juan Meza of ASCAC, consists of (from ASCAC) William Lester and Margaret Wright, and (from outside ASCAC) Michael Colvin of Lawrence Livermore National Laboratory, John Guckenheimer of Cornell University, and Bruce Hendrickson of Sandia National Laboratories; the latter three are experts in various aspects of computational biology. For more than a year, members of the subcommittee collected information about developments in biotechnology related to computation, conducted interviews with experts, and attended relevant workshops, some sponsored by the Department of Energy. The subcommittee's report includes five recommendations, all of which were unanimously endorsed by its members, as well as a detailed discussion of the rationale for the recommendations.

A draft report of the Biotechnology Subcommittee was presented to the members of ASCAC at the most recent ASCAC meeting, held on October 17 and 18, 2002. There was an extended discussion of the report at the meeting, and a further discussion by email. The final version of the report that accompanies this letter is the result of those discussions. All of the report's recommendations except one were approved unanimously by email by the members of ASCAC.

The exception is the first recommendation of the subcommittee:

The Office of Advanced Scientific Computing Research should not develop any specialized computational facilities for computational biology in the near term, as the underlying research problems are still too poorly understood. Instead OASCR should expand upon the SciDAC program to bring biologists, chemists, mathematicians, and computer scientists together as a team.

Nine members of ASCAC support this recommendation. However, three members of ASCAC (John Connolly, Jill Dahlburg, and Ellen Stechel) support the following alternative statement:

The Office of Advanced Scientific Computing Research should expand upon the SciDAC program to bring biologists, chemists, mathematicians, and computer scientists together as a team. Because underlying research problems in computational biology are still poorly understood, development of specialized computing facilities is likely inappropriate at this time. Consequently, any proposal for specialized facilities in computational biology should be rigorously justified with competitive peer review in terms of programmatic value, including cost effectiveness, before any specialized hardware funding is approved.

Please let Juan Meza know if you have any questions about the subcommittee's report, and please let me know if you have any questions arising from this letter.

With best wishes.

Sincerely,

Margaret H. Wright
Chair, Advanced Scientific Computing Advisory Committee

P.S. Your speech at SC2002 was great!

Encl: Biotechnology subcommittee report (as email attachment)

Dr. Raymond Orbach
Director, Office of Science
U.S. Department of Energy, SC-1
19901 Germantown Road
Germantown, MD 20874

Dr. Orbach,

I write to give you the report of the Advanced Scientific Computing Advisory Committee (ASCAC) Biotechnology Subcommittee. Our report is in response to the charge given to the chair of ASCAC by Dr. Decker in which he asked the committee to advise the Office of Science on the issue of computational biotechnology. In response to the letter, Dr. Margaret Wright, the chair of ASCAC, formed a subcommittee, chaired by myself, to address the two specific questions that were outlined in the letter:

1) What are the areas on which the Advanced Scientific Computing Research (ASCR) program should target investments to have maximum impact on the underlying science? Possible examples of areas include specialized facilities for biological computation, basic research in underlying mathematical algorithms, or advanced computer science related to data management.
2) How to most effectively couple research supported by the ASCR program with discipline-specific research carried out by biologists?

In brief, the subcommittee's recommendations are:

- **The Office of Advanced Scientific Computing Research should not develop any specialized computational facilities for computational biology in the near term, as the underlying research problems are still too poorly understood. Instead OASCR should expand upon the SciDAC program to bring biologists, mathematicians, and computer scientists together as a team.**
- **The ASCR program should continue to invest in biophysics and biomolecular simulations, which are already having an impact in the underlying science;**
- **Computational biology will drive new fields of mathematics and computer science. The ASCR program should address these new areas through investments in fundamental mathematical and computer science algorithms;**
- **The ASCR program should develop new database and scientific data management infrastructures that can be used for computational biology;**
- **The Office of Advanced Scientific Computing Research should help to develop training programs for the next generation of computational biologists;**

The rest of this memo provides details on our findings and recommendations from our individual research, the various workshops, and our discussions with other researchers in this field.

The Office for Advanced Scientific Computing Research has already undertaken an ambitious program in computational biology known as the Genomes to Life (GTL) program. The GTL program has as one of its primary goals to "establish, within a decade, a national infrastructure to transform the tremendous outpouring of data and concepts into a new computationally based biology". [1] In particular, this program has four technical goals of which the fourth is to develop the computational methods and capabilities to advance understanding of complex biological systems and predict their future. As part of our research we attended several of the GTL workshops to discuss the challenges in mathematics, computer science and their infrastructure. These workshops spanned a period of several months from August 2000 through March 2001. Some of these findings have been documented in a separate draft report[2] that was presented at the ASCAC meeting of May 2-3, 2002. Another workshop that several members of the subcommittee attended focused on systems biology and was organized by Dr. Eric Lander on September 6-7, 2001 (the Lander report is included as an appendix). Finally, we interviewed researchers in the field as to their thoughts and visions for the future of computational biology. The balance of this letter is organized around the original set of questions posed to the committee.

What are the areas on which the Advanced Scientific Computing Research (ASCR) program should target investments to have maximum impact on the underlying science? Possible examples of areas include specialized facilities for biological computation, basic research in underlying mathematical algorithms, or advanced computer science related to data management.

The first issue that the subcommittee addressed was the definition of "computational biology" itself. This term has many different meanings and includes many different research areas including "databases, sequence annotation, protein structure prediction, biochemical simulations, metabolic network modeling and many others"[3]. The subcommittee also decided to focus on those biological areas that were most relevant to the mission of the DOE and those mathematical and computational areas where OASCR could have the greatest impact.

The full Advisory Committee discussed the question of specialized facilities for computational biology at our May 2-3, 2002. This topic was also discussed in a separate memorandum[4] sent to you from the Advisory Committee on May 21, 2002 in which we stated our views on the related concept of topical centers. We therefore only summarize

[1] Genomes to Life Accelerating Biological Discovery, DOE/SC-0036, April 2001, http://doegenomestolife.org

[2] http://www.newbiology.org/draft/index.html

[3] The "Lander" report, "Visions for Computational and Systems Biology, A Report on the DOE Workshop", (see Appendix B)

[4] Memorandum sent to Dr. Raymond Orbach on May 21, 2002 (see Appendix A)

those findings here. The consensus among the Advisory Committee members is that specialized computational facilities devoted to a single discipline do not offer any advantages over general facilities. The committee also expressed serious concerns over related concepts described at various times as either topical centers or topical facilities. Two specific concerns that we raised during the May 2-3, 2002 meeting included choosing the correct structure for such a topical center so that it doesn't lead to a dead end and the inability to work with researchers from other fields if the center is defined as too narrow or becomes isolated.

In the subsequent memorandum, we also described several characteristics that we deemed important to any successful computational center. We highlight three of those points here: 1) a need for a defined focus, 2) a need for a center to be collaborative and multidisciplinary and 3) a need for competitive peer review. Examples of a defined focus could include organizing along the lines of related disciplines, such as a group of scientific applications that share a common set of algorithms or related computational models, such as data-intensive applications. The need for the center to be collaborative and multidisciplinary was deemed just as important to its success. The Committee believes that providing the environment to build critical mass and to bring researchers together to work across disciplines, including application developers, applied mathematicians, and computational, computer, and algorithmic scientists is critical to its success. Finally, the Committee strongly believes that any process that OASCR should choose in implementing these strategies be done through a competitive peer review process.

One of our major findings is that more research is required in understanding the fundamental mathematical and computer science issues in computational biology and that the primary emphasis should be placed on addressing these issues. We believe that it is premature to develop any computational facilities devoted to computational biology before understanding these issues. These thoughts were paraphrased in a quote attributed to Donald Knuth[5] that, "Premature optimization is the biggest source of programming inefficiency".

It is our recommendation, that the Office of Advanced Scientific Computing Research should not develop any specialized computational facilities for computational biology in the near term, as the underlying research problems are still too poorly understood. In many ways, however, the ideas mentioned above for a computational center are already embodied in the very successful SciDAC program. It is our recommendation that OASCR should expand upon the SciDAC program to bring biologists, mathematicians, and computer scientists together as a team.

The ASCR program is already having an impact on the underlying science in the areas of biophysics and biomolecular simulations. A variety of chemical simulation methods have been developed that trade off accuracy for computational cost. At one extreme are

[5] Literate Programming, "We should forget about small efficiencies, about 97% of the time. Premature optimization is the root of all evil".

quantum mechanical methods that can in principle predict any biochemical property to high accuracy but are computationally limited to determining static properties of modest-sized molecular systems. At the other extreme is classical molecular dynamics (MD) that can simulate the motions of much larger biochemical systems, but using simplified “ball and spring” force fields that have limited accuracy. These different methods are complementary—for example, classical molecular dynamics can provide information on large-scale conformational changes in biological macromolecules, while quantum mechanical simulations can determine the effect of these changes on their activity. Both of these methods are now well developed and widely used in the study of biology systems. Nevertheless, there is much room for continued investments in algorithmic improvements that will depend on fundamental mathematical advances. Two such examples are the development of algorithms that allow for longer time steps in molecular dynamics and linear scaling algorithms for quantum chemical methods. Such improvements would have a large impact on the field for the very reason that they are so widely used.

In addition to improving well-established biophysical simulation tools, algorithmic research could help enable new simulation methods on the horizon that show great promise. For example, recent improvements in computer speeds and algorithms now allow a new type of simulation that combines the accuracy of static quantum chemical methods with the ability to simulate the motions of atoms in biochemical systems. This so-called “first principles molecular dynamics” requires teraflop-speed computers to simulate even a few hundred atoms for a few picoseconds. Despite this computational cost, these methods constitute a nearly exact simulation of nature, and offer the promise of transforming our ability to understand dynamical biochemical processes. The results for small biochemical systems currently being simulated on teraflop scale computers provide tantalizing glimpses of the value of longer time and larger system size simulations that will be made possible with faster computers and improved algorithms. We recommend that the ASCR program should continue to invest in biophysics and biomolecular simulations, which are already having an impact in the underlying science.

It is also clear that there is a growing body of problems that will require new mathematical and computational techniques and that these areas will play a major role in making biology a predictive science. The types of simulations envisioned in biology are fundamentally different from the types of physical systems currently studied and modeled. Some examples include metabolic pathway analysis, the development of kinetics models, inverse modeling of a protein from its crystal form to its unfolded state, and model-based design of experiments. The characteristics of these new systems include inherently noisy, complex, and self-regulating systems and the experimental data that could be used to validate models is difficult if not impossible to obtain. In addition, the data comes from a wide variety of experimental techniques such as *in vivo* optical tracking, DNA and protein microarrays, as well as NMR, mass spectrometry, X-ray, and neutron scattering that must be combined. Finally, in many cases there is even a lack of understanding as to whether all of the underlying components are being modeled. This aspect was described in the Lander report as "graduating from cartoons of multiprotein machines to a real understanding". In these respects, it is clear that modeling biological systems will drive fields of mathematics and computer science that have not received much attention in the past and even entirely new fields. A few examples of areas where fundamental advances are needed to enable real understanding of systems biology include:

- **Stochastic dynamical systems:** Many molecules within a cell are present in such small numbers that deterministic modeling of the system as a system of chemical reactions is inadequate. The fluctuations inherent in processes involving small numbers of molecules are large enough that there is macroscopic variability in the dynamics of the biological processes.

- **Parameter estimation:** The robustness of cells in adapting to changes in their environment or to genetic variation suggests that the values of many parameters may range widely with little effect on a cell's behavior. On the other hand, there are other parameters to which cells show exquisite sensitivity. The problem is to use the observed output of the whole cell under varying environmental conditions to estimate unknown parameters in the models.

- **Multiple time scales:** We frequently describe cellular processes in terms of switches; genes are "turned on" and "turned off". Models can represent this process as discrete events or involve multiple time scales. In either case, there are substantial computational issues in implementing model simulations. There also are substantial mathematical issues about how multiple time scales affect dynamics. Although theory bolsters our intuition about the system behavior, new phenomena that are still poorly understood occur in the presence of multiple time scales.

- **Model reduction:** Simulation of large, complex models with multiple time scales is computationally demanding. To focus upon the critical elements of system behavior in a particular phenomenon, it is desirable to develop reduced models that embody only the critical elements. This enables more extensive exploration, leading to insights that can then be tested both experimentally and with simulations of larger models.

- **Computational topology:** At the molecular level, the reactions that occur involve the shapes of macromolecules and complexes formed from these molecules. At a higher level, cellular components often have very complex morphologies that have likely evolved for a specific purpose. We have only a fragmentary understanding of how shape plays a role at the molecular or cellular level in regulatory processes. The use of topological properties could lead to new understanding of the behavior of these systems and their regulatory processes.

- **Hierarchy:** Cells have large numbers of subsystems, represented by organelles and molecular complexes. The functional role of this hierarchy in determining the robustness of a cell and how it regulates its activities is unclear. Abstractly, we would like to know how the characteristics of the graphs depicting gene regulatory networks and protein interactions affect dynamical processes within a cell and its interaction with its environment.

- **Data Interpretation:** While the resulting flood of new experimental data is enabling high-throughput biology, techniques for adequately interpreting this information are often lagging. One particularly noteworthy example is the enormous effort required to assemble a genome from shotgun sequence data. Other examples include protein fragment assembly from mass spectrometry data, gene network deduction from microarray experiments, and structure determination from NMR and x-ray crystallography analyses. In addition, even modest changes in experimental methodologies can have a dramatic impact on the computational techniques required to analyze the data. For many of these problems, discrete algorithms are a key part of the data interpretation, and continued development of applied combinatorics is required. Another recurring theme is the presence of significant noise in the data, which leads to a need for novel statistical techniques.

This is by no means a complete or exhaustive list, but is only meant to highlight some of the many new possibilities. Advances in mathematics and computation that address the issues listed above are needed to achieve a deeper understanding of systems biology. It is our finding that the Office of Advanced Scientific Computing Research can make significant contributions in these areas. In summary, computational biology will drive new fields of mathematics and computer science. The ASCR program should address these new areas through investments in fundamental mathematical and computer science algorithms.

The last issue addressed advanced computer science methods related to data management. The raft of new experimental methodologies has led to an explosion of new types and quantity of biological data including, but not limited to, genomics, protein sequence and structure, and gene and protein pathways. Mining this data has become a new and essential component of biological research and the committee identified these areas as requiring new investments. Today, many of the data bases used in the biological community are heterogeneous, distributed and usually contain many errors. To take full advantage of the flood of information that will soon be available, scientists need to be able to access, combine, and query these biological data sets easily and efficiently. An example from the Lander report cites several needs including the development of common, low-level data-interchange methods, the development of common ontologies, and the establishment of automated query access. This particular area could have a substantial payoff in terms of return on investment. It is our recommendation that the ASCR program should develop new database and scientific data management infrastructures that can be used for computational biology.

How to most effectively couple research supported by the ASCR program with discipline-specific research carried out by biologists?

An important finding of this subcommittee is that there is an urgent need to train the next generation of computational biologists who can serve to bridge the gap between biology and computer sciences. To quote from the report on Visions for Computational and Systems Biology Workshop for the Genomes to Life Program, September 6-7, 2001, "Without any individuals with expertise crossing the discipline boundaries, participants believed that there is little prospect that the necessary collaborations can be fostered." We recommend that the Office of Advanced Scientific Computing Research help to develop training programs for the next generation of computational biologists. One possibility might be to have a series of short courses on computational biology that were co-sponsored by OASCR and OBER. Another possibility would be to take advantage of the ongoing and highly successful Computational Sciences Graduate Fellowship program to attract students into these new areas.

I hope that you find these results of this subcommittee useful. Please contact me if you need amplification or clarification of my remarks.

Sincerely yours,

Juan Meza
Chair, ASCAC Biotechnology Subcommittee

For the ASCAC Biotechnology Subcommittee:
Michael Colvin
John Guckenheimer
Bruce Hendrickson
William Lester
Margaret H. Wright

Appendix A

E-mail to Dr. Raymond Orbach
May 21, 2002

Today's scientific world is poised to experience a paradigm shift to a world in which simulation and computation are equal partners with theory and experiment, a world in which no phenomenon is too complex to dissect and reconstruct and a world rich with currently unimagined new possibilities. A program such as SciDAC, which brings together computational scientists, application developers, computer scientists and applied mathematicians, is a significant and necessary enabler for driving this shift. Nevertheless, the challenge remains to create the complete set of necessary and sufficient enablers that will facilitate as well as accelerate the transition, without which the shift might stagnate or likely proceed too slowly to meet the next generation of major scientific thrusts. The latter will be increasingly complex, progressively more multidisciplinary, and span a much larger range of spatial and temporal scales than previous thrusts.

Continuous and significant investments in experimental facilities and expert personnel have been the Office of Science's primary strategy for advancing science. Computer hardware and software could and arguably should be viewed in an analogous manner to the experimental facilities. Similar to the latter, an increased focus on computational infrastructure should not be viewed as a substitute for equally necessary attention and growth of the foundational base in the computational and computer sciences, algorithms and applied mathematics. Nevertheless, advancement in computational infrastructure and further development of strong interactions and consequent synergies across disciplines and with the traditional theoretical and experimental sciences are critical and necessary enablers for maximizing value derived from the department's facilities and research investments.

Topical computing centers, targeted and focused on significant scientific challenges, form the basis for a strategy that creates a structure to develop the computational infrastructure, to assure critical mass, and to nurture interactions and synergies, in order to affect the completion of the paradigm shift.

The committee recommends that focused computing centers (Topical Centers) be

1. **Complementary to the Department's flagship center**
2. **Built from but not detract from the strength of the base programs**
3. **Managed as a diverse portfolio**
4. **Selected by competitive peer review**
5. **Sharing the following set of common characteristics:**

    - A defined focus. Examples could include organizing along the lines of
        - Related disciplines, such as a group of scientific applications that share a common set of algorithms.

    - Related computational models, such as data-intensive applications.

- Collaborative and multidisciplinary**.** Providing the environment to build critical mass and to bring researchers together to work across disciplines, including application developers, applied mathematicians, and computational, computer, and algorithmic scientists.

- A specific goal**.** The topical center should have at least one urgent, challenging and exciting end point that is identifiable, plausible, and timely. For example, to take a currently intractable problem with current technologies to a tractable problem through advances in algorithms, applied mathematics, computational science and computer science within three-five years.

- Tuned computer configuration**.** A topical center should cater and balance the computational architecture to best serve the needs of the defined focus area. It should also include research in advanced hardware architecture and/or systems balance.

Appendix B

A Report on the DOE Workshop Visions for Computational and Systems Biology

Introduction:

On September 6-7, 2001, 120 biologists and computational scientists met in Washington DC for a workshop entitled: “Visions for Computational and Systems Biology”. The central theme of this workshop 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 collected and assembled into predictive computational models. This 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 in order for biology to achieve its full promise for improving human well-being. This report summarizes the key findings from this workshop. It describes the long-term goals and major scientific drivers behind computational and systems biology, as well as the discussions related to overcoming the existing barriers in biosciences research. The clear conclusion of this workshop 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. This workshop 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 this vision.

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 capable of accurately predicting all relevant properties of the phenomena. 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 a 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, e.g. mycoplasmas. (i.e. being able to predict what will happen 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 datasets 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 each interaction
5. A map describing the subcellular localization of each interaction

For this data to be used effectively in predictive models of high-level cellular function, it will need to satisfy many criteria. It 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 information 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 quantitatively the models. There was also 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, design of experiments, 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 hypotheses.

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 requires 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, and consequently the products can be very expensive to outside users (if available at all) and are often 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 are usually 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 system 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 must typically 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 indicate at least that simply constructing solicitations that encourage multiple disciplines among the PI's 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 discipline. 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 only arise 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 thedevelopment of aircraft and satellites, to the large DNA sequencing efforts in the Human Genome Program. 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. Need 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 will also 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 and it is usually easy to estimate the improvements that will be achieved by specific investments in software or hardware. 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 justthe 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, the Department of Energy (DOE) is developing new computational and systems biology programs to support DOE 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 it 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 datasets and the necessary computational infrastructure, software and algorithms must be developed to effectively use these datasets. A new generation of life scientists must be trained who are facile with both the methods of 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 from 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 involved the coordinated efforts of many government agencies, research and educational institutions, industries, and researchers from many scientific disciplines.

Appendix I: Workshop agenda

Thursday September 6, 2001

Keynote Talks on Visions for Computational and Systems 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
Bruce Alberts, NAS -- *Some Thoughts on the Future of Cell*

11:00 - 11:30
Gene Myers, Celera Genomics

11:30 - 12:00
Michael Eisen, Lawrence Berkeley National Laboratory

12:00 - 1:00
***Lunch* (Catered)**

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* (Catered)**

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

Appendix II: Workshop attendees

Bruce Alberts
National Academy of Sciences
Carl Anderson
Brookhaven National Laboratory

Steve Ashby
Lawrence Livermore National Laboratory
Ray Bair
Computational Sciences and Mathematics, Pacific Northweest National Laboratory

Michael Banda
Lawrence Berkeley National Laboratory
Yaneer Bar-Yam
New England Complex Systems Institute

Mina Bissell
Ernest Orlando Lawrence Berkeley National Laboratory
Elbert Branscomb
Joint Genome Institute

Michelle Broido
University of Pittsburgh
Eugene Bruce
BIO/IBN, National Science Foundation

Carol Bult
The Jackson Laboratory
William Camp
Computation, Computers and Mathematics, Sandia National Laboratories

Denise Casey
Human Genome Management Information System, Oak Ridge National Laboratory
Marvin Cassman
NIGMS/NIH

Su Chung
geneticXchange
Dean Cole
U.S. Department of Energy

Michael Colvin
Biology and Biotechnology Research Program, 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
IBM

Michael Eisen
Lawrence Berkeley National Laboratory
Leland Ellis
U.S. Department of Agriculture, Agricultural Research Service

Michael Elowitz
The Rockefeller University
Brendlyn Faison
U.S. Department of Energy

Dan Fraenkel
Harvard Medical School
Marvin Frazier
Office of Biological and Environmental Research, 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
Bioscience Division, Los Alamos National Laboratory

Peter Good
National Human Genome Research Institute
Frank Greene
Division of Integrative Biology and Neuroscience, National Science Foundation

John Guckenheimer
Cornell University
Frank Harris
Oak Ridge National Laboratory/UT-Battelle

Maryanna Henkart
Division of Molecular and Cellular Biosciences
Daniel Hitchcock
U.S. Department of Energy/OASCR

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/OASCR
Peter Karp
Bioinformatics Research Group,
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
Life Sciences Division, Oak Ridge National Laboratory
William Lorenson
General Electric

Peter Lyster
NIH
Lee Makowski
Biosciences Division, Argonne National Laboratory

Natalia Maltsev
Argonne National Laboratory, MCS
Reinhold Mann
Life Sciences Division, Oak Ridge National Laboratory

Betty Mansfield
Human Genome News, 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
Genomics, Monsanto Company
Edward Monachino

NIH/NCI/OTIR

Gary Montry
Southwest Parallel Software
John Moult
CARB

Gene Myers
Celera Genomics
Thomas Ndousse-Fetter
U.S. Department of Energy

Magnus Nordborg
University of Southern California
Edward Oliver
Office of Advanced Scientific Computing,
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
Department of Energy
John Rice
IBM

Victoria Roberts
The Scripps Research Institute
Daniel Rokhsar
Lawrence Berkeley National Laboratory/JGI

Charles Romine
U.S. Department of Energy/OASCR
Joh Von Rosendale
U.S. Department of Energy

Kenneth Rudd
University of Miami School of Medicine
David Schneider
Cornell Theory Center

Mary Anne Scott
U.S. Department of Energy
Arend Sidow
Stanford University

Richard (Dick) Smith

Pacific Northwest National Laboratory
Temple Smith
Biomolecular Engineering Research Center

Sylvia Spengler
Biological DB and Informatics, 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
Office of Biological and Environmental Research, 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
Computer and Computational Sciences Division, 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

Human Genome Management Information System, Oak Ridge National Laboratory
Adong Yu
Marshfield Medical Research Foundation

Thomas Zacharia
Oak Ridge National Laboratory