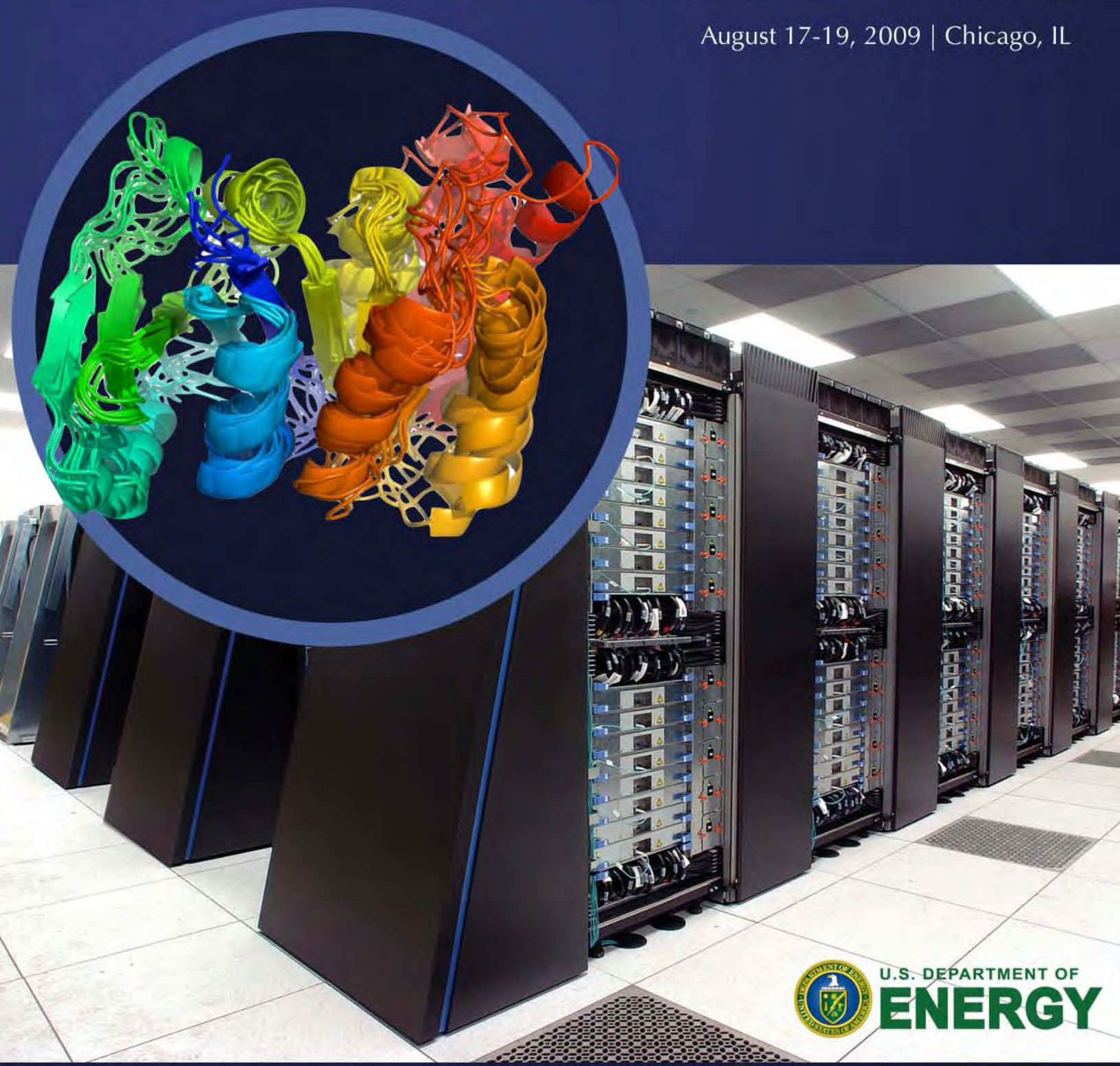
Scientific Grand Challenges

Opportunities in Biology at the Extreme Scale of Computing

August 17-19, 2009 | Chicago, IL



U.S. DEPARTMENT OF ENERGY

Sponsored by:

The Office of Biological and Environmental Research

The Office of Advanced Scientific Computing Research

DISCLAIMER

On the cover: Argonne National Laboratory's IBM Blue Gene/P supercomputer
Inset visualization of ALG13 courtesy of David Baker, University of Washington

AUTHORS AND CONTRIBUTORS

Sponsors and representatives

Susan Gregurick, DOE/Office of Biological and Environmental Science
Daniel Drell, DOE/Office of Biological and Environmental Science
Christine Chalk, DOE/Office of Advanced Scientific Computing Research

Workshop co-organizers

Mark Ellisman, University of California, San Diego
Rick Stevens, Argonne National Laboratory

Workshop session leads

Macromolecular Proteins and Protein Complexes
Mike Colvin, University of California, Merced
Tamar Schlick, New York University

Pathways, Organelles, and Cells
Adam Arkin, Lawrence Berkeley National Laboratory
David Galas, Institute for Systems Biology

Populations, Communities, Ecosystems, and Evolutionary Dynamics: Genomics and Metagenomics
Edward Delong, Massachusetts Institute of Technology
Gary Olsen, University of Illinois at Urbana-Champaign

Tissues, Organs, and Physiology Modeling
John George, Los Alamos National Laboratory
George Karniakadis, Brown University

Data Analysis, Imaging, and Visualization
Chris Johnson, University of Utah
Nagiza Sematova, North Carolina State University and Oak Ridge National Laboratory

CONTENTS

EXECUTIVE SUMMARY

"Thinking big comes naturally to many biologists. Pursuing biological research on a monumental scale traditionally has not."[1]

The history of computing is an accelerating evolution of speed, chips, power, and capital, set in the context of grand challenges and pursued for their national and intellectual value by engaging a diversity of the brightest minds across scientific disciplines and institutional boundaries. Just ahead in this decade are commercial supercomputers more than 1,000 times faster than any current offering, with petaflop performance – computers capable of completing more than 1,000 trillion floating-point calculations (flops) per second. Now the Department of Energy (DOE) and supercomputer engineers aim to leapfrog that benchmark by several orders of magnitude, targeting one million trillion calculations per second to reach the goal of exascale, or to be less limiting, extreme scale computing. (*Exa* is the metric prefix for quintillion, or 10^{18}.) Whatever on earth demands such spectacular capacity? The new integrated biology, and the complex problems it seeks to answer.

In their presentation, **Opportunities in Biology at the Extreme Scale of Computing**, the authors show that dauntingly monumental computational scale can and will be achieved; foresee its revolutionary impact on the futures of energy research, challenges to the environment, and health issues; and offer focused recommendations for developing an exascale computational biology program.

DOE's Offices of Biological and Environmental Research (BER) and Advanced Scientific Computing Research (ASCR) have a joint interest in integrating experimental research with computational modeling, including simulating new developments in the algorithmic and mathematical sciences, to fully understand and utilize a systems level biology for our missions in energy and the environment. This partnership includes both basic research into novel algorithms and methods as well as research to develop algorithms that can scale to leadership class computing facilities. Nevertheless, one cannot reach the goal of discovering and developing systems-level biological properties by a linear advance in computing, either from the hardware or algorithmic side. Therefore, this workshop addresses the fundamental sea change needed in the computational biological sciences that will push systems biology into a realm of truly integrating extreme scale computing with extreme scale grand challenges.

In the past three decades, the fields of computer science and telecommunications were dominated by transformational technologies that rapidly impacted every segment of society. Here, leading bioinformatics and computational biology researchers assert that the next wave of world-altering scientific breakthroughs will be a child of the marriage of bioscience and an unprecedentedly powerful cyberinfrastructure. Indeed, the union of biology and mega-computers was a long time in the making:

> In fact, the application of computational techniques to the simulation and analysis of biological systems has a history dating to the earliest analog and even mechanical computers. The recent resurgence of interest in quantitative modeling can be attributed at least partly to the greater power afforded by modern information technology, but even more to the explosion of data brought about by modern molecular techniques. It is now clear to many

[1] Collins, F.S., M. Morgan and A. Patrinos, "The Human Genome Project: Lessons from Large-Scale Biology," *Science*, vol. 300, pp. 286-290, April 11, 2003.

researchers that future progress in understanding biological function rests inescapably in the development and application of computational methods.[2]

The ultimate goal of exascale computing applications to challenges in modern biology is to go from atoms to organs or from microbes to ecosystems: for example, to enable an understanding of how the brain works as an energy efficient, biologically-based information system, or to understand microbial processes and their impact on the geosphere. In the process, these newly enlarged scales of computing will resolve unfathomably complex research issues in a host of fields as diverse as neuroscience and microbial metagenomics.

Within the next decade we expect to have the complete genome sequence of more than 10,000 bacteria and archaea and other single-celled microbes. Exascale computing platforms will make it possible in principle to systematically reconstruct the metabolic networks of all sequenced microbes through automated comparative analysis, to reconstruct their regulatory networks by integrating a variety of data sources, and to combine these reconstructions into functional models of cellular states.

These integrated models can be used to further integrate and validate experimental datasets, but they can also push forward the ability to predict elements of a cell's phenotype from its genome and knowledge of the environment *for all sequenced organisms*. Exascale computing can employ an iterative method of detailed comparative analysis of the resulting reconstructions and models; that is, perform 10,000 x 10,000 comparative studies simultaneously, to gain insight into the evolutionary trends and decisions that gave rise to the diversity of observed variations.

Biological research is poised for dramatic accelerations by the advances in computing systems over the next decade. The benefits of fully exploiting these systems to advance our understanding will accrue not only to science but also to society at large. DOE's Offices of BER and ASCR co-sponsored the international workshop, giving rise to this report to present the biological community with the opportunity to shape the scientific frontiers in fundamental biological research.

Overall this report addresses what may be termed the 'Grand Challenge Issues' in biology for extreme computing. These include:

- Biophysical simulations of cellular environments, either in terms of long time dynamics, crowded environments and the challenges therein, or coarse graining dynamics for entire cellular systems.
- Cracking the 'signaling code' of the genome across the tree of life, implying a reconstruction of large-scale cellular networks across species, and across time points and environmental conditions.
- Re-engineering the units of biological function at a whole organism scale.
- Correlating observational ecology and models of population dynamics in microbial community systems, and from this basis, to apply these models to microevolutionary dynamics in microbial communities.
- Reverse engineering the brain to understand complex neural systems, as well as to provide new methods to integrate large-scale data, information and simulation dynamics,

[2] Hucka, M. et al., "Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project," Syst. Biol., vol. 1, no. 1, pp. 41, June 2004.

and ultimately to provide societal benefits for health. And perhaps most importantly, to learn how nature's design of a biologically-based knowledge machine beats all estimates for "low power" systems to compute or store information when compared to even the most power efficient hardware systems now being conceived for construction in the future.

In these opening chapters, very specific computational challenges are explored that illustrate the role that computational modeling can play in understanding complex physiological processes and systems.

We begin with an exploration of exascale opportunities in macromolecular proteins and protein complexes—the "nanomachines" involved in virtually all elementary processes of life—covering four areas of biological systems and problems in terms of their increasing complexity of temporal and spatial dimensions. The opportunities are: macromolecular folding, including protein folding and RNA folding; biochemical binding and reaction mechanisms, such as enzyme catalysis and protein/ligand interactions; macromolecular pathways, including DNA replication and repair fidelity, protein synthesis, chromatin organization, and RNA editing; and supramolecular cellular processes, such as protein signaling networks, plant cell-wall formation, and endocytosis.

Though methodological advances have enabled researchers to model the largest assemblies in the cell, they cannot model the length scale of an entire cell in reasonable time frames using current leadership systems. However, new scalable tools that admit a variety of time, space, and trajectory sampling methods (and fully exploit the hundreds of millions of cores expected on an exascale machine) will enable long time integrations, implicit solvation conditions, and mixed molecular mechanics and quantum mechanics models. A large biochemical network within a full-scale model of a eukaryotic cell could be modeled in the span of few hours.

Understanding macromolecular complexes is crucial to drug discoveries that counter certain biothreats as well as advancing bioenergy applications. The discussion in **Chapter 1** includes the substantive technical gaps in achieving exascale biomolecular modeling, with a focus upon threshold capabilities and major technical gaps.

The use of exascale computing to advance our post-genomic era understanding of how the one-dimensional genome translates the code of life into three-dimensional cells frames the discussion in **Chapter 2**. This "signaling code" of the genome includes how it organizes populations of kin and non-kin cells that are communicating and differentiating. At the pathway level, scientists operate under uncertainty about all aspects of the system—from boundary conditions to inputs and outputs. This necessitates an understanding of the factors impeding the prediction, control, and design of cellular behavior, which include data accuracy, analysis, community databases, and data simulation. The main issue delaying the translation of genotype into phenotype through a better knowledge of pathways is a simple lack of data, while the easiest problem to address may be the annotation of genomes to determine the molecular functions they encode, which is nevertheless extremely challenging.

Building models of cells and organisms in the context of evolution—the large-scale analysis and reconstruction of cellular networks—will require the integration of genome analysis tools with modeling environments, high-throughput computing frameworks, and the availability of mixed integer optimization tools in a massive parallel environment.

Directly addressing the challenges that exascale computing is best suited for, *metabolic model reconstruction*—something that is impossible to achieve using today's computational

capabilities—is one of the top ten problems in systems biology, and essential for scientific advances in these fields. It is also a prerequisite for any meaningful reconstruction of transcriptional regulatory networks: the metabolic model is the scaffold on which scientists can assemble and implement regulatory networks. Another likely area for success is the prediction of feasible *parameter values for dynamic models of metabolism* that would enable scientists to design organisms that would perform a variety of tasks. These models might also contribute to the development of treatments for emerging types of infections. This chapter concludes with a comprehensive discussion of future work, and all its myriad possibilities.

As genome sequencing technologies improve, and the importance of decoding the microbial community structure and metabolic potential in terrestrial and pelagic ecosystems is realized, it will become imperative that we develop the means to sequence important environments deeply enough to develop a comprehensive understanding of the microbial communities that are present. **Chapter 3** surveys the challenges presented by populations, communities, ecosystems, and evolutionary dynamics (genomics and metagenomics) as well as how research in each area will benefit from the exascale effort. The emphasis here is on understanding bacteria as well as microbial cells with a nucleus or other organelles in them, for example, eukaryotes. Ultimately this research will inform our understanding of ecosystems and how humans are integrated into these systems.

Building multi-scale whole cell models for cellular populations of 10 billion cells or greater—a distinctly exascale computational problem—would build up from the abstract molecular networks, via molecular interaction models, through course graining of these models to mesoscale process models of the major elements of the cell.

Four primary challenges confronting the biology community are: developing integrated ecological observatories and modeling them; modeling microbial macro- and microevolutionary dynamics and processes; modeling and engineering complex multispecies biological systems for the lab environment; and developing accurate analysis and annotation of genome and metagenome sequences.

One of the challenges in microbial ecology is to understand how the actions and interactions of numerous taxa sustain a complex microbial ecosystem. This challenge is especially evident in well-mixed environments such as the ocean. Among the questions related to this central problem are: How do such communities assemble (deterministically or founder-effect dominated)? And how do communities change over time? Modeling provides testable predictions and scientists use two approaches to model ecosystems. The first involves agent-based models, where agents act as organisms. The agents are virtual organisms that live in a computing (CPU) environment and here they live, replicate, and die. The second approach is metabolic reconstruction and/or flux-based analysis of an entire community. In both cases, the models generate hypotheses, the hypotheses can be critically tested, and the results incorporated back into the models to iteratively improve them.

New modeling innovations moving this field ahead can be divided into two categories: *ecological* and *computational* drivers. Ecologically, scientists need to comprehend the relevant measures of the organisms—genomic, metabolic, and proteomic. Computational drivers are needed to create better models of organisms that react to changes in the local environment, including complex models based on the genomic information in an ecological system. These models should include a linkage between chemical, environmental, geographical, and physical data.

Coupling models of terrestrial and marine microbial communities with the exchanges in the geosphere and with relevant atmospheric and oceanographic ecosystem processes is a multi-disciplinary, multi-scale problem that also requires advances in the underlying scientific understanding. However, the construction of integrated models is critical to exploring parameters and advancing our understanding of global climate, and the global carbon, nitrogen and phosphorus cycles.

The second major challenge here is modeling microbial macro- and microevolutionary dynamics and processes that can enhance scientists' understanding of how diversity in the biosphere reflects evolutionary processes. While engineering systems require a working knowledge of our actions, biological systems are self-replicating. Thus, researchers cannot model biological systems using the same techniques used to model chemical and physical systems, such as partial differential equations.

A third significant challenge is the modeling and engineering of complex multispecies biological systems for the lab and environment. This effort is likely to involve bioreactor dynamics, host-viral interactions, host-predator processes, and process manipulation, and will also include an associated evaluation of theories, such as community assembly, functional redundancy, and parameter sweeps.

The fourth challenge is developing accurate annotation and analysis of genes in genomes and metagenomes. A number of sequence similarity-based tools perform functional assignments in current annotation pipelines. Consequently, the functions of new proteins are extrapolations of existing annotations. As a result, there are two interrelated parts to the problem of accurate gene annotation: accurately annotating a set of genes or genomes and extending the annotation to new genes, genomes, and metagenomes.

Chapter 4 focuses on research efforts to model the brain in order to probe how it functions as a biological system, including the ambitious *Blue Brain Project's* simulations at the human brain scale. This chapter also includes discussion of synthetic cognition and neuromimetic systems as well as the value of lessons to be learned from figuring out how the brain manages and uses knowledge.

This rapidly-morphing subject matter represents not only an exascale computational challenge in the context of tissues, organs, and physiology modeling, but reverse engineering the brain, specifically, is a leading Grand Challenge of the National Academy of Engineering. Current modeling and simulation capabilities provide only a local view of individual processes of this complex system without interaction between modalities and scales—essential for a comprehensive understanding of an organ system.

Chapter 5 explains the various new data, image, and visual analysis techniques that enable researchers to understand massive biological databases that are more often than not multidisciplinary. We are in an era of pervasive parallelism. As the number of transistors doubles, the number of cores will double. This means that software of the future will be very different from the sequential programs that scientists use today.

Most problems in biology involve analyzing large quantities of data and/or many loosely coupled, "multitask" computations. These problems require more than a good message-passing interface (MPI) or an MPI/OpenMPI implementation and support for parallel I/O. Experiments and simulations now routinely produce petascale datasets, a prelude to the even larger, extreme-scale datasets that will be common in the future. The mere fact that scientists have data is not sufficient for analysis; it must become knowledge to be of any real value. Scientific data analysis

and representation are central to this transformation – critical links in the chain of knowledge acquisition.

A well-developed "Summary of Technical Recommendations" is proffered that recapitulates the findings above into pragmatic steps forward. We then discuss the research areas with the most activity and promise, illustrating the immediate need for extreme-scale capabilities. The ability and opportunities of computing at this scale offer not only the DOE but the nation and society as a whole unprecedented resources to understand biology from single cells to complex cellular communities to whole organ and even higher brain function. Meeting these grand challenges in biology cannot be accomplished by one individual or even by one field of research but require collective experimental programs that are coupled to a more powerful computing sciences.

FOCUSED RECOMMENDATIONS

Exascale computing capabilities open up considerable possibilities for addressing the complexity of biological systems and offer the promise of moving towards a predictive science of systems biology. However, to leverage this capability the biological sciences community will need to embark on a push for next generation codes and tools explicitly targeting the requirements of exascale systems. A key observation is that exascale computing capabilities will for the first time enable a truly multiscale attack on biological problems, coupling molecular level details to cellular networks, to populations and communities and coupling multiple communities and ecosystems to the global biosphere. Here are four areas where significant impact could be expected during the next decade.

- *Simulating the mechanical, electronic, and chemical mechanisms of biomolecular complexes, DNA-protein, RNA-protein, DNA, RNA and other biopolymers, and protein-protein interactions.* Development of scalable tools that will admit a variety of time, space, and trajectory sampling methods to fully exploit the hundreds of million of cores expected on exascale computing platforms. This next-generation integrated code suite will enable longtime integrations, implicit solvation conditions, and mixed molecular mechanics and quantum mechanics models and exploit the expected architectural features of exascale systems.
- *Building models of cells and organisms in the context of evolution (large-scale analysis and reconstruction of cellular networks).* Integrate genome analysis tools with modeling environments, high-throughput computing frameworks, and the availability of mixed integer optimization tools in a massive parallel environment.
- *Building multi-scale whole cell models for cellular populations of 10 billion cells or greater.* Develop models and simulation tools that enable the bottom up construction of populations and communities, and the interactions between modalities and scales.
- *Integrating improved biogeosphere processes with existing high-resolution climate models.* Construction of integrated models that enable the exploration of parameters and provide the coupling to large-scale global climate simulations.

Chapter 1: Macromolecular Proteins and Protein Complexes

Chapter One *opens this assessment exploring exascale opportunities in macromolecular proteins and protein complexes, covering four areas of biological systems and problems, in terms of their increasing complexity of temporal and spatial dimensions. It also examines the substantive technical gaps in achieving exascale biomolecular modeling, with a focus upon threshold capabilities and major technical gaps.*

Macromolecular complexes are elementary in all biological systems and encompass the translational and transcriptional machinery, such as the ribosome, RNA polymerase, and DNA polymerase. Already, researchers have achieved impressive methodological advances that permit them to model the largest assemblies in the cell, for instance, the ribosome, membranes and membrane proteins, chromatophores in photosynthetic bacteria, and small viruses, for short times. Systems biologists have also successfully coupled genomic and other "omics" (such as transcriptomics, proteomics, metabolomics, and interactomics) data to map cellular networks and predict their functional states.

Unfortunately, even with these advances, researchers cannot model the length scale of an entire cell—or even a small bacterium—in reasonable time periods on current leadership computer systems. For example, using proteomics and imaging data, scientists are studying the reaction-diffusion kinetics inside a full-scale model of an *E. coli* cell. The model includes approximately 20,000 ribosomes together with approximately 1 million other protein:protein and protein:nucleic acid complexes. On a 1-teraflop platform, investigators can calculate only a few coupled biochemical pathways for each 24 hours of computing time. On an exascale machine, it is estimated that they could model a large biochemical network within a full-scale model of a eukaryotic cell in one hour.

Exascale computing would therefore enable new scientific discoveries about macromolecular complexes. Understanding these complexes is crucial to the development of antibiotics that counter biothreats including anthrax and plague, because they target the ribosome. Being able to simulate the kinetics of cells is also essential for bioenergy applications that include cellulase complexes and the cellulosome.

But before researchers can fully exploit the computational power offered by exascale computing, several technical challenges must be met. New models will be essential. Current analytic models cannot adequately analyze the dynamics of complex living systems. Improving these models requires considerable advances in computational power and techniques. Current models are unlikely to scale to the size of a single cell, even a small bacterium, for relevant times, such as minutes or hours. If we want to perform simulations at longer times, we must turn to coarse-grained models that permit scaling up of macromolecular pathways and supramolecular cellular processes. Scientists also must develop new methods. Doing so will require, for example, a focus on force fields and enhanced sampling techniques.

This chapter explores four topics in terms of their importance, the factors currently limiting advancement, and the way in which exascale computing can lead to new scientific discoveries (see Figure 1.1):

1. Macromolecular folding
2. Biochemical binding and reaction mechanisms, such as enzyme catalysis and protein/ligand interactions
3. Macromolecular pathways, including DNA replication and repair fidelity, protein synthesis, chromatin organization, and RNA editing
4. Supramolecular cellular processes, such as protein signaling networks, plant cell-wall formation, and endocytosis

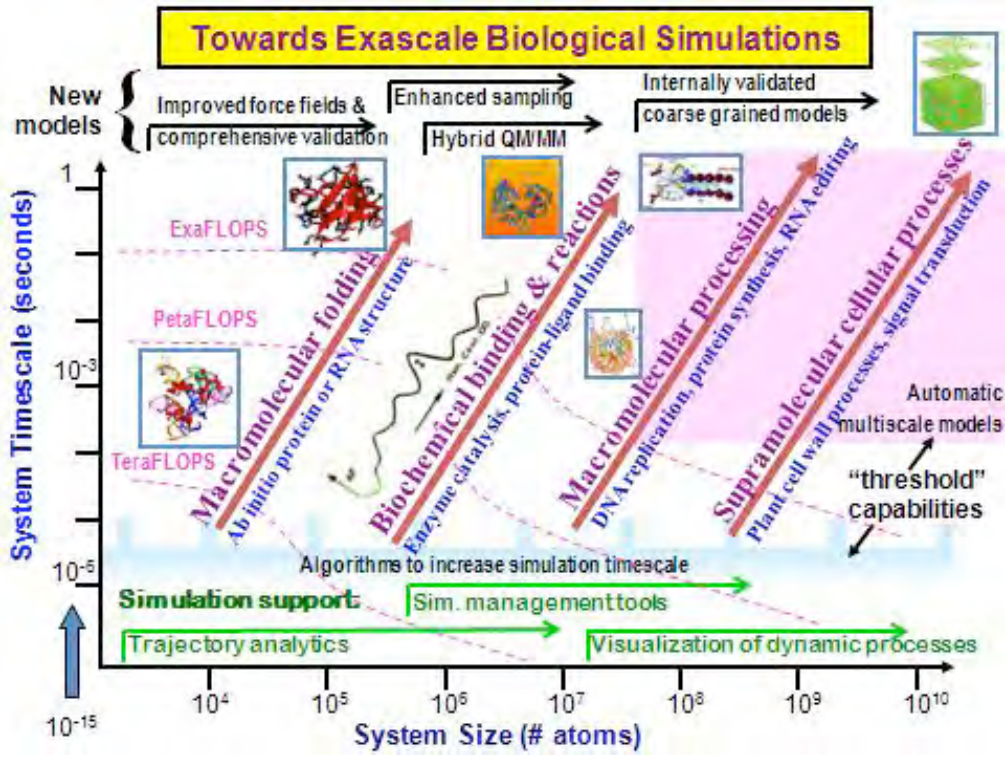


Figure 1.1. Four areas benefitting from exascale computing

1.1 Macromolecular Folding

Almost all proteins and many other biological macromolecules are active only after they adopt a specific, three-dimensional structure. Using high-speed genome sequencing, scientists are beginning to identify protein structures, but the rate of sequencing still outstrips scientists' ability to elucidate important structural information by several orders of magnitude.

One of the great promises of biomolecular simulations is to map sequences to structure to function by using simulations of macromolecular folding. Understanding how ligands bind to targets and how enzymes catalyze reactions is critical to achieving further advances in molecular biology and the energy biosciences. The development of a working model of protein functional

surface homology would be another important advance in the study of protein interactions. With such a model, scientists could take the structure of a protein and identify regions that match functional surfaces of well-characterized proteins. Further, scientists would be able to identify binding sites and suggest their function. All of these advances, however, face key computational challenges.

Coarse-Grained Models. Ideally, one would like to explore structural changes such as macromolecular folding at the atomistic level. Because such folding mechanism involves large numbers of degrees of freedom, however, traditional methods such as molecular dynamics are infeasible at this time. One alternative to all-atom simulations is the use of coarse-grained techniques.

Unlike models of a decade ago, current coarse-grained models are far more realistic. While simplistic, they provide at least two benefits. They can reproduce the essential features of protein folding and give an idea of the organizing principles regulating macromolecular processes.

Recently, scientists have been exploring strategies to use coarse-graining in conjunction with multiscaling methods. Multiscaling uses the results of atomistic simulations to develop coarse-grained models at larger scales. This is a way to create a seamless connection between the molecular, cellular, and supracellular worlds. It permits researchers to simulate long time-scale and length-scale processes. The challenge in multiscaling is to identify the appropriate coarse-grained particles and functional interaction forms that provide the crucial information required. Specifically, at least two issues must be addressed: At what level of resolution can researchers match functionally relevant details? And at what level can researchers distinguish the fine-scale aspects of function, such as specificity?

A related challenge is to determine how to decouple processes on different length- and time-scales efficiently. Among the methods proposed are: a “protein ensemble method” whereby one zooms in on interesting regions and develops large ensembles based on one initial protein representation; an adaptive multiscale method in which particles are freely exchanged between regions having different resolutions; and a “model hopping” method that switches between different levels of detail during a simulation.

Force fields. The force field is considered a cornerstone in macromolecular modeling. Not unsurprising, then, scientists have developed numerous methods for addressing force fields. Yet the development of accurate force field techniques remains problematic. For example, a recent study using multi-microsecond simulations with state of-the-art force fields to fold a small β-sheet protein failed to produce folded structures with any β-sheet content. Careful analysis indicated that the problem was in the force field, which was biased toward α-helical structures.

Among the most popular means for describing a macromolecular reaction has been a combined approach in which quantum mechanical methods are used to handle the atoms in the active site, while classical molecular modeling methods handle those outside the active site. Another approach is quantum mechanics/free energy. In both cases the challenge is how to handle the link between the two regions. The QM/MM approach has been limited principally to minimization, while the QM/FE approach has not allowed full coupling between the QM atoms and the environment.

An exciting new area of research has been in polarizable force fields, which have been called one of the most significant developments in the next-generation force fields used in biomolecular simulations. The principal challenges here have been the cost of computing the electrostatic energy and force of a polarizable model, but recent work has suggested techniques for addressing this problem. For example, researchers at the University of Illinois have devised and

implemented a noniterative method based on an approximation to the electrostatic potential energy that is suitable for long time simulations.

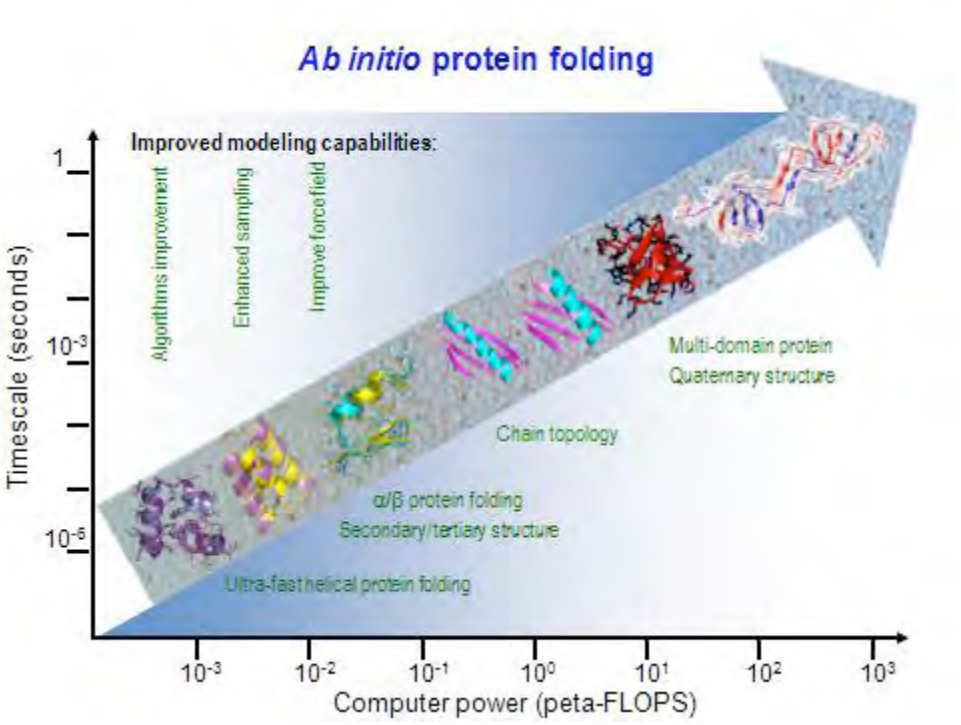


Figure 1.2. Predicted advances in protein folding with high-performance computing

1.2 Biochemical Binding and Reactions

Understanding how ligands bind to targets and how enzymes catalyze reactions is critical to achieving further advances in molecular biology and the energy biosciences. Here again, innovative computational techniques are essential if scientists are to determine the mechanisms that enzymes use to catalyze reactions.

Sampling Techniques. Both enzyme reaction calculations and ligand binding also involve extensive sampling methodologies. These methodologies can be applied in parallel to discover reaction pathways and to sample free energy profiles as a function of the pathways found. To simulate the full range of biological timescales—extending from picoseconds to years—scientists will need simulation methods to sample a conformational space framework. Such methods will need to go beyond simple brute-force, long-timescale calculations of the classical motions at the atomistic scale. To tackle the computations of free energies, transition rates, and statistical populations of states, researchers must use mathematical formulations that break computations into smaller parts that are not strongly coupled to one another. Among the possible formulations are the “string method” for refining conformational transition pathways, umbrella sampling, replica-exchange Monte Carlo, thermodynamic free energy perturbation, and steered MD coupled to coarse variables. Although researchers have begun to develop these frameworks for use on the most powerful machines, the frameworks have not yet been run on these machines. Such methodologies thus lend themselves naturally to exascale computing, which is likely to

improve the accuracy of enzyme reaction and ligand binding calculations dramatically, enhancing drug and inhibitor discovery and enabling rational, structure-based protein engineering. See Figure 1.3.

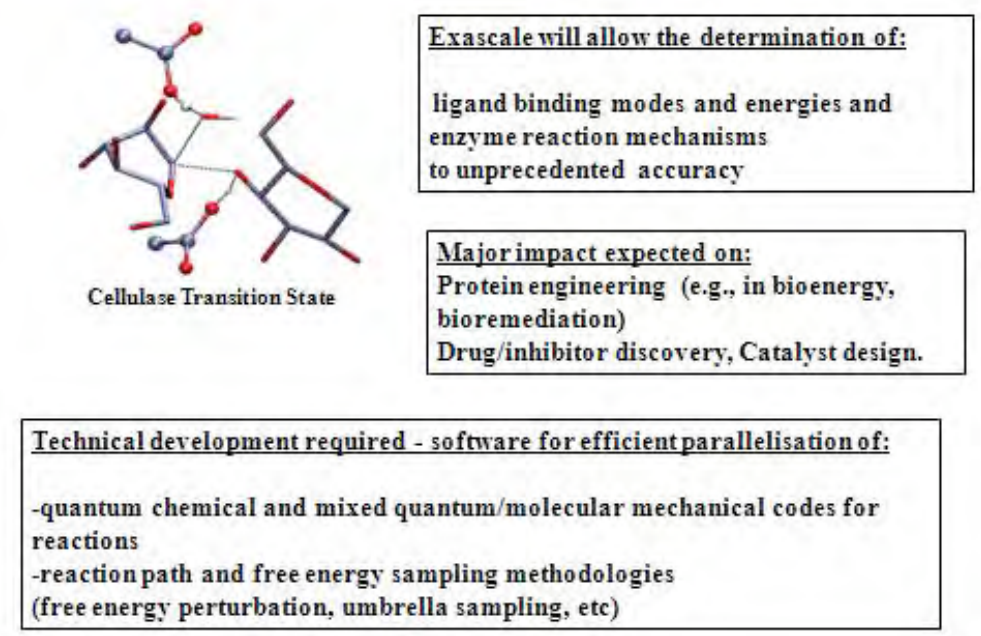


Figure 1.3. Impact in biochemical binding with exascale computing

1.3 Biomolecular Machines and Pathways

Macromolecular complexes are fundamental in all biological systems and encompass the translational and transcriptional machinery, such as the ribosome, RNA polymerase, and DNA polymerase. These systems are crucial to the development of antibiotics that counter biothreats including anthrax and plague, because they target the ribosome. The systems are also essential for bioenergy applications that include cellulase complexes and the cellulosome.

Several technical gaps must be addressed, however, before such applications can be realized. Many of these gaps are similar to those facing advances in biochemical binding (Section 1.2).

Visualizing the dynamics of a macromolecular complex in atomic detail is one of the transformative challenges of molecular biology, biochemistry, biophysics, and structural biology. Over the past three decades, major efforts have been made to simulate a variety of biomolecular systems in such detail. Nevertheless, the current state of the art still falls short of achieving physiologically relevant time scales for studying macromolecular complexes. Visualization is discussed in more detail in Chapter 5.

Scaling is another challenge**.** Today, most molecular dynamics simulation techniques use fast Fourier transforms to treat long-ranged electrostatic forces. But it is difficult to scale this algorithm to large numbers of processor cores in the range of 10,000 to 100,000. Some success for weakly charged systems has been reported for generalized reaction field potentials, but these neglect long-range interactions beyond a certain distance.

Validation also remains a substantial issue for several reasons. Force fields and time scales may be validated by comparing simulations against experiments for amino acids, peptides,

proteins, and macromolecular complexes. Researchers can also validate coarse-grained models by testing them against explicit solvent molecular dynamics simulations on shorter time scales. Further, the convergence of the simulation must be tested to ensure simulation times are long enough for the system in question and provides for sufficient sampling.

1.4 Supramolecular Cellular Processes

Supramolecular complexes comprise proteins linked to other binding partners forming interaction networks. These complexes are transient structures whose dynamic assembly and disassembly inside living cells is only beginning to be understood. Supramolecular structures dedicated to signal transduction exist in the plasma membrane. The importance of such structures for cellular signaling and disease has been demonstrated by experiments involving, for example, cholesterol depletion. Supramolecular assemblies also open new possibilities for the extraction of heavy elements from spent nuclear fuels.

To identify the design principles of supramolecular cellular processes, scientists at the University of Florida have recently developed techniques to quantify the binding kinetics of specific proteins inside focal adhesions. These techniques combine high-resolution, *in situ* molecular imaging with mathematical modeling of transport and reaction. Nevertheless, the correct representation of supramolecular machines is still out of reach of both current simulation methods and computers.

Coarse-grained models of the chemical kinetics within the cell are needed to predict and control the time-varying responses of living cells. One of the technical challenges here is to connect cellular processes and energy networks to systems biology models. Another technical challenge is the legitimate decoupling of complex systems, as shown in Figure 1.4.

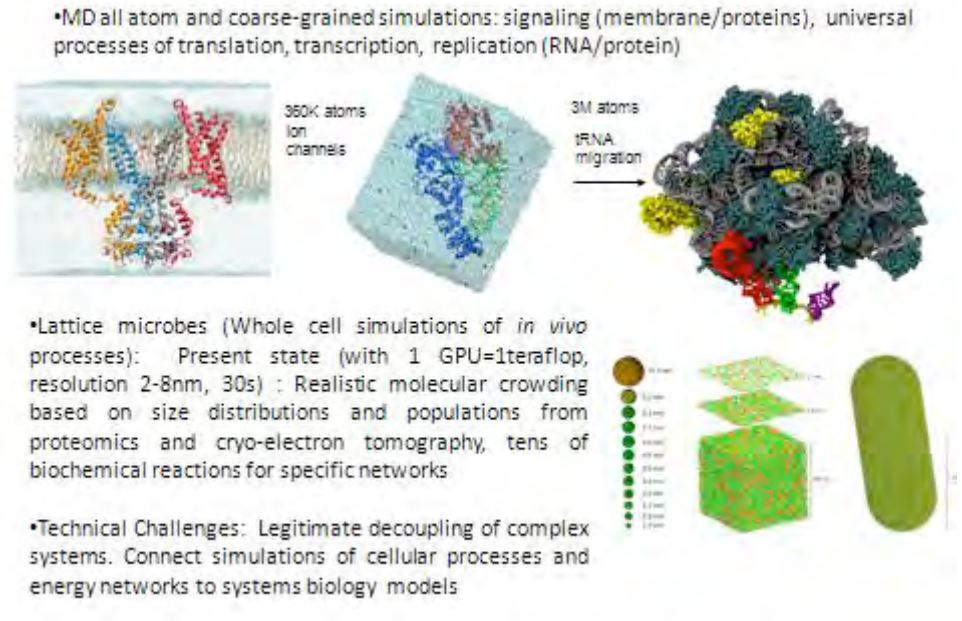


Figure 1.4. Connecting simulations of cellular processes and energy networks to systems biology models.

Visualization algorithms pose another technical challenge. Such algorithms must fully leverage both multicore CPUs (expected to be typical of exascale computers) and accelerator

GPUs (graphics processing unit). This capability will enable scientists to extract low-level properties of the structure and use them to dynamically update graphical representations (e.g., the ring structure of the carbohydrates).

1.5 The Road to Transformational Science

The different biological areas discussed in this chapter have underscored the need for new technical capabilities that, once solved, will transform exascale simulation of macromolecular processes.

Force-Field Quality and Validation. There exists considerable quantitative experimental data to validate the atomic models of systems that are either simple or of moderate complexity. Examples of work with such systems are the solvation free energy of small molecules (few thousands of molecules) and the conformational propensity of short polypeptides (< 20 residues). At this time, only a fraction of this information has been used to validate the current models. In the future, improvements to such models may permit them to use polarizable or nonpolarizable force fields; see Section 1.1 for a discussion of progress in force fields, and Section 1.3 for a discussion of the need for improved force field validation in biomolecular complexes.

For larger biological systems, scientists have less experimental data on quantitative rates and relative conformational populations that are necessary for force field validation. They are performing more experiments to amass data on the dynamics of biomolecular systems that might be used to validate new force fields. Here, molecular dynamics can play an important role but could require tens of teraflops of computer power for several months. Quantitative rates and relative populations are not well known. Sometimes, even mechanisms are unknown and computations can help calculations converge toward the correct mechanism by proposing a smaller number of possibilities that need to be tested. For example, scientists know the rate between all the steps of the duty cycle of a P-type ion pump from experiment results. Consequently, producing a quantitative movie of all those steps that are in agreement with experimental results could have a great impact. Doing so would require an estimated 35 to 75 teraflops for several months.

Sampling Methodologies. Most analyses of biomolecular molecular dynamics simulations involve a combination of calculating structural and dynamical properties averaged over the trajectory, combined with viewing movies of the results while monitoring structural properties such as specific, interatomic distances. This approach works well when the research question can be easily related to a specific structural change and the simulation does not take a long time, that is, when the trajectory is relatively homogeneous in time and space. By contrast, with multi-microsecond simulations of complex biomolecular systems, many changes in the structure, dynamics, and state of different components can occur. Current approaches involve a time-consuming, labor-intensive process to determine whether such changes are due to errors in the simulations (e.g., an anomalous precipitation of counter ions) or reveal true, but unanticipated, properties of biomolecular systems (e.g., a residual structure in unfolded proteins). In the near future, scientists will need new sampling techniques that can handle multi-microsecond simulations of macromolecular processes. For example, the development of parallel umbrella sampling and free-energy sampling methodologies could have a major impact in scientists' understanding of biochemical reactions (see Section 1.2).

Automated Annotation and Interpretation of Large-Scale Simulations. In the past, simulations were conducted on a cluster or supercomputing resource. These were downloaded to

a scientist's local facility and examined by using fairly ubiquitous algorithms, for instance to study a solvent accessible surface area. Today, the process is reversed. A large collection of trajectories is uploaded to a supercomputer for analysis using novel algorithms and strategies to handle the scale of thousands of proteins, or tens or hundreds of thousands of trajectories. With this change, computing time for analysis must be added to or displace the time used for simulation.

The scientific challenges for large-scale, high-throughput modeling projects include deciding which techniques to use, identifying individual trajectories in chemistry or biochemistry, and determining how biomolecules behave. Molecular dynamics trajectories of complex biomolecular systems would partition the system into its different chemical components (solvent, counter ions, and biomolecules; see Section 1.4) and then annotate the trajectory in several ways. First, it would estimate the inherently dynamic properties of the components (e.g., solvent viscosity, counter ion residence times, and magnitudes of biomolecular motion) and map them onto the trajectory. This step would permit researchers to match individual frames to the time-dependent behavior of the system. Additionally, this analytic system could identify spatial and temporal regions where the structural and dynamical features of each component do not match the reference values that are stored in a pre-existing knowledge base of features.

The main focus of protein biophysics has been the structure and interactions of the natively folded protein. Here, molecular dynamics techniques are useful. MD studies of protein dynamics describe the ensemble of conformations, or structural arrangements. They also have sampled, in terms of probability distributions, standard descriptors of folded proteins, such as the radius of gyration or a fractional secondary-structure composition. Structural decorrelation time and principal component analysis are examples of new measures of protein dynamics that researchers have developed.

Now, however, there is growing interest in the properties of denatured protein structures that are associated with the mechanisms of protein folding and the functioning of so-called intrinsically disordered proteins. Here, researchers need new theories and software tools to enhance the existing molecular dynamics techniques. Such tools will provide improved metrics that concisely describe the dynamics or natively structured proteins. Moreover, these tools will provide data that can help researchers design better molecular dynamics studies that can measure how efficiently different simulation lengths or the numbers of simulations imitate the sample conformational space.

Chapter 2. Pathways, Organelles, and Cells

Chapter Two *covers extreme-scale computing for pathways, organelles, and cells, emphasizing that in the postgenomic era, our goal is to understand how the one-dimensional genome translates the code of life into three-dimensional cells, including how it organizes populations of kin and nonkin cells that are communicating and differentiating. Directly addressing the challenges that exascale computing is best suited for, metabolic model reconstruction is one of the top ten problems in systems biology, and essential for scientific advances in these fields. It is also a prerequisite for any meaningful reconstruction of transcriptional regulatory networks: the metabolic model is the scaffold on which scientists can assemble and implement regulatory networks.*

A major issue in the post-genomic era is to understand how the one-dimensional genome translates the code of life into three-dimensional cells, including how it organizes populations of kin and nonkin cells that are communicating and differentiating. Macromolecular structure acts as a key intermediary between the genome and the pathways that put the genetic code into effect. The dynamic functioning of these proteins, interacting with other entities and moving actively or passively within and about the cells that encode them, creates observed behaviors. Mapping genotype to phenotype requires an understanding of how these entities function together as an integrated whole rather than as independent parts. The development of this understanding is the next natural challenge after the genome project.

Previous efforts in the sequencing and annotation of DNA are facilitating our ability improve this understanding by modeling the dynamic behavior of integrated biomolecular networks. The "pathway level" of biomolecular networks serves as a conduit between genome and behavior. Here, there is a qualitative change in the type of data that scientists can collect. Unlike sequencing, which represents basic truths because base sequences are valid, a three-dimensional structure is a "near-true" guide to the actual structure of a macromolecule. The reason is that in a molecule of interest to scientists, pathway-level data is highly conditional, often noisy, and always incomplete. The three-dimensional structure of molecular complexes can also suffer from issues that arise at the many-body pathway level.

Four significant challenges confront any efforts to analyze pathways in cells:

- T*he discovery of or inference about cellular pathways from direct and indirect data.* This challenge can range from determining all the genome regions that affect a given phenotype without asserting any causality in pathway structure, to constructing causal and mechanistic models of dynamic function.
- *The prediction of the global behaviors of a biological system based on often-incomplete knowledge of the molecular components that comprise the system.* Methods to predict behavior from molecular-level data underlie all the challenges that extend beyond inference. They are the key enablers for the experiment/theory/compute cycle.
- *The determination and/or control of the external structures that must be placed in a cell (perhaps, even in a cell that exists in a complex environment) in order to create a desired function.* This might include the remediation of a contaminated site or ways to produce a

desired, naturally synthesized molecule. In both cases, a model developed to address the first two challenges might be used to determine which external structures are feasible.

- *The design of new cellular functions.* Cells have not evolved to serve the goals of biotechnology. Consequently, it is likely that scientists will consider ways to modify cells to meet specific needs. In order to produce a well-known chemical, it may be necessary to determine the functions that can be transplanted into a given host. This modification could be done to support agricultural or environmental balance or to produce a desired material. One question this approach poses is how to create a predictable system in a target organism that can affect its behavior in a predictable way but be guaranteed to be safe.

Computational challenges are related to clarifying the role of pathways and how they are organized into the dynamic operation of organelles and cells. At the pathway level, scientists operate under uncertainty about all aspects of the system—from boundary conditions to inputs and outputs and the natural operating environment of a system to its structure, precise mechanisms, and the dynamics of its internal operation. Work is needed in three areas: (1) minimizing uncertainty, (2) propagating uncertainty in models of cellular processes and how they are used to predict function, and (3) developing modeling and simulation frameworks capable of capturing experimentally measurable system dynamics and understanding the implications of approximation on the simulated dynamics.

Scientific research is a discovery process: Not only must the reliability of the simulation execution be ensured, but the model representation and the associated distortion of reality implicit in simulations must also be minimized. At present, these goals can be accomplished only through parallel investments in experimentation as computational simulation is brought to the exascale,

Experiments need to explore how pathways operate and how they are organized in the dynamic operation of organelles and cells. This knowledge must be tightly coupled to computational developments. For example, molecular physics is not studied through experiments on thermally equilibrated ensembles; rather, technology to achieve near-zero temperatures, isolation in molecular beams, and monochromatic laser pumps/probes had to be developed before calculations could be interpreted meaningfully.

Just as progress in physics has been partially predicated on designs for experiments that meet theoretical understanding and computational tractability, likewise, biological experiments must to be tailored to complement the available capabilities of computation.

2.1 Grand Challenges

Combining high-performance computing with larger-scale experimentation will provide an opportunity for scientists to understand a whole-life form and its interaction with its environment. Scientists will then be able to place a life form in its evolutionary context, a key step toward understanding how the life form and its parts arose from their original, inorganic components. This knowledge could lead to a discovery of the cellular behaviors that could be exploited for the benefit of humans. Thus, by harnessing this knowledge for the benefit of mankind, success in this area can have a profound impact on our way of life.

Challenge 1. Model Phenotype from Genotype across the Tree of Life

In this case, the rapid reconstruction of cellular networks across multiple time and space scales will permit scientists to "crack the signaling code" that characterizes large communities of microorganisms. This will also improve scientists' understanding of how evolution continually shapes and reshapes cellular networks to establish niches in new environments. In addition, scientists might also infer the biogeochemical features that characterize an environment from the genomes of organisms that live there. Advances in this area may also allow scientists to predict the metabolism and optimal growth conditions for any microbe, once they know its sequence. Ultimately, scientists could optimize the natural system activity of individuals and consortia based on small differences in genomic composition.

Challenge 2. Discover and Engineer Units of Biological Function at Network Scale

Scientists will discover the principles of cellular network architecture and perhaps even how such architectures facilitate and lead to the observed patterns of evolutionary inheritance and change. This may be accomplished by the inference of conserved modules of function found in a variety of configurations in existing organisms. The diversity of relationships with other parts of the organism, the various adaptations these subsystems have for the various niches in which they are placed, and the combinations of such modules that seem to be selected in varying environments will provide scientists with greater insight into how to engineer these.

2.2 Factors Impeding the Prediction, Control, and Design of Cellular Behavior

Data Accuracy. A lack of data is the main issue delaying the translation of genotype into phenotype through a better knowledge of pathways. The easiest problem to address may be the annotation of genomes to determine the molecular functions they encode; yet this is still extremely challenging. Genomes have a good deal of "dark matter." Studies are annotating an average of only 60 to 70 percent of the open-reading frames.[3] Many of these annotations are vague, misleading, or simply wrong. Even under the best conditions, scientists are making educated guesses when they create annotations using homology/phylogeny. The quality of the annotation depends on the amount of sequence in the family of that open-reading frame (ORF), the variability of that family, the amount of primary data about the structure and function of members of that family, and the evolutionary distance of the given ORF to the nearest well-annotated member of that family.

The function of genes is important in understanding cellular behavior, not only because of the unknown functions and networks formed by mysterious genes, but also because, when

[3] "Regions of DNA that encode proteins are first transcribed into messenger RNA and then translated into protein. By examining the DNA sequence alone we can determine the sequence of amino acids that will appear in the final protein. In translation codons of three nucleotides determine which amino acid will be added next in the growing protein chain. It is important then to decide which nucleotide to start translation, and when to stop; this is called an open reading frame. Once a gene has been sequenced it is important to determine the correct open reading frame (ORF). Every region of DNA has six possible reading frames, three in each direction. The reading frame that is used determines which amino acids will be encoded by a gene. Typically only one reading frame is used in translating a gene (in eukaryotes), and this is often the longest open reading frame. Once the open reading frame is known the DNA sequence can be translated into its corresponding amino acid sequence. An open reading frame starts with an atg (Met) in most species and ends with a stop codon (taa, tag or tga)." University of Wisconsin System, Board of Regents, "Translation and Open Reading Frame Search," http://bioweb.uwlax.edu/genweb/molecular/seq_anal/translation/translation.html

expressed, these genes impact other cellular systems. This may only occur when they draw small amounts of resource for their own expression, at the cost of others. The genetic composition of a bacterial strain isolated from its environment represents just one instance of these species. Genomes are dynamic entities that are constantly mutating, rearranging themselves—even in the laboratory—and exchanging DNA with the outside environment. These changes affect pathway function and the key phenotypes and fitness of an organism as it operates in its environment. Clearly, challenges at the pathway level are deeply connected to those at the macromolecular and environmental/evolutionary level.

Experimental efforts to measure the functions of genes, especially ones of unknown function, are very challenging. The resolution of these measurements can vary. At one end of the spectrum are specific assays for chemical activity and structural studies. These are difficult to apply to the 50 to 200 genes that are not yet annotated in each new organism. At the other end of the spectrum are options that have special data or measurement requirements. For instance, "guilt by association" studies infer the function of an unknown gene from its physical association on the genome (or from an interaction assay) or from its similarity of expression or phenotype in knockout/overexpression studies under a variety of conditions. These studies require either large sequence databases (for the former association) or large-scale, functional, genomic measurements recorded under a broad range of conditions. Besides experimental costs, there are computational costs associated with the use of clustering algorithms and the statistical analysis of their significance. The computer memory and processing time required for this type of analysis scale with the amount of available data by a power of 2–3 depending upon the clustering algorithm used.

Analysis. Analyzing the available data can be demanding because it is heterogeneous in type, amount, error modeling, and physical meaning. Different approaches to inference and prediction are required for different questions. These can involve the collection of assorted data. The actual prediction of the three-dimensional biophysical functions of a cellular pathway in single cells is at one end of the scale. At the other end is the prediction of positive or negative population growth with a specific feedstock in a well-stirred reactor. The former requires exceptionally detailed and accurate biophysical measurement of the cell. Experiments at this level are not yet common, and special expertise is required to obtain the highest-quality data. Experimental design is also not standardized. It needs to be specially organized for each subsystem that scientists investigate. Indeed, the creation of functioning, noninvasive/nonperturbing fluorescent reporters is particular to every molecule and perhaps every condition.

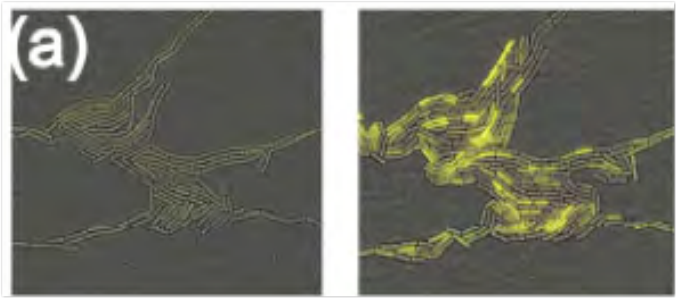


Figure 2.1 Time-lapse microscopy of microcolony growth in *B. subtilis*. The cells have fluorescent markers in a communication system (RapA/phRA, in yellow) and a sporulation commitment (spoIIA). The population shows heterogeneous development with a subpopulation expressing RapA/phrA, continuing to divide, and only sporulating after the other subpopulation has completed its sporulation.

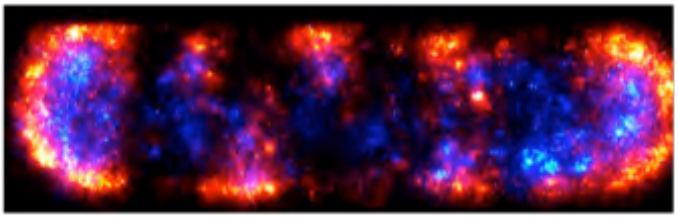


Figure 2.2 Single-cell and single molecular analyses shed light on key microbial processes that are missed when researchers use bulk measurements of expression and fitness. These analyses lack the whole-genome nature of other measurement approaches. New measurement techniques are needed to further our understanding of the function and structure of pathways. Today, the computational tools needed to extract information rapidly from these images and movies are not very sophisticated. As a result, the analysis requires extremely large computational runs. PALM image of Tar receptor in *E. coli* that was used to infer key principles of organization function of this key bacterial sensing system for chemotaxis. (Courtesy of Jan Liphardt, University of California, Berkeley)

Detailed image analysis, statistics of extracted trajectories, and time-series analyses to infer dynamical pathway function are also involved. In many cases, scientists must perform such tasks many times to estimate the optimal parameters for an algorithm.

Community Databases. Not only is the analysis of the data difficult, but turning "one-off" experiments into a community resource, such as the Systems Biology Knowledgebase[4], is not trivial. The interchange and annotation of such data types are a challenge. In addition, the comparison and clustering of movies and images are a developing area with many computational challenges. Chapter 5 discusses these challenges in greater depth.

Data Simulation. The modeling of processes that extend from cell growth and communication to the formation of macromolecular superstructure is at a very early stage. This includes modeling macromolecular structures' mechanical interactions with membranes. As a consequence, it is not possible to simulate multiscale, multiphysical, and multiresolution models, even with extreme-scale computers. The task is particularly difficult when the uncertainty in model structure and parameters needs to be integrated into the predictions such models can make.

2.3 Challenges for Exascale Computing

We address here several biological challenges that exascale computing can address effectively.

2.3.1 Metabolic Model Reconstruction

Metabolic model reconstruction is one of the top ten problems in systems biology, metabolic engineering, and biomedicine. Indeed, metabolic model reconstruction is essential for scientific advances in these fields. It is also a prerequisite for any meaningful reconstruction of transcriptional regulatory networks: *The metabolic model is the scaffold on which scientists can assemble and implement regulatory networks.*

Arguably, there exist well-established, experimental methods to *produce data* for the validation of metabolic models and regulatory networks. These include biological phenotype arrays, gene knockout experiments, mRNA microarrays, mRNA sequencing, C labeling

[4] The DOE's Systems Biology Knowledgebase (Kbase) is an emerging open cyberinfrastructure that integrates systems biology data, analytical software, and computational tools. As a unified framework linking otherwise disparate systems, Kbase will accelerate biological discovery in support of DOE missions and provide insights and benefits that can ultimately serve numerous application areas.

experiments,[5] and ChIP-on-chip.[6] Petascale computing is essential, however, if scientists are to *cope with all of the new data.* Petaflop computing power will speed the automated reconstruction and stewardship of a new generation of metabolic models. These models will include microbial communities and plants. They will be used to study metabolic flux elucidation (using labeled substrates); large-scale, isotope, metabolic reconstructions; and the *de novo* design of microbes for targeted purposes.

At the present time, however, few metabolic network models and optimization codes can use 1,000, 10,000 or 100,000 computing cores. Substantial efforts are needed to re-engineer existing codes and derive new ones that can take advantage of more substantial computing power.

To this end, investigators have recently developed software to reconstruct, optimize, and analyze genome-scale metabolic models on a massive scale using the Blue Gene/P system.[7] These scientists are creating new algorithms that will use parallel, mixed-integer, linear-programming optimization engines. They are also developing stochastic simulation software that will run on massively parallel architectures. This software will perform parameter scans and simulate large-scale biological systems, which will require 10,000 or more CPUs.

More such efforts are needed. Biological systems are too diverse, dynamic, and adaptable to understand through experimental observation alone. Although it is possible to use experiments to measure the instantaneous concentration of an enzyme or a metabolite in a cell, these concentrations change so quickly that they are difficult to measure. Computational models are necessary to discover the principles that govern the behavior of such systems. Once these principles are known, scientists will be able to rely on a small number of measurements to determine the long-term behavior of a system. In short, researchers have no alternative but to use high-performance computing in biology if the field is to continue to advance.

The missing pieces needed to advance metabolic model reconstruction are:

- Algorithms and software to apply HPC to solving extremely large-scale optimization problems with 10^6 variables and constraints. These algorithms will be valuable to optimize biological models and systems so that they fit experimental data. A number of areas of biology can benefit from this software, such as annotation, genome-scale reconstruction, and dynamic modeling. The development of this software is probably low risk and of medium cost and importance.
- Algorithms and software to predict protein structure, protein-ligand binding energies, and enzyme kinetic constants from protein sequences. Such software will be far more efficient and effective than measuring these parameters experimentally. This effort is also

[5] T. Szyperski, "Biosynthetically Directed Fractional 13C-Labeling of Proteinogenic Acids: An Efficient Analytical Tool to Investigate Intermediary Metabolism," *European Journal of Biochemistry*, vol. 232, no. 2, pp. 433–448. September 1, 1995. http://www.ncbi.nlm.nih.gov/pubmed/7556192

[6] "ChIP-on-chip, also known as genome-wide location analysis, is a technique for isolation and identification of the DNA sequences occupied by specific DNA binding proteins in cells. These binding sites may indicate functions of various transcriptional regulators and help identify their target genes during animal development and disease progression. The identified binding sites may also be used as a basis for annotating functional elements in genomes. The types of functional elements that one can identify using ChIP-on-chip include promoters, enhancers, repressor and silencing elements, insulators, boundary elements, and sequences that control DNA replication." Chiponchip.org, http://www.chiponchip.org/

[7] C. Henry, F. Xia, and R. Stevens, "Application of High-Performance Computing to the Reconstruction, Analysis, and Optimization of Genome-scale Metabolic Models," *Journal of Physics***,** Conference Series, vol. 180, 2009**.**

high risk because it is unclear whether such progress is possible, given researchers' present understanding of the problem.

- Algorithms and software to apply HPC to massively scaled, dynamic simulations of biological systems. At the present time, no software can partition a large-scale simulation of a biological system among 10,000 or more processors. This software will be essential if scientists are to model biological systems beyond a minimal level of detail, size, and complexity. Its development is probably low risk and low cost.

2.3.2 Prediction of Feasible Parameter Values for Dynamic Models of Metabolism

Accurate, dynamic modeling of single-cell and multicell systems is one of the main long-term objectives of computational biology. Such models, if truly predictive, would enable scientists to design organisms that would perform a variety of tasks. The models might also contribute to the development of treatments for emerging types of infections.

Stoichiometric models of metabolism play a key role in systems biology. Many labs around the world have built these models, but their use has proven difficult. Linear, convex, and mixed-integer optimization problems often arise when the models are used. Some new advances suggest the possibility for further innovations. They include stoichiometric modeling of nonmetabolic networks for signaling, regulation, and macromolecular synthesis.[8] For example, researchers have developed a macromolecular synthesis model for *E. coli* that includes approximately 400 genes, 10,000 model components, and 14,000 model reactions.[9] And recently, an integration of macromolecular synthesis with metabolism resulted in a rectangular matrix, a stoichiometric matrix that represents the functions of almost 2,000 *E. coli* genes, more than 60,000 components, and about 80,000 model reactions.[10]

Yet this stoichiometric matrix does not scale well. This means that entries in the matrix, that is, the stoichiometric coefficients, are distributed over four orders of magnitude, while the corresponding variables or reaction fluxes vary over 7 orders of magnitude. This poor ability to scale is a more general problem common to many biological systems. It occurs because (1) many metabolic precursors, such as nucleotides and amino acids, must combine to form a RNA species or a protein; and (2) metabolic reactions occur much faster than macromolecular reactions. For example, the solution to a linear optimization problem for a typical metabolic model takes less than 1 second of compute time on a standard PC (Dual core, Quad CPU, 2.83 GHz, 8 GB RAM, 32 bit). To reach an optimal linear programming (LP) solution for an integrated model of *E. coli's* metabolic and macromolecular synthesis network takes as much as 20 minutes of computer time on the same machine. These problems cannot be solved with standard network flow algorithms because the matrix is not completely unimodular.

Scientists need to develop new algorithms to improve the accuracy and speed needed to solve optimization problems. This can be done by (1) improving preprocessing and/or scaling routines, (2) improving the solvers that employ stoichiometric matrices to define the topology of a hypergraph

[8] A. M. Feist, M. J. Herrgard, I. Thiele, J. L. Reed, and B. O. Palsson, "Reconstruction of Biochemical Networks in Microorganisms," *Nature Reviews*, vol. 7, pp. 129–143. 2009.

[9] I. Thiele, N. Jamshidi, R. M. Fleming, and B. O. Palsson, "Genome-Scale Transcriptional and Translational Machinery: A Knowledge Base, Its Mathematical Formulation, and Its Functional Characterization," *PLoS Computational Biology*, vol. 5, no. 3, 2009.

[10] I. Thiele, R. M. T. Fleming, A. Bordbar, R. Que, and B.O. Palsson, "An Integrated Model of Macromolecular Synthesis and Metabolism of *Escherichia coli*," in preparation.

instead of using an arbitrary matrix, and (3) using large-scale mixed-integer linear programming and convex optimization solvers that are modified to work with stoichiometric matrices.

2.4 Future Work

Extreme-scale computing can enable key discoveries in biology. But much more than a high-performance platform is needed.

User Friendliness. The requirement to understand the details of parallel computing hardware and software prior to using such resources limits their use. Increased accessibility to the skilled domain scientist should be emphasized by software development teams with the requisite computing and domain expertise. Widely available and user-friendly modeling platforms with command-line interfaces can serve as archetypes for high-performance simulation software of more targeted simulation intent. With the appropriate job configuration middleware in place, computational biologists could remotely access parallel supercomputing resources from their desktops without explicit concern for how high-performance computing jobs are configured. NanoHUB activity[11] is another potential model where community contributions are mobilized for computation at all scales. Supercomputing experts, computational biologists, and researchers with dual bridging interests will need to work closely to implement the most standard, constraint-based modeling techniques. These are now widely used to study metabolic networks, yet they remain impractical for large networks.

Advanced Techniques. Models of multiple cells or microbial communities also must include more advanced techniques for matrix preprocessing and linear optimization algorithms. Today, the average metabolic model accounts for 1,000 to 3,000 reactions. Biologists expect this number of reactions and components to grow at the same rate as the number of cells they study. Furthermore, considering nonmetabolic functions or multicellular organism models, new efforts need to begin soon if scientists are to avoid the numerical analysis bottleneck that is likely to occur. Large-scale, multicellular models will also require faster, more efficient algorithms that can describe the properties of individual cells within the community, as well as the community's behavior.

Optimization Algorithms. Genome-scale metabolic models have had considerable impact in many areas of biotechnology and biomedicine, and more complex biological models will have a similar, if not more profound, impact. If these complex models are to be used extensively; however, they will depend on the availability and usability of computational algorithms that are able to reliably and efficiently solve the resulting optimization problems. Broader use will also depend on the ability of these algorithms to assess the significance of constraints when an objective is not known *a priori*.

Multiscale, Multiphysics Algorithms. There exists a need for fast, scalable, and easy-to-use algorithms or well-encapsulated algorithms for multiscale, multiphysical, and hybrid models of cellular function. Current workflows to simulate biological systems are too complex to scale out to the general computational biology modeling and simulation community. Inferring and simulating models of pathways and the cellular behaviors they drive require computationally intensive algorithms. Such tasks also require a significant amount of interaction with substantial amounts of genome-scale and accurate single-cell data.

Combinatorial Capability. Simulation models can currently solve many small-scale problems. If researchers want to use them to handle more than a few dozen variables, however, they must

[11] Osamu Tabata, "Introduction of MEMS Activity at Nano/Micro System Engineering Lab," http://nanohub.org/resources/3243/

develop new algorithms. Even steady-state, metabolic flux models do not have the ability to explore combinatorial inputs and knockouts while also calculating flux and growth, because they cannot scale to this level.

Data Exploration. Cellular behavior responds to complex combinations of inputs. It is combinatorially affected by large numbers of genes. Thus, any exploration of models or testing them against data remains difficult. If a new effort is to have a profound impact on biology, scientists will need high-throughput, easy-to-use methods that present data concisely. They will also need data, models, and an ability to compare the quality of predictions to the data.

Data Testing. Data remains a central bottleneck in the inference and testing of models. Also needed are experiments that can scale alongside computational resources. The larger and more complex the model, the greater the number and variety of experiments that are needed to infer and/or parameterize it or to test its predictions. Researchers also need to understand that data can be "fragile" and must be of a rigorously controlled quality, instantly accessible, and well annotated. Computation should be complementary to experimentation, so that models begin to test the consistency of data generated by many labs in real time. At present, computational throughput is too slow and complicated. Thus, models are not as useful as a knockout or microarray.

Access to Exascale. Fast and universal access to exascale computing from a researcher's benchtop should be viewed as a core enabling technology. Delivering data, models, and the means to use them to a larger population of scientists has the possibility of creating a "network effect." This could improve both data and models much faster than current trends. Open libraries and testing frameworks for new algorithms should underpin this effort. New investments in rigorously designed experiments are needed if scientists are to succeed in this area, as is providing processed data to the larger community.

Artificial Intelligence. Investment in tying the substantial knowledge around artificial intelligence (AI) to the specific problems of functional and topological inference should be considered. The current process of one-gene, one-experiment to validate gene function is simply not scalable to the entire tree of life if we foresee scientific progress on a timescale of decades. Current computational approaches to biological inference are fairly limited, and a greater interaction between the computational biology and AI communities could be fruitful. To accomplish such advances, scientists may need to create "gold standard" data sets and models. These would be used to test algorithms and computational architectures. They might also help demonstrate what it means to understand the function of a cell through the dynamic functioning of its networks

Chapter 3: Populations, Communities, Ecosystems, and Evolutionary Dynamics: Genomics and Metagenomics

Chapter Three *is an important review of the challenges presented by populations, communities, ecosystems, and evolutionary dynamics (genomics and metagenomics), and how research in each area will benefit from the exascale effort. The emphasis here is on understanding bacteria as well as microorganisms with no cell nucleus or other organelles in their cells, e.g., eukaryotes, ranging from microflagellates and fungi and to whales and trees. Ultimately this research will inform our understanding of ecosystems and how humans are integrated into these systems.*

The emphasis in this chapter is on understanding bacteria as well as microorganisms with no cell nucleus or other organelles in their cells. We begin with eukaryotes[12]. Ultimately, this research will inform our understanding of ecosystems and how humans are integrated into these systems.

Four primary challenges confronting the biology community in order to gain such an understanding are:

- Developing integrated ecological observatories and modeling
- Modeling microbial macro- and microevolutionary dynamics and processes
- Modeling and engineering complex multispecies biological systems for the lab and environment
- Developing accurate analysis and annotation of genome and metagenome

Before reviewing each of these challenges, we note that much of the difficulty in analyzing and understanding these systems lies not only in their common features, such as large volumes of data and a wide range of time and spatial scales, but also in the diversity of elements in the systems. For example, not only do environments typically include a large number of species, but there is also diversity among the individuals that constitute the species that scientists have not yet defined. We also note that each challenge might benefit from being integrated with other areas such as cellular modeling, molecular structure and functional relationships, geology, oceanography, and climatology.

These scientific challenges are not insurmountable. They do, however, remind us that advances are likely only when science approaches them as long-term goals.

3.1 Challenge 1: Developing Integrated Ecological Observatories and Modeling

Scientists model ecosystems because one of the challenges in microbial ecology is to understand how the actions and interactions of numerous taxa sustain a complex microbial ecosystem. This challenge is especially evident in well-mixed environments, such as the ocean. Among the questions related to this central problem are the following: How do such

[12] "Animal cells are typical of the *eukaryotic cell* [emphasis added], enclosed by a plasma membrane and containing a membrane-bound nucleus and organelles. Unlike the eukaryotic cells of plants and fungi, animal cells do not have a cell wall. This feature was lost in the distant past by the single-celled organisms that gave rise to the kingdom Animalia." "Animal Cell Structure," http://micro.magnet.fsu.edu/cells/animalcell.html

communities assemble (deterministically or founder-effect dominated); and how do communities change over time?

3.1.1 Traditional Modeling

Scientists can sample a wide range of features to describe ecological systems. These include the genomes and metagenomes of the organisms that are present, rainfall, temperature, water salinity, total biomass of organisms, nutrient levels, and wind patterns. Scientists can reduce an ecosystem to a set of variables that describe a dynamic system. Sets of mathematical functions describe the variables in such a system. Each function represents the relationship between these variables.

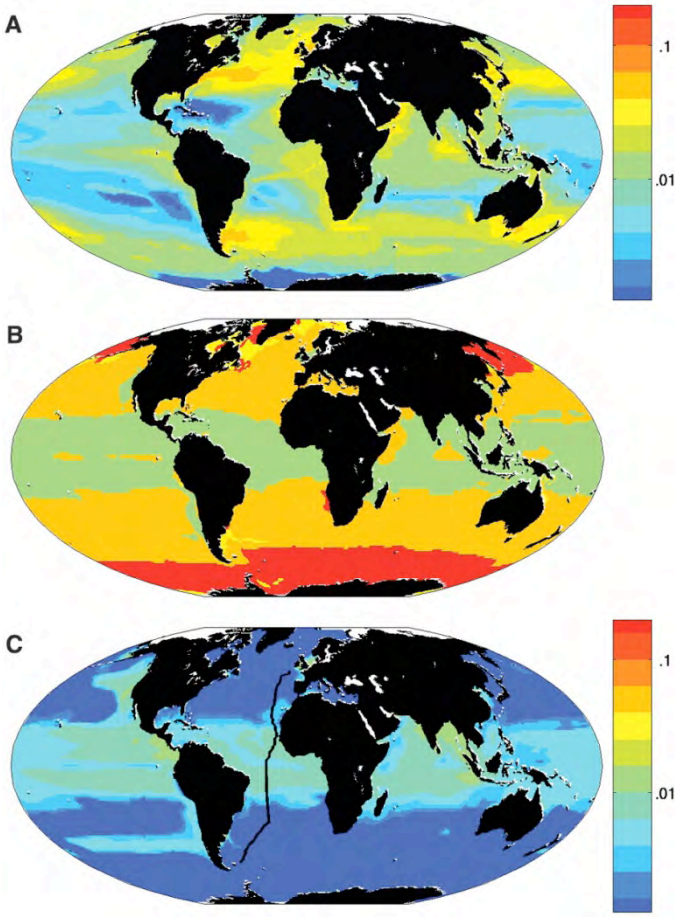


Figure 3.1. Annual mean biomass and biogeography from a single integration.
(A) Total phytoplankton biomass (µM P, 0 to 50 m average).
(B) Emergent biogeography: Researchers categorized modeled photoautotrophs into four functional groups; the color-coding reflects the local group that dominates the annual mean biomass. Green represents analogs of *Prochlorococcus*[13]; orange, other small photoautotrophs; red, diatoms; and yellow, other large phytoplankton.
(C) Total biomass of *Prochlorococcus* analogs (µM P, 0 to 50 m average). The black line indicates the track of Atlantic Meridional Transect 13 (AMT13).
Source: M. J. Follows, S. Dutkiewicz, S. Grant, and S. W. Chisholm, "Emergent Biogeography of Microbial Communities in a Model Ocean," *Science*, volume 315, pp. 1843–1846, 2007.

[13] Prochlorococcus are "the world's smallest photosynthetic organisms, microbes that can turn sunlight and carbon dioxide into living biomass like plants do." They "dominate the phytoplankton of the oceans" and "play a critical role in the regulation of atmospheric carbon dioxide." Understanding the phytoplankton better "will aid studies on global climate change." So-called blueprints of these microbes offer the promise of insights into oceans." *MIT News*, August 13, 2003. http://web.mit.edu/newsoffice/2003/plankton.html

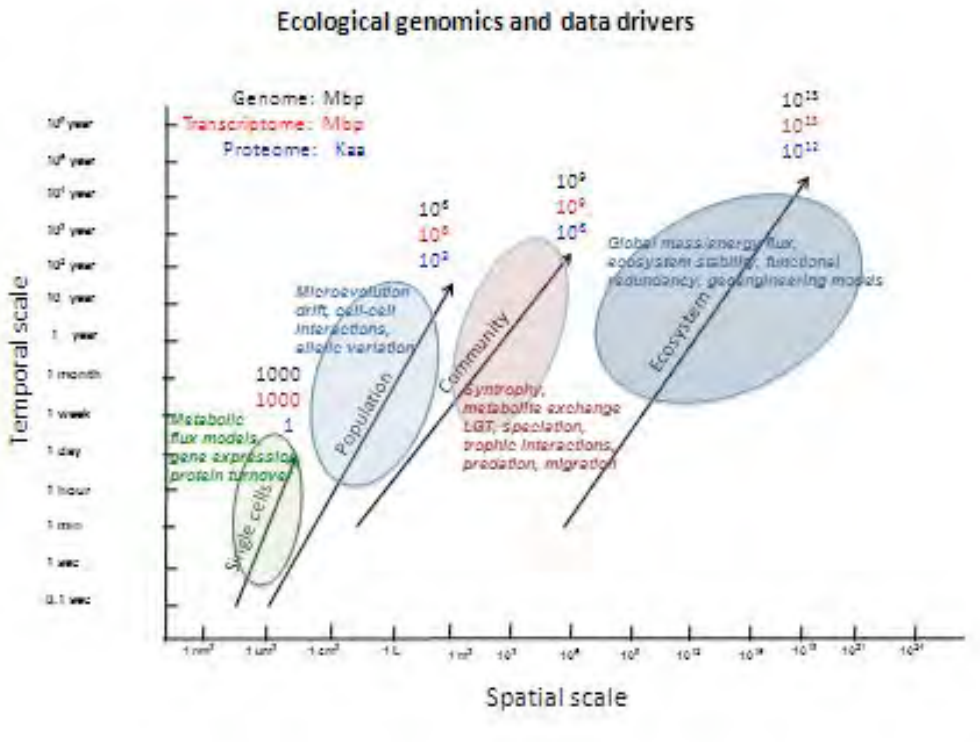


Figure 3.2. Ecological genomics and data drivers. There are different spatial and temporal scales and pertinent processes associated with each ascending level of biological complexity (cells, populations, communities, and ecosystems – see the ovals in the figure). The data presented here are conservative estimates of the number of nucleic acid base pairs, or amino acids in peptides, that occur at each level of complexity. Mbp are millions of base pairs of nucleic acids. Kaa are thousands of amino acids. (Source: DeLong, 2009, unpublished, with modifications.)

Scientists model ecosystems using two approaches. The first involves *agent-based models*, where agents act as organisms. The agents reside in CPUs; they live, replicate, and die. The second approach is *metabolic reconstruction* and/or flux-based analysis of an entire community. See Figures 3.1 and 3.2.

Currently, there are four types of research models, each severely limited in their capabilities:

- Static models, which can depict simpler environments and offer a detailed understanding of the metabolism present. These cannot vary concentrations to depict the environment's dynamics.
- Box models of an ecosystem, which treat organisms as black boxes. For example, models lump together large groups of organisms, such as microbes, phytoplankton, trees, and grasses.
- Mathematical models of environments, which are crude because scientists cannot integrate details of the substructure—pathways, the abundance of organisms, dynamics, physics, mixing, and motility.
- Flux models, which lack metabolic detail and biological realism.

Once scientists possess realistic models, however, they will be able to use simple, real-time measures of environmental parameters to evaluate and project the state of the environment. A good example is recent work that couples computational, global ocean circulation models with models of emergent biological complexity in different environmental contexts. In these images, ecological models of competitive exclusion, when overlaid on global ocean circulation models, are able to recapitulate known, global microbial distributions. As a consequence, these coupled computational models can predict global biological features successfully.

3.1.2 Innovations in Modeling

The specific advances that researchers need to achieve to move this field ahead can be divided into two categories: ecological and computational drivers.

Ecological Drivers

- Scientists need to open the black box and fill it with relevant measures of the organisms—genomic, metabolic, proteomic, and so on.
- Researchers need to use genomic data to assess the diversity of species present in an ecological niche. This includes their metabolic activities as well as their use of and production of nutrients.
- Researchers need a better understanding of organism-to-organism interactions. This encompasses cross feeding, antagonism, predation, and bottom-up versus top-down control.
- Researchers need to identify the rules and mechanisms of community assembly. For instance, what differentiates deterministic assembly from strong interactions? What distinguishes stochastic assembly from weak interactions (control) and from dispersal dominated systems?
- Researchers need better measures of metabolites, compounds, and the like in environmental samples.
- Researchers in biochemistry or annotation need a more complete understanding of metabolic fluxes in the different organisms in the environment.
- Researchers need to integrate the physical and chemical parameters that scientists measure in the environment.
- Researchers must ensure that the community includes all trophic levels, from viruses to eukaryotes.

Computational Drivers

- Researchers need to create better models of organisms because to a first Born approximation[14] the environment is static. That is to say, this is a model that reacts to changes in the local environment.
- Researchers need to calculate complex models based on the genomic information in an ecological system. These models should include chemical, environmental, geographical,

[14] "In scattering theory and, in particular in quantum mechanics, the Born approximation consists of taking the incident field in place of the total field as the driving field at each point in the scatterer. It is the perturbation method applied to scattering by an extended body. It is accurate if the scattered field is small, compared to the incident field, in the scatterer." "The Born approximation," Wikipedia, http://en.wikipedia.org/wiki/Born_approximation

and physical data. There is a need for linking between chemical, biological, and physical models.

- Researchers need better data integration and organization as well as rapid release of results.
- Researchers need to include large-scale data fusion and diverse data types.

3.1.3 Computational Challenges

The main computational challenges lie in four areas: population density, systems biology, genomics, and metabolites.

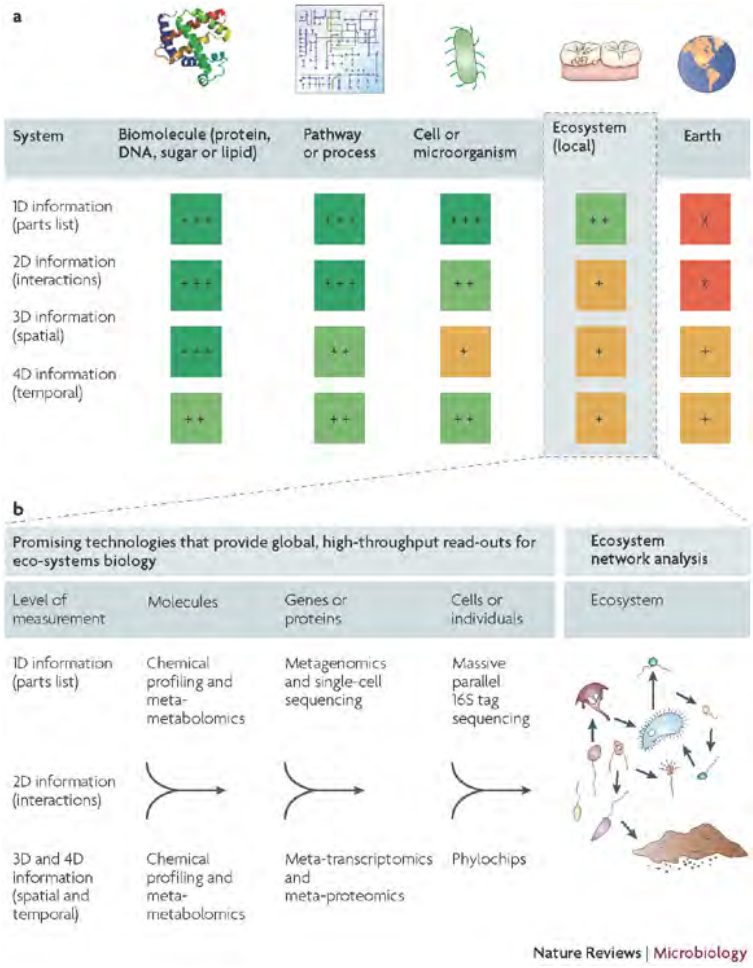


Figure 3.3 (a) The different spatial scales at which biologists can analyze systems biology (using the "dimension" definitions in C. Fraser, W. P. Hanage, B. G. Spratt, "Recombination and the Nature of Bacterial Speciation," *Science*, vol. 315, no. 5811, pp. 476–480, 2007. The columns show the data availability for each scale. The rows indicate the aspect of the system that the data targets. (1) +++ indicates ample data is available and biologists have good knowledge of the system aspect; (2) ++ indicates a number of high-throughput data sets are available and biologists have a fair knowledge of the system aspect, but more data is needed to build comprehensive models; (3) + indicates that a few, scattered, non-high-throughput data sets are available and that biologists need to restrict model building to case studies; (4) x indicates that there is almost no data available. (b) At the ecosystem level, biologists can refer to read-outs at different levels: molecules (ranging from trace elements, to small signaling

compounds, to metabolism intermediates), genes or proteins, and cells or individuals. Illustrated here are a few of the more promising high-throughput approaches that can generate data useful for ecosystems biology. At present, no high-throughput tools can map interactions. Biologists will infer this information from other data sources. (Source: J. Raes and P. Bork, "Molecular Eco-Systems Biology: Towards an Understanding of Community Function," *Nature Reviews Microbiology*, volume 6, pp. 693–699, 2009.)

Population Dynamics. Scientists use realistic models of population dynamics in a microbial ecosystem. These combine physical models of fluid flow with stochastic evolutionary dynamics. The fluid flow models may have multiple spatial scales, ranging from the cellular to the global. The evolutionary dynamics can include colonization, selection, and migration. In addition, recombination can generate new genotypes within populations. As a result, modeling a large number of recombinant genotypes could become an important way to answer basic questions facing bacterial population genetics.

Systems Biology. Biologists are moving inexorably toward systems-biology modeling at the community and ecosystems level (see Figure 3.3). This trend is happening in spite of the fact that observational, experimental, and data collection capabilities currently are inadequate for measuring biological complexity. The data and algorithm complexity that is likely to emerge over the next decade dwarfs any contemporary computational drivers.

Genomics. Scientists have used high-throughput sequencing approaches to characterize microbial ecosystems at the DNA level. This work has achieved an unprecedented level of detail. Single experiments produced gigabases of DNA data corresponding to, in one example, tens of millions of individual marker genes that describe the taxonomic makeup of a community. Newer generations of DNA sequencers hold the promise of expanded throughput, namely, the $100 human genome. It is possible this will put microbial population genomics on a new footing that could sequence thousands or hundreds of thousands of microbial genomes in each environment and/or time point. Significant increases in computing power will allow researchers to combine models of microbial metabolism, community structure, physical structure, fluid flows, and evolutionary dynamics.

Metabolites. Complete genome sequences are the raw material for understanding the metabolic role of bacteria in the environment. But scientists need new algorithms to reconstruct metabolic reactions in order to do metabolic modeling. These would be new approaches to "metabolic hole-filling." Scientists need to reduce high-resolution, cellular metabolic models that are based on complete genome sequences to a smaller list of consumed and excreted chemical classes. Once completed, this list would be included in broader ecosystem models that are tied to physical models of the environment.

The rate limiting factors are as follows:

- The need to integrate diverse data types from environmental, geospatial, organismal, temporal and sequence information for each observatory.
- The ability to assemble, annotate, and reconstruct the metagenomic samples. This depends on predictions of the metabolic outcomes of environmental perturbations on an organismal community.
- The ability to process and model these systems. For instance, as an upper bound, to model one milliliter of seawater, a researcher would need as many as 106 organisms or genomes, N^2 interactions, and tens of thousands of different substrates.

To place this in context, it takes 9000 CPU-hours to annotate the metagenomic reads from a single 454 pyrosequencer run. For 250 intensively studied biomes, with 1,000 measurements per biome, and 24 points in time per day, it would require 3.6 x10^{10} CPU-hours per day. This will require 1.3124511 x 10^{15} CPU-hours.

Without petascale computing, performing simple, preliminary analysis, such as genome assembly, gene prediction and/or annotation, and homolog detection may prove difficult. More sophisticated analyses also require greater computational resources, so they can include the detection of natural selection or rates of recombination among strains or populations.

With petascale computing, researchers will be able to create multiscale simulations. To ensure that newly combined models accurately describe natural systems, scientists will perform new computational, empirical, and theoretical studies. These will facilitate making quantitative predictions about how species levels and metabolites may change over time. In addition, researchers will be able to replicate simulations with parameter variations; these are essential for sensitivity analysis.

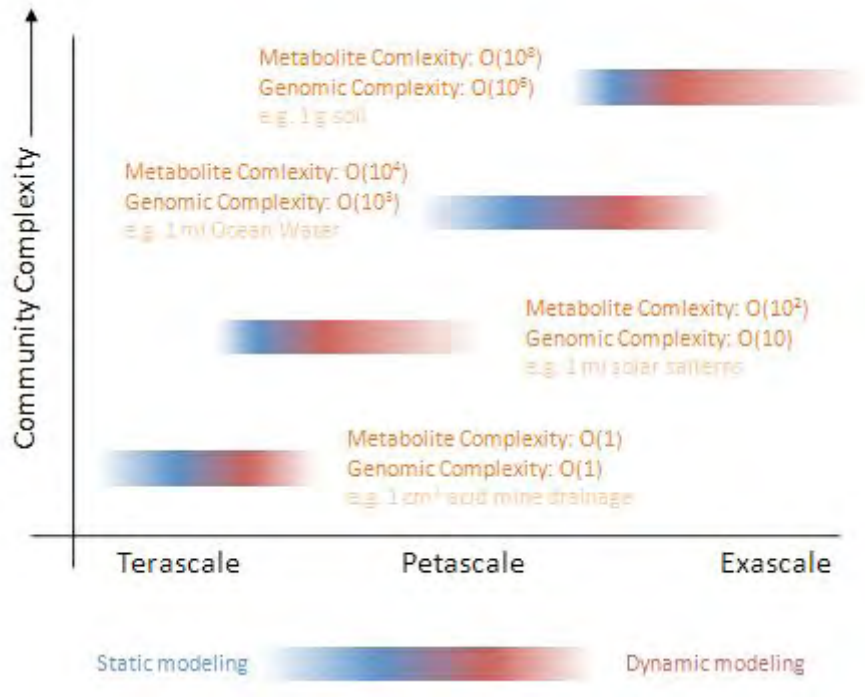


Figure 3.4. Community complexity and modeling scale. The scale of computing resources needed to capture and model the genomic complexity in various microbial communities or ecosystems. These communities can range in size from a cubic centimeter in a simple community to complex, soil, microbial communities. (Source: Rob Edwards, unpublished)

3.2 Challenge 2: Modeling Microbial Macro- and Microevolutionary Dynamics and Processes

Why should scientists study evolutionary dynamics? One reason is that such a study can enhance scientists' understanding of how diversity in the biosphere reflects evolutionary processes that have produced a wealth of biodiversity. While engineering systems require a working knowledge of our actions, biological systems are self-replicating. Thus, researchers

cannot model biological systems using the same techniques used to model chemical and physical systems, such as partial differential equations. Scientists need to understand the basic processes that underlie the evolutionary dynamics of species, populations, and clones.

Population and species evolution is not currently incorporated into broader efforts to understand natural communities. Even within genomics, it is not easy to address phenomena that vary across time and space. Consequently, researchers treat population biology and dynamics in a theoretical framework separate from species evolution.

Scientists draw upon many theories to describe microevolutionary processes, yet with few exceptions they have not validated these theories with actual data. Inexpensive, genome-sequencing technology offers a chance to test existing theories and to develop and refine quantitative models. Computing also might enable researchers to bridge this gap because it would facilitate tracking systems over long time periods, enough for significant changes to occur.

The following list summarizes the central challenges scientists face in studying environmental and host-associated bacterial communities:

- *Limited understanding of gene flow patterns and population boundaries in bacteria.* Science lacks a universally accepted species definition for animal taxa. This problem is worse for bacteria, where the rates and genetic and/or ecological bounds of gene flow are unknown. To complicate matters further, different natural populations probably occupy a continuum of rates of recombination that range from clonal to "panmictic."[15]
- *Lack of approaches that can detect recent, positive selection.* Scientists have statistical tools based on allele frequencies or haplotype structure. These can detect natural selection in sexual eukaryotes. But researchers have not studied which, if any, of these tests can be adapted for use in prokaryotic species.[16]
- *Unknown role of the "peripheral" genome.* A major part of the genetic diversity within microbial lineages resides in the "peripheral" or "flexible" genome. These strains are nearly identical in nucleotide sequence at orthologous loci, but they may differ by genomic islands containing megabase pairs of strain-specific DNA. Scientists do not understand how this extraordinary diversity contributes to adaptive evolution.

3.2.1 Issues in Genomics

Scientists need to address several issues before advances in genomics will occur.

Methods Development. The required computations will center on phylogenetics. If scientists use only sequence data, building sequence-based phylogenies with nonrigorous methods might require N^3 in terms of their complexity. It would be difficult to develop such phylogenies for more than 10^6 sequences because doing so requires graph analysis approaches that are not easy to parallelize. In addition, these problems have large memory requirements and must have sufficient bandwidth to that memory for compute nodes.

To study evolutionary dynamics using genomics or metagenomics data, scientists must

[15] Panmictic populations are unstructured, random-mating populations. See: Panmictic definition. http://www.biochem.northwestern.edu/holmgren/Glossary/Definitions/Def-P/panmictic.html

[16] University of Arizona, Department of Biochemistry and Molecular Biophysics, The Biology Project, "Prokaryotes, Eukaryotes & Viruses Tutorial," http://www.biology.arizona.edu/Cell_BIO/tutorials/pev/page2.html

modify existing methods or, more likely, develop new methods. Sequence data will probably be growing exponentially. Thus, the new methods will be needed in order to build phylogenies, explore their parameter space, and evaluate the results, all rigorously. The need to incorporate non-sequence-based data (such as environmental parameters) into phylogenies will greatly increase the complexity of these methods. Sequence-based phylogenies will be central to evolutionary models that can test hypotheses of genetic diversity, genome stability, and adaptation.

Organismal, Genomic, and Metagenomic Data. Scientists need to develop single-point and longitudinal measurements of organisms, genes, and genomes in environments. These must have as accurate an environmental description as is possible. Such data will be the foundation to test whether short- and long-term evolutionary models result in populations with similar structures of organism and gene abundance and in organism and gene diversity, both in type and in allelic variation.

Computational Models. Researchers will need to model microevolutionary processes that can predict speciation events and describe population dynamics. The work of R. Lenksi[17] on the laboratory evolution of communities provides a good example of the benefits of such studies. Lenski's work has helped scientists understand gene flow, migration, and dispersal. It defines the frequency and rules for horizontal gene transfer and provides insights into predictive diagnostics and parallel, cross-ecosystem, predictive, and comparative analysis. It also improves scientists' ability to test existing theories of evolutionary processes.

Computational models will also enable scientists to infer the function of genes, pathways, and cellular systems from their evolutionary history and context, as well as to infer the timing, origin, and distribution of extant protein domains and families.

Models will be used to identify new catalytic capabilities. One result will be to build a tree of life that comprehensively defines, or interrelates, all known taxa with genome dynamics and/or relationships. Computing will need to deal with triangle inequality as evident from the lack of reciprocity among pan-genomic members. Computing power will help researchers model, visualize, and describe the state and dynamic. It will also help identify the function of unknown proteins.

Computational models will not only begin to predict species abundances but also forecast the emergence of new recombinant genotypes. The input needed for this will include species abundance data and genotype and/or allele "fitness" across environments. The output required will be species' abundances at subsequent time points, including the abundance and identity of recombinant genotypes.

In addition to all these predictive capabilities, computational models will be used to test theories or compare them. Scientists have developed mutually incompatible theories to describe microevolutionary processes. However, most of this work did not use genomic data from real populations. These models can critically test existing theories that include the ecotype hypothesis, metapopulations, and the neutral theory of biodiversity. For example, Fraser et al.[18] used such models to study population sizes of 10^5. The researchers estimated the changes due to

[17] T. F. Cooper, D. E. Rozen, and R. E. Lenski, "Parallel Changes in Gene Expression after 20,000 Generations of Evolution in *E. coli*," in *Proceedings of the National Academy of Sciences, USA*, volume 100, pp. 1072–1077, 2003.

[18] Christophe Fraser, William P. Hanage, and Brian G. Spratt, "Recombination and the Nature of Bacterial Speciation," *Science*, vol. 315, no. 5811, pp. 476–480. January 26, 2007. On multilocus sequence typing (MLST), see C. Fraser, W. P. Hanage, and B. G. Spratt, "Neutral Microepidemic Evolution of Bacterial Pathogens," *Proceedings of the National Academy of Sciences, USA*, vol. 102, no. 6, pp. 1968–1973, 2005.

mutations. They also included populations that were homogenous in terms of their fitness, recombination, and other characteristics.

Modeling metapopulations demands simultaneous simulations for an enormous number of microenvironments. This is difficult, even with the simplified model that Fraser et al. developed. A substantial increase in computing power will be essential in order to expand these models to cover more realistic population sizes. If this power is unavailable, model will continue to place unrealistic restrictions on the level of complexity.

Memory constraints will also be a major concern. Each recombinant genotype of interest will need to be tracked. Models might exclude some loci that are not ecologically relevant for a particular simulation or environment. But for a simple model with N diallelic loci of interest, scientists would need to consider 2^N recombinant genotypes. In addition, some alleles may have more than two forms with fitness differences.

3.2.2 Issues in the Study of Bacterial Genetics and Evolution

Research is also needed to address fundamental questions in bacterial genetics and evolution. Scientists have far more limited knowledge of microbial population genetics than of eukaryotic systems.

It is almost impossible to observe microevolutionary processes such as speciation. Scientists (see, e.g., Kondrashov[19]) have studied these processes almost exclusively by using theory and simulation. For microbial systems, the problem is more challenging because of the idiosyncratic nature of horizontal gene transfer and very large population sizes. Moreover, the microscopic scale makes it impossible to track individual organisms experimentally in the same way that metazoans are studied.

Most of traditional theory for population genetics applies to eukaryotes. Its validity in prokaryotic systems is not established. Scientists have not made a substantial effort to develop computational methods to improve their ability to run large-scale simulations. For the most part, the largest simulations that researchers have run have been very restricted. They do not include fitness differences among genotypes and have targeted individual populations within ecosystems with a carrying capacity of 10^5–10^7.

Considering the analysis of future data, if researchers want to test its consistency with theory, the following issues exist:

- Sequence alignment tools exist, but mostly for relatively simple models.
- Phylogenetic tree inference has likelihood and Bayesian implementations, but generally not with “recombinations.”
- The integration of sequence alignment tools and phylogenetic trees is even more limited.
- Tree comparison tools exist, but they do not scale to very big data sets. There are also limits to current statistical mulitree methods.

In principle, the simulations could be verified against metagenomic data. Doing so would, however, require additional computing power. Moreover, researchers would need to obtain significant ecological metadata to validate their models.

[19]A. S. Kondrashov and F. A. Kondrashov, “Interactions among Quantitative Traits in the Course of Sympatric Speciation,” *Nature*, vol. 400, pp. 351–354, 1999

3.3 Challenge 3: Modeling and Engineering Complex Multispecies Biological Systems for the Lab and Environment

Today, few model systems provide researchers with full control of all the microbial entities and environmental parameters, the nutrient input and temperature. Current systems have the following shortcomings:

- Limited experience in measuring and quantifying carbon and energy fluxes
- Very little correlation of metadata with molecular data
- Little experience in simulating natural systems
- Poor performance in doing functional prediction for natural systems
- Inability of metabolomics to be done in a high-throughput fashion
- Lack of mathematical and computational tools.

Researchers will need to improve their ability to understand, model, predict, engineer, and, when feasible, manipulate microbial multispecies communities. This effort is likely to involve bioreactor dynamics, host-viral interactions, host-predator processes, and process manipulation. It will also include an associated evaluation of theories, such as community assembly, functional redundancy, and sweeps. This work will rely on real-time sequencing, omics,[20] and a parallelization of manipulations. In addition, new methodologies will be needed in order to include nutrient responses, nutrient limitations, and community dynamics recorded in laboratory model systems, as well as experimental manipulation in the environment.

Research might begin with simulations based on laboratory systems at a low level of complexity. Work could then progress to interspecies systems, including metabolic modeling and modeling of environmental interactions. The outputs would include predictive, multiscale simulation models of low- to medium-complexity microbial communities and complex models of simple communities. A major goal of this work is likely to be the definition and modeling of energy and carbon flux through these communities.

Several key areas of interest to DOE thus would benefit from this new knowledge:

- Carbon sequestration and cycling (global climate change)
- Bioremediation
- Corrosion and materials decay

3.3.1 Issues

In terms of technology engineering needs, a key requirement will be accurate annotation across the protein universe. To understand the omics data, scientists will need an accurate and constantly updated annotation of proteins. The annotation will need to include the inference of function, annotation, signatures, and unknown vs. known proteins. Scientists will also need to integrate different types of data such as metadata, high-throughput genomics, and other omics

[20] Omics includes genomics (the quantitative study of genes, regulatory and non-coding sequences), transcriptomics (RNA and gene expression), proteomics (protein expression), metabolomics (metabolites and metabolic networks), pharmacogenomics (the quantitative study of how genetics affects hosts' responses to drugs), and physiomics (physiological dynamics and functions of whole organisms) and neutrigenomics. “OMICS: A Journal of Integrative Biology.” http://www.liebertpub.com/products/product.aspx?pid=43

data. Their major concerns will include management, quality control, and assessment of these data. If researchers are to understand how environments respond to change, they will need to incorporate real-time data of many types, ranging from phylogenetic to omics, into metadata. Once this task is done, it should facilitate environmental surveillance and responses to perturbation.

Currently, scientists already can predict community metabolism in a laboratory's model systems and in the field, for very simple systems. Indeed, researchers have completed some community flux and balance studies and developed a few simple three-dimensional models for biofilm formation. Researchers cannot, however, predict community metabolism for complex, mixed-culture systems.

Scientists want to use multiscale modeling and computation to understand mixed systems, model them, and predict metabolic activities and fluxes. The issue is whether they will be able to do real-world estimation, control, and engineering.

3.3.2 Moving Forward

What do researchers need in order to move genomics toward its goals? First, they will need to study **model systems,** beginning with simple laboratory, multispecies models. These model settings provide a way to learn the rules and gradually increase the complexity of the bioreactors being used. Such models offer a way to deal with three or four species with defined carbon and energy levels. They also permit scientists to develop hypotheses and test them systematically. An important goal of such efforts would be to simulate an *in silico* bioreactor with full spatio-temporal resolution, such as three-dimensional lattice simulations with diffusion dynamics. For such an approach, complexity could be increased systematically by introducing additional species, environmental or temperature changes, and variations in nutrients.

Once work with simple, bioreactor and chemostat systems is completed, scientists will need to work with more complex laboratory systems. A **layered systems approach** might be beneficial because it is more complex than bioreactors, but it involves multiple species. Such model systems provide opportunities and challenges for understanding and modeling fluxes of carbon and energy. Layered communities are relevant to the U.S. Department of Energy because of their role in biofilms, biofuel production, corrosion, and bioremediation. Examples of such communities are layered microbial communities in sediment columns, such as oceans and lakes. Simulations using such systems can grow to a level of complexity that requires significant computational power.

To establish an *in silico* bioreactor and layered communities that can be simulated *in silico* with full spatio-temporal simulation, scientists will need to use well-established techniques. These might include two-dimensional or three-dimensional lattice simulations with diffusion dynamics. But they will need to be used at **unprecedented, spatio-temporal resolutions** in order to attain exascale levels. Scientists will also need to apply microscopic simulations game theory to such simulations and to simulate individual agent interactions. This work will provide quantitative and qualitative as well as spatio-temporal results that could be validated using lab experiments or observed and measured in natural environments. Chemical reaction networks might also provide greater insight into the interactions that are occurring. When such models are coupled with large-scale, multidimensional, optimization frameworks, such as simulated annealing, scientists would be able to infer the underlying reaction rates (i.e., the inverse problem) and compare them or fit them to experimental results. For large-scale simulations,

researchers might draw upon simulation techniques from astrophysics such as star or supernova simulations and adaptive grid and mesh calculations. They might also draw on fluid dynamics.

Even before researchers develop new computational methods, they can begin to learn to gather and integrate field data and how to **select sites** that are appropriate for environmental systems analysis.

Achieving these capabilities will depend on **advances in measurement and modeling** in laboratory systems. Eventually, such measurement and modeling will be possible with natural systems. Simulations of laboratory systems will allow scientists to verify whether they can predict complex interactions, as well as energy and carbon flow, in experiments.

Perhaps most significant, while initial laboratory experiments can be analyzed with desktop systems, researchers will need to turn to significant **computing power** for predictive simulations. Exascale computing will also be needed to handle the data generated by these simulations.

3.4 Challenge 4: Developing Accurate Annotation and Analysis of Genes in Genomes and Metagenomes

A number of sequence similarity-based tools perform functional assignments in current annotation pipelines. Consequently, the functions of new proteins are extrapolations of existing annotations. As a result, there are two interrelated parts to the problem of accurate gene annotation: accurately annotating a set of genes or genomes and extending the annotation to new genes, genomes, and metagenomes.

The better the solution to the first part of the problem, the easier will be the solution to the second one. Although annotations are specific, they are often not very sensitive. As a result, sequence similarity-based models do not cover many regions in protein sequences. The result is incomplete annotations or no annotations. Errors usually occur when there is a simplified assignment of function based on minimal information that relies on general sequence similarity. Improving sequence similarity-based models takes a long time because researchers improve models incrementally only once a year. By contrast, database size increases every day. Structural information that might significantly improve the quality of function prediction is not readily available. Consequently, the number of solved structures is increasing at a very slow rate and has not even doubled over the past 10 years.

Prediction of protein functions might be dramatically improved by the following:

- Implementing more sensitive sequence similarity searches
- Bringing structural considerations, such as fold recognition, structure prediction, ligand docking, etc., into the annotation process
- Building new and improved models by integrating sequence- and structure-based information, and using these models to incorporate error detection and correction in the annotation system.

Scientists will need to use massively parallel computing to implement each of these improvements. Researchers will need to conduct sequence- and structure-based searches often to ensure that new genomic information is incorporated into the models because the data change daily. Such searches are costly: Sequence-based improvements are likely to require hundreds of millions of processor hours, and structure-based improvements might demand billions of processor hours.

Chapter 4: Tissues, Organs, and Physiology Modeling

Chapter Four *extends this rapidly-morphing subject with a discussion of tissues, organs, and physiology modeling, starting with the Blue Brain Project's simulations at the human brain scale, initially concentrating on models of the rat, mouse, cat, and primate brain systems. This includes discussions of synthetic cognition and neuromimetic systems.*

The computational modeling of biological tissue, organs and physiological systems is an area of increasing interest, importance, and impact. Such a hierarchical structure sets higher, complex organisms apart from the simplest living creatures. Indeed, relatively subtle, functional changes within a single organ—the brain—distinguish humans from other forms of life. Although scientists have traditionally modeled these systems in exceedingly simplified form, the introduction of powerful computing platforms will enable the development of models that depict the interactions of the parts, instead of simply summarizing high-level functions.

The simplest tissues incorporate all of the ultrastructural and biochemical complexity of the cell, including the cytoskeleton and cell membrane, systems for maintaining ionic and osmotic homeostasis, linked and highly regulated pathways for metabolism and reproduction, and biochemical synthesis and repair. During development, tissues and organs add functional and structural specialization and a differentiation of cell types and self-organization.

Modeling challenges can arise from the sheer size of systems, i.e., brain models require billions to trillions of elements or more, depending on the scope and desired level of detail; or from system complexity, i.e., the presence of multiple variables, each spanning a range of spatial and temporal scales, or the combination of both. For instance, the function of neural systems depends on structural scales ranging from tens of nanometers to meters and temporal scales that are from microseconds to hours. Other related systems cover scales that extend this envelope. For example, the macromolecular function of ionic channels must be modeled on scales of angstroms and nanoseconds or picoseconds. The circulatory system is up to hundreds of kilometers in length. Developmental or pathophysiological processes can require timescales of years or decades.

Current modeling and simulation capabilities provide only a local view of individual processes, without interaction between modalities and scales. Modeling at multiple scales—from atomic through cellular—is needed to represent the various macroscopic subsystems for a comprehensive understanding of an organ system.

Detailed computational models of complex cellular systems let us ground theories of physiological function in the facts gleaned from exploring biological implementation. The granularity and resolution of the required models depend on the nature of the questions to be examined. Modern biological science is built on a wealth of data. Besides genomic, proteomic, and genetic expression data, it includes statistics about cells and about macroscopic systems that cover their anatomy, physiology, patterns of connectivity, and interaction. In addition to physiological systems' specialized functions, scientists must also take into account basic biological imperatives, such as energy metabolism and self-repair, generic functions that may be at the root of system failures and disease.

Inadequate detailed knowledge often limits the development of system models. Scientists may lack an understanding of the essential components and of the equations and parameters required to describe the function of a system. New computational models can often drive new

experiments for model building and validation and can facilitate the exploration of a parameter space that may have not be suitable for experimental study.

Three computational grand challenges illustrate the role that computational modeling can play in understanding complex physiological processes and systems:

- ***Reverse engineering of the brain:*** Bottom-up models that incorporate all available physiological detail to capture the biological function of the brain can predict the consequences of activation and pharmacological or electrophysiological intervention. Simpler, large-scale models of information encoding and processing elements, architectures, and spatiotemporal dynamics allow scientists to identify computational principles and primitives of the brain. This can lead to a synthetic cognition[21] in neuromimetic systems. Understanding the complexities of the brain is a grand challenge and many visionaries identify this as the key science opportunity of this century.

- ***Multiscale physiological models of organ systems:*** Scientists use systems of coupled models to understand the function and functional interactions between cells, organs, and physiological systems. They also use these systems to quantify how such interactions change in disease states. An international effort in the Physiome Project, a worldwide public domain effort to provide a computational framework for understanding human and other eukaryotic physiology, is developing multiscale computational models of 12 major organ systems. These models will include mechanical, biochemical, electrophysiological, and other functional characteristics, as well as the embryonic and childhood development of organ systems. They will improve diagnostic understanding and help practitioners individualize clinical practice using custom therapeutic interventions.

- ***Health risks of occupational or environmental insults:*** Some of the most dangerous modern health risks, such as cancer, are linked to byproducts of energy production. These include radiation, organic chemicals, and emerging concerns such as engineered nanoparticles. With models of organ systems, scientists will be able to explore normal functions and to describe the physical consequences of exposure as well as the long-term consequences of rare, stochastic events. Such events occur in a complex microenvironment, influenced by changeable genetic backgrounds and different selective pressures.

[21]One example is synthetic cognition systems that can emulate the functional architecture of the primate visual cortex: "By using petascale computational resources, combined with our growing knowledge of the structure and function of biological neural systems, we can match, for the first time, the size and functional complexity necessary to reproduce the information processing capabilities of cortical circuits. The arrival of next-generation supercomputers may allow us to close the performance gap between state of the art computer vision approaches by bringing these systems to the scale of the human brain." Los Alamos National Laboratory, "Synthetic Cognition," http://synthetic-cognition.lanl.gov/index.php

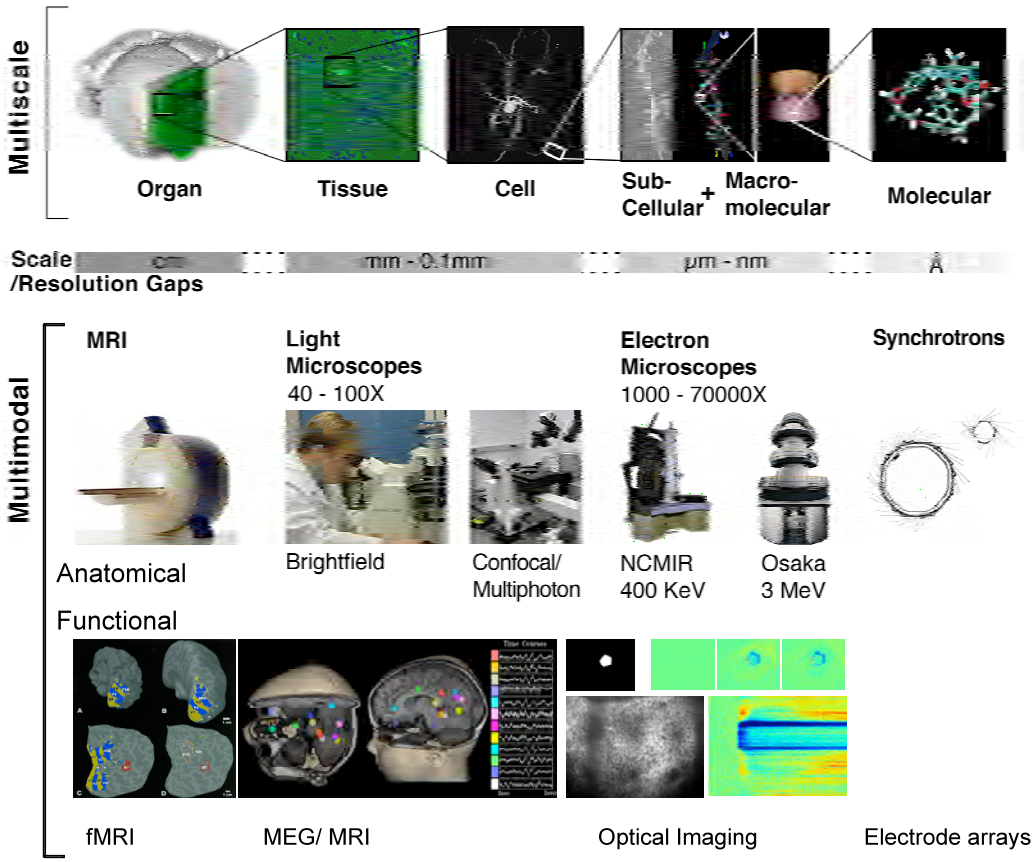


Figure 4.1 Data sources and scales for whole brain anatomical and physiological modeling

In the remainder of this chapter, we focus on three challenges in simulating the human brain. The National Academy of Engineering identified reverse engineering of the brain as one of the leading Grand Challenges <http://www.engineeringchallenges.org/cms/8996/9109.aspx>. This effort is likely to attract a wide range of professionals, including neuroscientists, physicians; psychologists; physicists; materials scientists; electrical, computer, and biomedical engineers; and computer and information scientists. Already, it has drawn substantial interest and investment within the United States and across the world. The scale of this effort will expand.

The human brain represents an incredibly complex network that if understood could be used to enhance computer chip development leading to enhanced communications and more powerful computing machines, to develop aspects of artificial intelligence, and to develop more precise methods for testing biotechnology solutions to brain disorders. In all likelihood, the effort to understand the function of the brain will define the scientific legacy of this century, just as subatomic physics, the human genome project, and the development of digital computation defined the last century.

Major hurdles remain to be surmounted to enable acceleration in the rate of progress toward more complete knowledge of brain structure and function. For example, we must find better ways to obtain large amounts of multiscale data and represent the complexity of brain networks and pathways continuously from the level of molecules to the complete functional systems of the

brain. Expanding the application of extreme scale computing to many aspects of this grand challenge without doubt deliver important collateral benefits to society.

Over the past 50 years, advances in experimental studies, computational modeling, and theory have increased our understanding of the individual processing elements (neurons) and the functions of some very simple circuits. In the past two decades, researchers have developed powerful new methods to study and manipulate the function of large neural populations. These methods cover both cellular networks and extended brain systems. Researchers have also created rich new sources of data to describe the cellular connectivity that supports these functions. In the future, however, progress is likely to be limited by the need to work with more realistic and more complex neuronal models. Such models will need to be sophisticated enough to incorporate the variety and gigantic numbers of neurons and the connections between them.

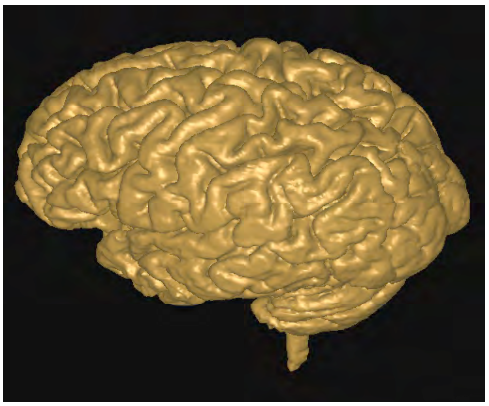


Figure 4.2 Volumetric model of the human brain based on 3D magnetic resonance imaging (MRI). Such geometric models of soft tissue anatomy define the substrate and provide an organizational framework for detailed functional neuroanatomy, circuitry and connectivity schematics. Such anatomical models enable more sophisticated analysis of data using functional neuroimaging techniques, such as electro- and magnetoencephalography (EEG and MEG), positron emission tomography, functional diffusion tensor MRI, and other emerging techniques.

The high-performance computational modeling envisioned by studying the brain is likely to have a significant impact on several areas of basic science, clinical medicine, and engineering.

This impact could even transform these disciplines through the following:

- Analyzing complex interactions within the system and predicting the consequences of intervention, for example, by pharmacological treatment or electromagnetic stimulation.
- Evaluating and interpreting experimental data from neural systems. This effort typically requires a combination of neurophysiological and biophysical modeling techniques.
- Understanding how information encoding and processing mechanisms support neural computation, thereby enabling synthetic sensory cognition and neuromimetic and neuromorphic electronic systems.

Each of these applications requires simulations that operate across several spatial and temporal scales. Neural simulations must include geometrically realistic models of neurons, neural networks, and neural tissue to predict experimentally observable responses. These responses must include individual cell responses; responses of networks of neurons that are spatially resolved; and the integrated, noninvasive responses of large neural systems. Various techniques exist to inform and evaluate such simulations. These include data from electrode

arrays or novel optical imaging techniques; noninvasive methods such as magnetoencephalography (MEG) or electroencephalography (EEG); and functional MRI.[22]

Nevertheless, researchers need better computational methods to explore the feasibility of measuring neural population response or of optimizing any proposed new methods for such measurement. These methods will need to include MRI techniques sensitive to neuronal currents[23] and optical tomography in scattering tissue. Scientists need to develop ways to simulate how artificial stimulation changes neural responses. These should include how scientists can use applied currents or magnetic fields in emerging systems for electroneural prostheses, to provide the engineering foundation for designing such systems.

4.1 Synthetic Cognition

Efforts to model the brain probe how the brain functions as a biological system. Such models may also provide insight into the computational function of the brain. Unlike invertebrate neural systems, which incorporate specialized neurons with very specific patterns of connections, the vertebrate brain tends to employ generic neurons and patterns of connectivity that are modified based on their patterns of use. Models of vertebrate systems cannot specify patterns of connectivity in detail; they must employ probabilistic, statistical descriptions of neuronal geometry and of patterns and strengths of interconnection that are a function of cellular type, geometric proximity, and patterns of activity. During development, vertebrates lay down the substrate for these extended circuits, which are fine-tuned during early experience by patterns of spatio-temporal covariance in their activity.

Artificial intelligence researchers have had some success in the study of vertebrate brain function. For example, Poggio and colleagues at MIT[24] have designed systems for machine vision that mimic the hierarchical architectures and specialized feature detectors of primate visual pathways. A team at Los Alamos is now using petascale hardware architectures to run model systems that capture the computational function of feedback, as well as feed-forward and lateral interconnections. These models obey sophisticated learning rules that exploit the information processing and encoding possibilities associated the neuronal spiking and synchronized population activity.

Because of the complexity of large neural systems, computational models offer the only path to understand collective function. Currently, however, neural modeling and neurophysiological experiments offer few tools or strategies to test and validate neural population behavior of system models. Indeed, the most successful models predict detailed patterns of firing only in individual cells or small networks.

Even these relatively simple brain models involve vast numbers of parameters and elements. Modeling human systems will need to draw on and extend techniques for sensitivity estimation, adjoint computation, large-scale optimization, and probabilistic estimation of model parameters developed for large-scale physics simulations of climate and mantle convection. New forms of data, from macroscopic physiological measurements and simultaneous measurements of many

---

[22] J. S. George et al., "Mapping Function in the Human Brain with MEG, anatomical MRI, and functional MRI," *Journal of Clinical Neurophysiology*, vol. 12, pp. 406-431, 1995.

[23] L. Heller et al., "Modeling Direct Effects of Neural Current on MRI," *Human Brain Mapping*, volume 30, pp. 1-12. 2009; see also K. Blagoev et al., "Modeling the Magnetic Signature of Neuronal Tissue," *NeuroImage*, vol. 37, pp. 137-148, 2007.

[24] T. Serre, A. Oliva, and T. Poggio, "A Feedforward Architecture Accounts for Rapid Categorization," in *Proceedings of the National Academy of Science (USA)*, vol. 104, pp. 6424–6429, 2007.

neurons in parallel, will be valuable for driving and validating model development. Such strategies will require the development of coupled biophysical models that can predict observable physical responses based on neural simulations. If scientists can begin to predict experimental observables based on neuronal network simulations, they will be able to optimize large-scale models to account for experimental data. They will also be able to conduct critical tests of these models. Further, they will be able to explore the parameter space, to characterize the response of a large system, and to trace the chain of causality that accounts for any observed response.

4.2 Neuromimetic Systems

Computational simulations, in addition to their important role in the interpretation of observational or experimental data, can be used to implement and explore neuromimetic systems.[25]

One example providing a link between sight and signal processing is the Artificial Retina Project.[26] This collaborative effort seeks to develop an implantable microelectronic retinal prosthesis that restores useful vision to people blinded by retinal diseases.

More specialized processing architectures will allow these systems to operate in demanding operational environments, such as autonomous vehicles, remote sensing platforms, and distributed surveillance environments. Drawing inspiration from biology is especially important for the design of systems for the front-end encoding, processing, and segmentation of environmental data that can achieve and eventually exceed the speed, accuracy, flexibility, and reliability of human sensory systems. Using a suitable mix of analog logic and digital communications, neuromorphic hardware may eventually achieve the remarkable performance of biological systems, with devices of comparable size, weight, and power consumption.

4.3 The Blue Brain Project

The Blue Brain Project, based in Ecole Polytechnique Fédérale de Lausanne, is an attempt to reverse engineer the brain to create a physiological simulation for biomedical applications. The Project is moving in stepwise fashion toward simulations at the human brain scale, initially concentrating on models of the rat, mouse, cat, and primate brain systems. The first phase started after 15 years of systematically dissecting the microanatomical, genetic and electrical properties of the elementary unit of the neocortex to construct a neocortical column model—the template circuit of the neocortex—which consists of 10,000 physiologically detailed, three-dimensional neurons and 30,000,000 synapses. The model was based on data from the somato-sensory cortex of a young rat and was simulated on an 8,192-core 4-rack Blue Gene/L system.

---

[25] Neuromimetic systems are electronic hardware or software systems that are inspired by or mimic the function of biological neurons or neural systems. Neuromorphic hardware goes a step further by emulating the form or functional design of such systems in silicon devices. C. Mead, "Neuromorphic Electronic Systems," in *Proceedings of the IEEE*, volume 78, pp. 1629–1636. 1990. K.A. Zaghloul and K. Boahen, "A Silicon Retina That Reproduces Signals in the Optic Nerve," *Journal of Neural Engineering*, vol. 3, pp. 257–267, 2006.

[26] Artificial Retina Project URL: http://artificialretina.energy.gov/

The present simulation facility is now being extended to support modeling of the subcellular domain, which will integrate additional levels of biological detail into the existing model. The project aims to develop a generic facility that could allow rapid modeling, simulation and experimentation of any brain regions, if the data can be measured and provided according to specifications.[27]

At the present time, researchers do not understand the brain-scale effect of any drug. But they know that the unique configuration of an individual's brain and which receptors and ion channels are coded in each individual's DNA (e.g., an individual's epigenetics) shapes the effectiveness of many pharmacological agents. Similarly, individual variations in development and experience can change the organization and functioning of the brain. If scientists can integrate genetic, functional, and structural data from an individual patient, the model brain can open the way to personalized medicine and therapeutics, while also increasing the likelihood that treatments will be successful and ameliorating negative secondary effects.

Validation. Researchers have developed a tool chain to construct, simulate, visualize, and analyze a neocortical column and used these tools to validate the model while also integrating new experimental measurements. Biological experimental data supply the key validation criteria for such simulations. The Blue Brain Project will validate all models by replicating experimental protocols and data. The range of data includes ion channels, neuron firing behavior, synapses, dendritic integration, morphological parameters, connectivity, polysynaptic loops, and emergent network activity. However, much work remains before scientists can establish community standards for verification, validation, and performance assessment of whole-brain simulation architectures. In contrast to modern physics simulations, researchers have not established precise benchmarks for valid whole-brain simulation, nor have they developed acceptable measures to characterize computational efficiency.

Computation. The algorithms and simulation architectures needed to implement simulations of the entire brain are still under development. Besides enhanced simulation engines, whole-brain simulation requires tools to integrate databases of source data with tools that can construct, simulate, and analyze models in ways that domain-experts can operate intuitively and understand. Simulating an entire human brain at the cellular level will require an estimated 1 exaflop of computational power. This whole-brain model simulation will include detail at the electrical level and will depict the dynamics of ion channels and synapses distributed across 100 billion unique neurons. Molecular-scale activity that characterizes gene expression, biochemical signaling, protein-protein interactions, the vasculature, glial cells, ion channels, receptors, and synapses will be linked to the electrical and cellular scale. Thus, the model will be able to depict the effect of pharmacological agents.

In sum, such simulations may require the coupling of large-scale, molecular-level models with brain-scale cellular models. Eventually, such efforts, coupled with genetic data, will enable the development of effective personalized medicine.

[27] Blue Brain URL: http://bluebrain.epfl.ch/ and H. Markram, "The Blue Brain Project," *Nature Reviews Neuroscience*, vol. 7, pp. 153–160, 2006.

Chapter 5: Data Analysis, Imaging, and Visualization

Chapter Five *explains the various new data, image, and visual analysis techniques that enable researchers to understand massive biological databases that are more often as not multidisciplinary. Experiments and simulations now routinely produce petascale datasets, a prelude to the even larger, extreme-scale datasets that will be common in the future. The mere fact that scientists have data is not sufficient for analysis; it must become knowledge to be of any real value. Scientific data analysis and representation are central to this transformation – critical links in the chain of knowledge acquisition.*

Humans are innately visual creatures. Indeed, we devote half of our brain to processing visual information. In computational terms, vision is our highest-bandwidth data path. It is not surprising, therefore, that data visualization[28] is a fundamental way to explore the interpretation of models and understand complex phenomena. We depend on our visual perception and cognition to detect patterns, assess situations, and rank tasks. Visual data exploration is an important way to reduce and refine data streams; it lets scientists winnow huge volumes of data into smaller data sets.

When visual data analysis is coupled with interactive interfaces, it can facilitate the detection and validation of expected results. It can also facilitate the discovery of unexpected scientific findings. Visual data analysis also allows for the validation of new theoretical models. It lets scientists compare models and datasets, do quantitative and qualitative querying, and improve the interpretation of data. Scientists use visual data analysis systems to explore "what if" scenarios, define hypotheses, and examine data from different perspectives. Visualization lets scientists identify what connections may exist between large numbers of attributes. It provides a way to assess the reliability of hypotheses and to quantify how reliable they are. In sum, visual data analysis is an integral part of scientific problem solving and discovery.

Petascale and extreme-scale computing and data acquisition from high-bandwidth experiments across the sciences are changing the data analysis environment. Our ability to produce data is rapidly outstripping our ability to use it. As Herbert Simon, the Nobel Laureate in economics, has noted, "A wealth of information creates a poverty of attention and a need to allocate it efficiently."[29] This statement summarizes the issue with petascale and extreme-scale datasets; we have far more data than we can explore and analyze in a lifetime with our current tools.

5.1 Research Areas

Data analysis, imaging, and visualization research is important to the success of extreme-scale biological science. Fundamental advances are needed if scientists are to be able to analyze

[28]Parts of this section of the report are from R. Johnson, R. Ross, S. Ahern, J. Ahrens, W. Bethel, K. L. Ma, M. Papka, J. van Rosendale, H. W. Shen, and J. Thomas, "Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale," Department of Energy Office of Advanced Scientific Computing Research, 2007. http://www.sci.utah.edu/vaw2007/DOE-Visualization-Report-2007.pdf

[29]Herbert Simon, *Computers, Communications and the Public Interest*, Martin Greenberger, ed., Baltimore: The Johns Hopkins Press, 1971, pp. 40–41.

and understand large, complex biological datasets that will result from experiments and from petascale and extreme-scale simulations.

Algorithms. Scientists need effective data analysis and visualization tools to support predictive simulations and scientific discovery. These tools should be based on strong algorithmic and mathematical foundations. New mathematical methods will help improve the scientific understanding of images. They will allow scientists to characterize salient features in their data reliably. Advances in optimization, topology, and uncertainty modeling will be critical for analyzing extreme-scale datasets. Because of the complexity of biological networks, scientists will need to develop the high-dimensional mathematics required to model and accurately predict behavior in large-scale, distributed networks that can evolve over time.

Topological Methods. Topological methods have the power to describe complex shapes at multiple scales. As a result, they are becoming increasingly important in the development of advanced data analysis. The introduction of robust combinatorial techniques for topological analysis has accelerated the use of topology to present known phenomena and to detect and quantify new features.

Probability Analysis. Today's data analysis capabilities lag far behind researchers' ability to produce data from simulations or observational records. While scalable, parallel analysis methods work across specific applications, scientists have no generalized tools for this purpose. Browsing or looking at data is not possible when scientists are working with extreme-scale data. Hence, there exists an enormous need for methods that can be used to analyze, organize, and present data dynamically across all application domains. One promising approach is to use high-dimensional probability distributions of quantities of interest. This approach will require new contributions from mathematics, probability, and statistics.

Feature Detection and Tracking. Scaling simulations to ever-finer granularity and time steps presents new challenges in visualizing the data produced. Scientists need to create smart, semi-automated visualization algorithms and methodologies to help filter the data or present summary visualizations. Such algorithms would allow scientists to characterize salient features in their data reliably. They also would permit scientists to analyze the immense data by following a more top-down, methodological path. Additionally, scientists need formal, semantic schemas, taxonomies, and ontologies to describe, characterize, and quantify features. Such schemas and taxonomies might also highlight areas of interest in massive, time-varying data, indicating where to look or to make additional, high-level queries. Feature-based techniques are also important for analyzing the results of different simulations. They can help scientists compare simulations and experimental data. Once scientists identify and measure features and their evolution, they need tools to identify interfeature relationships and the configurations of sets of objects and their interactions.

Uncertainty Analysis and Visualization. Researchers confront a significant problem when they try to treat uncertainty in simulations robustly. Numerical simulations are rife with sources of uncertainty. Simulations introduce uncertainty through numerical imprecision, inaccuracy, or instability. Uncertainty is inherently a part of predictions and forecasting because it arises from variations in the physical processes under study. Scientific experiments and measurements introduce uncertainty as calibration errors, differences in repeated measurements, and the like. The analysis and visualization of extreme-scale biological datasets can introduce uncertainty during processing, decimation, summarization and abstraction, as artifacts of creating much-condensed representations of data. The ability to quantify uncertainty in high-performance, computational simulations will open new opportunities for the verification and validation of

simulation codes. With a robust mathematical framework to trace the sources of uncertainty and its propagation throughout the simulation process, simulation might gain a strong, predictive capability. Handling uncertainty must be an end-to-end process, where the different sources of uncertainty are identified, quantified, represented, tracked, and visualized along with the underlying data. Hence, scientists need to develop uncertainty representation and quantification, uncertainty propagation, and uncertainty visualization techniques in order to provide credible and verifiable visualizations.

Image Analysis. Biological imaging provides information at various scales ranging from the organ level to the molecular level. Scientists depend on image analysis to explore many biological problems, such as the study of phenotypes and populations. In the case of reverse engineering of the brain, imaging provides morphology and connectivity information. Scientists' ability to create extreme multiscale and multimodal image data for biological applications is growing at an extraordinary pace. For instance, by using automated image acquisition, biologists capture approximately 4,000 high-resolution electron microscopy images in 24 hours (approximately 0.1 terabyte/24 hr). Unfortunately, scientists have not developed robust image analysis techniques and tools that can scale up to the exascale level. This lag has created a bottleneck for biological research. To transform extreme-scale image data into knowledge, scientists will need new and improved approaches to data management. They will also find that it is essential to have multimodal image registration and assembly and automated segmentation and annotation. Moreover, an essential capability will be automatic proof reading of results; requiring a user to check the entire image data set for errors negates the purpose of automatic annotation. Hence, scientists need to develop techniques that can direct a user's attention to error-prone areas in a multiscale manner.

Reverse engineering of the brain is a good example of the imaging challenges. To compile all the data needed to image an entire mouse brain at electron microscopy resolution requires 30 petavoxels.[30] The management and dissemination of this data are an important challenge. Furthermore, image data at a single scale is insufficient. For instance, scientists can rely on EM to provide detailed connectivity information for relatively small volumes. They need light microscopy to trace long-range, neuronal projections and magnetic resonance imaging to obtain a more global overview. The registration of these multiple modalities is another important challenge. But perhaps the most significant challenge is automatic annotation of the features in extreme-scale image datasets. Current state-of-the-art approaches to image segmentation cannot confer the robustness and computational efficiency that is sufficient for this problem.

5.2 Complexity of Biology Datasets

Scientific simulation codes are producing data at exponentially increasing size. But size is only one of the challenges facing scientists. The complexity of the data is a major challenge.

Multimodal Data Understanding. New approaches are needed that take into account the multimodal nature of the data. Such approaches will enable scientists to blend traditional, scientific and information visualization; perform hypothesis testing, verification, and validation;

[30] A voxel is short for volume pixel, the smallest distinguishable box-shaped part of a three-dimensional image. Voxelization is the process of adding depth to an image using a set of cross-sectional images known as a volumetric dataset." *Internet.com Webopedia*, http://www.webopedia.com/TERM/V/voxel.html

and address the challenges posed by the vastly different grid types that the various elements of multimodel codes use. Visualization and analysis experts also have a critical need for tools that leverage semantic information and hide details of dataset formats. These tools can provide a way for experts to focus on the design of these approaches and not become mired in the minutia of specific data representations.

Multifield and Multiscale Analysis. Many biological and computational models try to simulate phenomena that occur over a range of spatial and temporal scales. These can span several orders of magnitude. In addition, such models can attempt to capture the interaction of multiple variables or multivariate and multifield data. Scientists use the analysis and visualization of multivariate or multiscale datasets to discover hidden relationships among the data and the transient events that occupy only a small fraction of simulation time but can have a profound influence on simulation outcomes. Arguably, current visual data analysis technologies can process many types of adaptive, multivariate, multiresolution data. But scientists need improved techniques to support zooming into regions of interest and to generate geometrical structure with a high degree of accuracy. They also need to be able to display animations that are short enough to match a viewer's desired context in a way that offers sufficient detail to describe important, transient events. For multifield data, visualization cannot simply map different variables to different visual parameters. Such an approach would lead researchers to quickly run out of visual parameters. It also introduces visual overload for the user, thus hampering the user's ability to understand the data. In sum, researchers need to draw on a range of approaches to improve multi-field and multiscale analysis, such as visual analytics, projections and dimensionality reduction, database queries, feature detection, and novel visualization techniques.

Time-Varying Datasets. One factor contributing to the exponential growth of data is the ability to run very large scale, time-varying simulations. Unfortunately, most current analytical methods do not function well with targeted, time-varying data. New techniques are needed so that scientists can browse through different spatial and temporal scales interactively. They need to visualize and identify scientific phenomena of different temporal lengths and to isolate and track salient features in both time and space. Researchers also need to integrate multiresolution spatial and temporal data management and encoding techniques with current and future visualization algorithms. Such integration would enable visualization users to see the scale and location of time-varying data in a way that is completely transparent.

5.3 Advanced Architectures

When scientists develop new computational methods, they must consider the computational platforms involved. Emerging petascale and exascale architectures are both a blessing and a curse. They offer unprecedented computational power. But they also have the ability to produce results much faster than machines can store them—and much faster than researchers can visualize them. This is nearly as dramatic a shift as the one from vector computers to distributed memory supercomputers that occurred 15 years ago. Scientists took years to adjust to that shift.

As the number of transistors doubles, the number of cores will double. AMD and Intel already have product lines that include multicore processors, and their product roadmaps foresee continued expansion of the number of cores in processors. Moreover, chip producers are combining GPUs and CPUs in order to couple applications and graphics more tightly. This technology could prove to be the biggest change in the PC platform in decades and will have an

enormous impact on graphics and visualization. In fact, the visualization pipeline we know today will be radically different if it is to exploit the new architectures.

Because graphics is the driving force behind creating such high-performance chips, it is critical that the graphics and visualization community participate actively in their development. Over the near term, researchers need to focus on the integration of the CPU and GPU and the programming models they use. Future architectures most likely will be heterogeneous and include different kinds of processors on a single die. Visualization that can utilize multicore-CPU-style, thread parallelism, and GPU-style data parallelism will play a key role in understanding the results these heterogeneous systems will produce. We can also expect to see the emergence of an entirely new class of interactive visualization applications.

5.4 Data Storage and Analysis

As processing power grows, it increases the amount of data processed and generated. Faster computation rates make it possible to run simulations at higher fidelity, which yields even more data. Unfortunately, a cost-effective storage system that tracks the rate at which it consumes additional data is not available. Thus, a divide has been created between what computers are producing and what storage is capable of maintaining. Today, it is not uncommon for researchers who run simulations to discard over 90% of what they compute. Because simulation data can no longer be stored dynamically as the simulation progresses, one needs to consider new methods to utilize the data as it is being generated.

5.4.1 *In Situ* Processing

Collocating certain visualization and analysis algorithms within simulations may improve the effectiveness of the algorithm. At the same time, it may be a way to maximize the information stored in the data. For example, saliency analysis may be able to help the simulation make better decisions about what data to store and what data to discard. Feature extraction is much more effective when scientists have all variable information available. In addition, feature tracking is more reliable when temporal fidelity is high. Features can provide an analyst with a great deal more information. They also can help an analyst reduce the storage demands far below what they were originally. If analysts want these techniques to be integrated into the application and supported by the run-time environment, they will need to press for a great deal of interaction between the designers of programming models and the developers of systems software for advanced architectures.

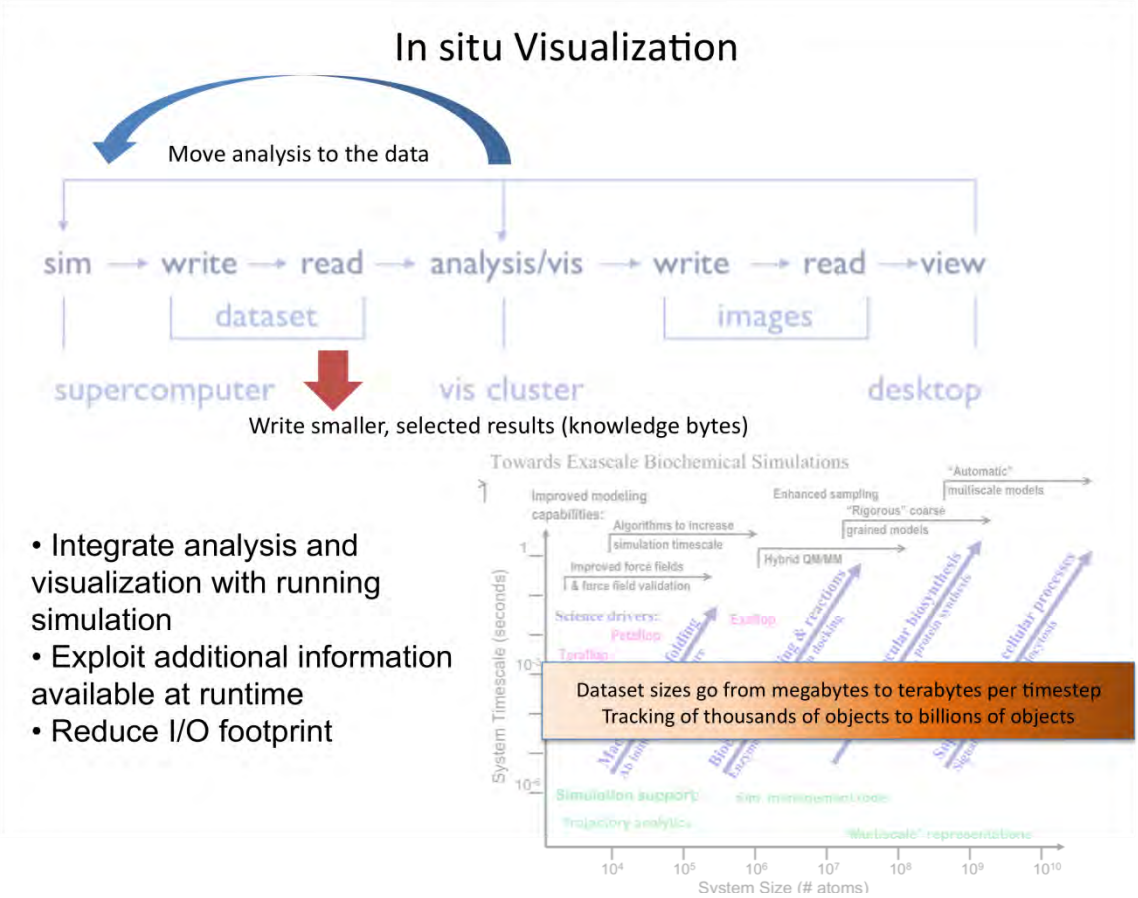


Figure 5.1. Simulation science pipeline. The figure presents the writing of data postsimulation for analysis. It also shows the generation and saving of images for visualization. *In situ* analysis moves the data analysis into the simulation. In this case, the data is processed as part of the simulation and only a small amount of more meaningful data is written to disk.

In situ processing can help scientists reduce the disparity between data generation rates and storage system capabilities. It is an important component in managing petascale and exascale datasets. Applications that will run on future systems will need to store an unprecedented amount of simulation data during their run time. Current practice relies on the postprocessing of datasets from leadership-computing applications. These are completed on separate visualization clusters. Most likely, this approach will not be suitable for analysis at the petascale level. Certainly, the approach will not work at the exascale level. Research into what other mechanisms might be used to process large datasets is critical for visualization at these scales. Approaches might include out-of-core mechanisms and streaming models of processing. In all likelihood, researchers will need to use these approaches in conjunction with *in situ* processing.

5.4.2 Data Format

Data models and formats are an important issue for applications. Decisions about defining these models and formats affect the ability of scientists to describe the results of their work. Such decisions also affect the efficiency with which data is moved to storage and subsequently

processed. The explosion of data formats and models present in the DOE application space is hampering scientists' ability to generalize tools for visualization and analysis. The use of multiple formats and models in applications that combine simulation and other data sources or that leverage coupled codes exacerbates the situation. The disconnect between data models used in simulation codes and subsequent postprocessing access patterns results in increased overheads in the I/O component of the visualization and analysis process.

5.4.3 Data Storage and Transfer

Scientists need to ensure that storage is optimal for state-of-the-art visualization algorithms. Storage will also need to map well to the systems on which data is to be processed. Achieving this objective will require a concerted effort by scientists, visualization experts, and storage researchers.

Reducing data within storage systems is one way to lessen analytical I/O requirements. Scientists are researching active storage technologies. These technologies could be an important enabler if they allow analysis primitives to execute within the storage system.

Where scientists prefer to view the results of remote simulations locally, minimizing the amount of data transferred is critical. Scientists need to do more research to understand how to integrate data reduction into remote I/O protocols in an efficient manner. This approach will let researchers reduce data requirements prior to the movement of datasets over long-haul networks.

5.5 Workflows and Provenance

Developing insightful visualizations often requires a combination of loosely coupled computational and data resources, specialized libraries, and Grid and Web services. Designing such a process involves data management and statistical analysis tasks that need to extract data from very large datasets, to transform or transpose data, summarize statistics, discover patterns, and reason analytically.

Many visualization and data analysis libraries and tools are available for these tasks today. They include VTK, VisIt, ParaView, and SCIRun. All of these can process very large data volumes in parallel. Some tools, like VisTrails, have advanced provenance, comparative, and multiview capabilities. Scientists routinely use statistical and plotting tools, including R, matplotlib, and IDL. Integrated environments, such as Matlab and Mathematica, are also popular. A number of data management tools, such as NetCDF and HDF5, support specialized data formats. Others, including FastBit, support specialized indexing methods that can perform value-based queries and subset extraction efficiently. Because these tools are not integrated, however, scientists are restricted in terms of the amount of visualization and data analysis they can perform. This is a major shortcoming.

Instead of developing a single, monolithic system with a wide range of capabilities, scientists need to integrate technologies and tools from different domains into a single framework. Such a framework not only would enable the interaction of various tools, but it would allow scientists to integrate existing and future software modules so that tasks might be completed in an end-to-end, continuous manner. Work is needed to ensure interoperability between visualization, data management, statistical, and reasoning tools. Additional efforts should begin to develop specialized workflow capabilities for visualization and data analysis. Developing these tools is

especially challenging because they would need to deal with expected petascale and exascale datasets and multiple scientific domains.

Most problems in biology involve analyzing large quantities of data and/or many loosely coupled, "multitask" computations. These problems require more than a good message-passing interface (such as MPI) or an MPI/OpenMPI implementation and support for parallel I/O. They require extreme-scale, data-intensive computing and scripting or workflow paradigms. Efforts to improve workflows include the following:

- **Human-in-the-loop**[28]: This represents the interactive steering of complex, multicomponent models. It also includes real-time, computational exploration, using exascale computers to solve petascale problems in real time and simple, intuitive interfaces that can aggregate information, from as many as 1,018 cores, for decision making.
- **Annotation:** Tagging relevant information to data is crucial at all levels as research moves to exascale computing with more automated workflows and tools. The annotation of components, data, and so forth needs to be able to facilitate provenance identification, high-level scripting of workflows, validation, and data archiving.
- **Exascale Models:** Scientists need to abstract scientific codes from new exascale technologies. These codes need to address issues such as fault tolerance—check pointing will not be the paradigm, parallelism (i.e., message-driven paradigms), and architecture heterogeneity (i.e., accelerators and memory hierarchies). Creating abstraction layers in high-level domain languages, component frameworks, and scientific libraries can help develop the needed codes. It will be critical for scientists to address the challenges created by complexity in the data structures and by the multiscale and multiphysics nature of models.
- **Power:** Exascale computing will be constrained by its power requirements. Decision making about the scope and scheduling of workflows will need to consider power consumption demands. Power-based criteria will be part of optimizing data analysis workflows.
- **Data:** File-based I/O is not expected to scale from petascale to exascale. If visualization and analysis are done entirely *in situ*, models will need to contain more intelligence to perform the appropriate validation and analysis at run time.

5.6 Summary of Technical Recommendations

Data, image, and visual analysis are essential technologies for the extreme-scale, biological-science, and discovery process. Work in the following technical areas should have the highest priority:

- *Pervasive Parallelism and Multiscale Analysis*. New developments in computer architecture will help scientists develop parallelizable, visualization applications that operate at multiple levels. Scientists require this capability if they are to maximize the time they have to analyze data.
- *Feature Detection and Tracking*. New algorithms will help researchers detect and track of features of scientific interest. This feature will help researchers discover important new areas that they can investigate further.

- *Multifield and Multimodel Data Understanding*. New approaches will enable scientists to compare and develop a combined analysis of multivariable data. This data is increasingly common in extreme-scale datasets.
- *In Situ Processing*. To maximize the effectiveness of large computational resources, scientists need to integrate visualization algorithms with simulation models.
- *Time-Varying Datasets*. New visualization techniques and user interfaces will help users understand extreme-scale, time-varying, multivariate datasets. Scientists need to browse through different spatial and temporal scales interactively. They need to identify scientific phenomena of different temporal lengths and to track salient features in both time and space.
- *Visual Analysis, Quantification, and Representation of Uncertainty and Error*. Scientists need new approaches to quantify uncertainty and error in the analysis process. This would provide immediate feedback once they select specific forms of visual analysis.
- *Workflows and Provenance*. Scientists need to be able to document and track the entire scientific process in a straightforward manner. End-to-end integration strategies must be developed that considers the entire simulation and analysis process as an analog to a biological experiment.

5.7 Summary of Research Areas

The research areas with the most activity and promise, illustrating the immediate need for extreme-scale capabilities include:

- *Fundamental Algorithms*. Scientists need effective data analysis and visualization tools to support predictive simulations and scientific discovery. These tools should be based on strong algorithmic and mathematical foundations. They ought to allow scientists to characterize salient features in their data reliably. New, mathematical methods will help improve the scientific understanding of images.
- *Topological Methods*. The introduction of robust combinatorial techniques for topological analysis has accelerated the use of topology to present known phenomena and for the detection and quantification of new features that are of interest to science.
- *Statistical Analysis*. Today's data analysis capabilities lag far behind researchers' ability to produce data from simulations or observational records, particularly in the mathematics that scientists need to convey analysis and estimation methodology into a data-parallel environment. The new mathematics must consider an entire estimation or analysis problem in a specific application and, thereby, assist in developing scalable, data-parallel algorithms for data analysis.
- *Feature Detection and Tracking*. Scaling simulations to ever-finer granularity and time steps presents new challenges in visualizing the data produced, so scientists need to create smart, semi-automated, visualization algorithms. These would permit scientists to analyze the immense data by following a more, "top-down," methodological path.
- *Uncertainty Analysis and Visualization*. Handling uncertainty must be an end-to-end process, where the different sources of uncertainty are identified, quantified, represented, tracked, and visualized along with the underlying data. Hence, scientists need to develop uncertainty representation and quantification, uncertainty propagation, and uncertainty visualization techniques in order to provide credible and verifiable visualizations.

- *Image Analysis*. Biological imaging provides information at various scales ranging from the organ level to the molecular level, and scientists depend on image analysis to explore many biological problems, such as the study of phenotypes and populations. Image analysis can similarly inform modeling: in the case of reverse engineering the brain, it provides morphology and connectivity information.
- *The Complexity of Biology Datasets*. Simulation codes are producing data at exponentially increasing sizes; spatial resolution is only one axis along which datasets are expanding. As computational scales reach the extreme scale, the codes are also increasing their temporal resolution, degree of code coupling, and extent of parametric exploration.
- *The Extreme Scale of Imaging Datasets and Models for the Next Phase of Simulation Science*. Presently fielded HPC platforms serve science communities with great large and scalable computational challenges but these same infrastructures struggle when the sizes of data volumes under study are in the teravoxel or petavoxel range. This is now the case as biology has in the last decade surpassed astronomy in the size and complexity of continuous imaging data across all scales. It well accepted that insight into many complex biological functions could be obtained by enabling computational work on multiscale representations of complex biological systems. Exascale architectures that accommodate this large data trend will be crucial to biological computing at the extreme.
- *Multimodal Data Understanding.* Visualization and analysis experts have a critical need for tools that leverage semantic information and hide details of dataset formats; a way for experts to focus on the design of these approaches and not become mired in the minutia of specific data representations.
- *Time-Varying Datasets*. The amount of data that simulations produce has grown exponentially in recent years. One factor contributing to this growth is the ability to run very large scale, time-varying simulations. Most current analytical methods do not function well with targeted, time-varying data, in spite of efforts to improve the visualization of very large datasets. New visualization techniques and user interfaces are needed to help users understand exascale, time-varying, multivariate datasets.
- *Advanced Architectures and Software*. Emerging petascale and exascale architectures are both a blessing and a curse, offering unprecedented computational power that can produce results much faster than machines can store them – often faster than researchers can visualize them.
- *Pervasive Parallelism*. As the number of transistors doubles, the number of cores will double. This means that software of the future will be very different from the sequential programs scientists use today. This revolution in computer architecture will have an enormous impact on graphics and visualization. In fact, the visualization pipeline we know today will be radically different if it is to exploit the new architectures.
- *In Situ Processing*. Faster computation rates make it possible to run simulations at higher fidelity, which yields even more data. Unfortunately, a cost-effective storage system that tracks the rate at which it consumes additional data is not available, yielding a divide between what computers produce and what storage is capable of maintaining.
- *Data Access*. Despite the challenges, *in situ* processing can help scientists reduce the disparity between data generation rates and storage system capabilities, an important component in managing petascale and exascale datasets. Applications run on future systems will need to store an unprecedented amount of simulation data during their run

time. Current practice relies on the post processing datasets from leadership-computing applications.

- *Workflows and Provenance*. Researchers need to analyze and understand scientific data, assembling complex, computational processes and developing insightful visualizations. This approach often requires a combination of loosely coupled computational and data resources, specialized libraries, and Cloud and Web services. Designing such a process involves data management and statistical analysis tasks that need to extract data from very large datasets, to transform or transpose data, summarize statistics, discover patterns, and reason analytically.

REFERENCES

"Animal Cell Structure," http://micro.magnet.fsu.edu/cells/animalcell.html

Blagoev, K., et al., "Modeling the Magnetic Signature of Neuronal Tissue," *NeuroImage*, vol. 37, pp. 137-148, 2007.

Blue Brain URL: http://bluebrain.epfl.ch/ and H. Markram, "The Blue Brain Project," *Nature Reviews Neuroscience*, vol. 7, pp. 153–160, 2006.

Bonneau, R., N. S. Baliga, E. W. Deutsch, P. Shannon, and L. Hood, "Comprehensive de novo Structure Prediction in a Systems-Biology Context for the archaea Halobacterium sp. NRC-1," *Genome Biology*, vol. 5, no. 8, 2004.

"The Born approximation," Wikipedia, http://en.wikipedia.org/wiki/Born_approximation

Chiponchip.org URL: http://www.chiponchip.org/

Collins, F.S., M. Morgan and A. Patrinos, "The Human Genome Project: Lessons from Large-Scale Biology," *Science*, vol. 300, pp. 286-290, April 11, 2003.

Cooper, T. F., D. E. Rozen, and R. E. Lenski, "Parallel Changes in Gene Expression after 20,000 Generations of Evolution in E. coli," in the *Proceedings of the National Academy of Sciences of the United States of America*, volume 100, pp. 1072–1077, 2003.

Duan, Y., L. Wang, and P. A. Kollman, "The Early Stage of Folding of Villin Headpiece Subdomain Observed in a 200-Nanosecond Fully Solvated Molecular Dynamics Simulation," in the *Proceedings of the National Academy of Sciences of the United States of America*, volume 95, no. 17, pp. 9897–9902, 1998.

Feist, A. M., M. J. Herrgard, I. Thiele, J. L. Reed, and B. O. Palsson, "Reconstruction of Biochemical Networks in Microorganisms," *Nature Reviews*, vol. 7, pp. 129–143. 2009.

Fraser, C., W. P. Hanage, and B. G. Spratt, "Recombination and the Nature of Bacterial Speciation," *Science*, vol. 315, no. 5811, pp. 476–480. January 26, 2007.

Fraser, C., W. P. Hanage, and B. G. Spratt, "Neutral Microepidemic Evolution of Bacterial Pathogens," in the *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 6, pp. 1968–1973, 2005.

George, J.S. et al., "Mapping Function in the Human Brain with MEG, anatomical MRI, and functional MRI," *Journal of Clinical Neurophysiology*, vol. 12, pp. 406-431, 1995.

Heller, L. et al., "Modeling Direct Effects of Neural Current on MRI," *Human Brain Mapping*, vol. 30, pp. 1-12, 2009.

Henry, C., F. Xia, and R. Stevens, "Application of High-Performance Computing to the Reconstruction, Analysis, and Optimization of Genome-scale Metabolic Models," *Journal of Physics***,** Conference Series, vol. 180, 2009.

Hucka, M., et al., “Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project,” *Systems Biology*, vol. 1, no. 1, pp. 41, June 2004.

Johnson, R., R. Ross, S. Ahern, J. Ahrens, W. Bethel, K. L. Ma, M. Papka, J. van Rosendale, H. W. Shen, and J. Thomas, “Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale,” Department of Energy Office of Advanced Scientific Computing Research, 2007.
http://www.sci.utah.edu/vaw2007/DOE-Visualization-Report-2007.pdf

Kondrashov, A. S. and F. A. Kondrashov, “Interactions among Quantitative Traits in the Course of Sympatric Speciation,” *Nature*, vol. 400, pp. 351–354, 1999.

Mead, C.,“Neuromorphic Electronic Systems,” in the *Proceedings of the IEEE*, vol. 78, pp. 1629–1636. 1990.

MIT News, August 13, 2003. http://web.mit.edu/newsoffice/2003/plankton.html

OMICS: A Journal of Integrative Biology http://www.liebertpub.com/products/product.aspx?pid=43

Panmictic populations are unstructured, random-mating populations. See: Panmictic definition.
http://www.biochem.northwestern.edu/holmgren/Glossary/Definitions/Def-P/panmictic.html

Prokaryotes, Eukaryotes, & Viruses Tutorial, The Biology Project, University of Arizona, Department of Biochemistry and Molecular Biophysics
http://www.biology.arizona.edu/Cell_BIO/tutorials/pev/page2.html

Serre, T., A. Oliva, and T. Poggio, “A Feedforward Architecture Accounts for Rapid Categorization,” in the *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 6424–6429, 2007.

Simon, H., “Computers, Communications and the Public Interest,” Martin Greenberger, ed., Baltimore: The Johns Hopkins Press, 1971, pp. 40–41.

“Synthetic Cognition,” Los Alamos National Laboratory, http://synthetic-cognition.lanl.gov/index.php.

Szyperski, T., “Biosynthetically Directed Fractional 13C-Labeling of Proteinogenic Acids: An Efficient Analytical Tool to Investigate Intermediary Metabolism,” *European Journal of Biochemistry*, vol. 232, no. 2, pp. 433–448. September 1, 1995.
http://www.ncbi.nlm.nih.gov/pubmed/7556192.

Tabata, O., “Introduction of MEMS Activity at Nano/Micro System Engineering Lab,”
http://nanohub.org/resources/3243/.

Thiele, I., R. M. T. Fleming, A. Bordbar, R. Que, and B. O. Palsson, “An Integrated Model of Macromolecular Synthesis and Metabolism of Escherichia coli,” in preparation.

Thiele, I., N. Jamshidi, R. M. Fleming, and B. O. Palsson, “Genome-Scale Transcriptional and Translational Machinery: A Knowledge Base, Its Mathematical Formulation, and Its Functional Characterization,” *PLoS Computational Biology*, vol. 5, no. 3, e1000313, 2009.

"Translation and Open Reading Frame Search," University of Wisconsin System, Board of Regents, http://bioweb.uwlax.edu/genweb/molecular/seq_anal/translation/translation.html

voxel, Internet.com Webopedia, http://www.webopedia.com/TERM/V/voxel.html

Wolfram MathWorld, "NP-Hard Problem," http://mathworld.wolfram.com/NPHardProblem.html

Zaghloul, K. A., and K. Boahen, "A Silicon Retina That Reproduces Signals in the Optic Nerve," *Journal of Neural Engineering*, vol. 3, pp. 257–267, 2006.

APPENDIX A: Workshop Charge

Department of Energy
Washington, DC 20585

MAY 13 2009

Professor Rick Stevens
Computing and Life Sciences
Argonne National Laboratory
9700 South Cass Avenue
Building 221
Argonne, Illinois 60439

Dear Professor Stevens:

I write to ask if you will co-organize and conduct an international workshop to examine the scientific opportunities in advanced modeling and simulation at the exascale in the biological sciences. You will be joined by co-organizer Professor Mark Ellisman of the University of California San Diego.

A key goal for the workshop is to present the biological community with the opportunity to shape the appropriate role for scientific computing at the exascale in the quest to advance the scientific frontiers in fundamental biological and ecology research and to examine the role of extreme scale computing in applied biological research such as bioenergy, bioremediation and the understanding of the global carbon cycle.

This workshop will build on the series of the ASCR sponsored town hall meetings held in the spring of 2007, examining the potential of exascale computing for applications in energy and the environment, the BER workshops addressing scientific opportunities in research in the global carbon cycle, GTL knowledgebase, and bioenergy held in 2008 and 2007, the workshop sponsored by NSF in 2006 looking at the opportunities in biology at the petascale and the recent national academy studies on The *Impact Of High-End Computing On Illustrative Fields* (evolution), *The Role Of Theory In Advancing Biology* and *The Frontiers At The Interface Between Computing And Biology*.

The report from your workshop is expected to be a document that should not exceed 100 pages and should be completed by the end of August 2009 if at all possible. The workshop will be a collaborative effort between BER and ASCR and we encourage you to involve the appropriate members of the community to create a balanced yet forward looking discussion and report. We would also welcome any other recommendations on program content, emphasis, or balance. This effort, we realize, is a large undertaking. However, we believe that biological research is poised to be dramatically influenced and accelerated by the likely advances in computing over the next decade and the benefits of fully exploiting future computing systems to advance our understanding will accrue not only to science but to society at large and thus justifies the effort.

A desired outcome of these meetings is to develop a short list of "global challenge" computational problems. Solving these problems should have the potential to transform

our understanding of science and its impacts and to improve our ability to apply knowledge in applications important to science, engineering, industry, and society. We anticipate that a final workshop report will address these global challenge computational problems.

This list of topics to consider include: *atomistic level biomolecular modeling, protein complexes, cell and organelles level modeling, protein folding, protein and pathway engineering, computational genomics and genome scale high-throughput data analysis, computational evolution, community and population level simulation, ecosystems modeling, artificial life and evolutionary computation, computational neuroscience and organ and tissue level simulation for complex organisms and computational approaches to imaging for biological systems.*

An effort should be made to identify the general scope of the funding required to achieve success.

As co-chair of this endeavor, you will play a critical role in ensuring the success of the workshop. It is a major responsibility: with your help, and that of your colleagues, you will enable ASCR and BER to identify key problems of national interest, document both the science and the computational case rigorously, and contribute to a focused DOE program for the next fifty years.

Dr. Susan Gregurick and Dr. Daniel Drell of the Office of Biological and Environmental Research and Mrs. Christine Chalk of the Office of Advanced Scientific Computing Research are the program managers responsible for this workshop. Susan, Dan and Christine will be contacting you shortly to discuss the schedule, deliverables, logistics and administrative needs.

If, at any time, you have questions about current plans, priorities and strategies, please fell free to contact us. Many thanks for your willingness to lead what we hope will be a landmark workshop in the field.

Sincerely,

Anna Palmisano
Associate Director of Science
for Biological and Environmental Research

Michael Strayer
Associate Director of Science
for Advanced Scientific Computing Research

APPENDIX B: Workshop Agenda

Opportunities in Biology at the Extreme Scale of Computing

AGENDA
August 17-19, 2009
Hotel: Sheraton Chicago Hotel and Towers
Sunday, August 16, 2009

Time	Session	Lead
6:00 pm	Pre-Workshop Meeting for Organizers, B/O Co-Leads, & Plenary Session Speakers	Mark Ellisman/Rick Stevens

Monday, August 17, 2009

Time	Session	Lead
7:30-8:00 am	Registration/Morning refreshments	
8:00-8:30 am	Opening remarks from Organizers (Mark Ellisman/Rick Stevens), BER (David Thomassen), ASCR (Michael Strayer)	
Plenary Talks		
8:30-9:00 am		David Galas
9:00-9:30 am	Opportunities in Large-Scale Computation for Biology	Anne Trefethen
9:30-10:00 am	Break	
10:00-10:30 am	Image-Based Biological Modeling, Simulation, and Visualization	Chris Johnson
10:30-11:00 am	Future Directions in Computing Environments	Rick Stevens
11:00-11:30 am	Workshop Kick-off and Instructions	Mark Ellisman/Rick Stevens
Time	**Session**	**Lead**
11:30 am-3:30 pm	***Breakout Sessions (All sessions run in parallel)***	
	Tissues, Organelles, and Physiology Modeling	John George/George Karniakadis
	Pathways, Cells, and Organelles	Adam Arkin/David Galas
	Macromolecular Proteins and Protein Complexes	Mike Colvin/Tamar Schlick
	From Genomics and Populations to Ecosystems and Evolutionary Dynamics	Ed Delong/Gary Olsen
	Imaging and Computing in the Loop	Chris Johnson/Nagiza Sematova
11:30 am	Working Lunch	
3:30 pm	Break	
4:00-5:00 pm	***Continue Breakout Sessions (return to breakout rooms)***	
5:00-5:30 pm	Organizers and B/O Leads meetings	

Tuesday, August 18, 2009

Time	Session	Lead
8:00 am -5:00 pm	***Breakout Sessions (All sessions run in parallel)***	Adam Arkin/David Galas

	Tissues, Organelles, and Physiology Modeling	John George/George Karniakadis
	Pathways, Cells, and Organelles	Adam Arkin/David Galas
	Macromolecular Proteins and Protein Complexes	Mike Colvin/Tamar Schlick
	From Genomics and Populations to Ecosystems and Evolutionary Dynamics	Ed Delong/Gary Olsen
	Imaging and Computing in the Loop	Chris Johnson/Nagiza Sematova
10:00-10:30 am	Break	
12:00 pm	Working Lunch	
3:00-4:00 pm	Break	
5:00 pm	Dinner on your own	
5:00-5:30 pm	Meet with co-chairs for summary of Days 1 and 2	

Wednesday, August 19, 2009

Time	Session
8:00-8:30 am	Report Out: Tissues, Organelles, and Physiology Modeling
8:30-9:00 am	Report Out: Pathways, Cells and Organelles
9:00-9:30 am	Report Out: Macromolecular Proteins and Protein Complexes
9:30-10:00 am	Break
10:00-10:30 am	Report Out: From Genomics and Populations to Ecosystems and Evolutionary Dynamics
10:30-11:00 am	Report Out: Imaging and Computing in the Loop
11:00-12:00 pm	Closing remarks and adjournment

APPENDIX C: Workshop Attendees

Opportunities in Biology at the Extreme Scale of Computing

August 17-19, 2009

Sheraton Chicago Hotel and Towers

List of Attendees

94 attendees

Allen, Gabrielle
Louisiana State University
gallen@lsu.edu

Arkin, Adam
Lawrence Berkeley National Laboratory
aparkin@lbl.gov

Baker, Nathan
Washington University in St. Louis
baker@biochem.wustl.edu

Baliga, Nitin
Institute for Systems Biology
nbaliga@systemsbiology.org

Banda, Michael
Lawrence Berkeley Lab
mjbanda@lbl.gov

Barcellos-Hoff, Mary Helen
New York University School of Medicine
mhbarcellos-hoff@nyumc.org

Beck, David
University of Washington
dacb@u.washington.edu

Beckman, Pete
Argonne National Laboratory
beckman@mcs.anl.gov

Cannon, William
Pacific Northwest National Laboratory
william.cannon@pnl.gov

Carter, Jonathan
Lawrence Berkeley National Laboratory
jtcarter@lbl.gov

Chalk, Christine
U.S. Department of Energy
christine.chalk@science.doe.gov

Chang, Christopher
National Renewable Energy Laboratory
christopher.chang@nrel.gov

Chatterjee, Lali
U.S. Department of Energy
lali.chatterjee@ascr.doe.gov

Cohen, Robert
Cohen Communications Group
bcohen@bway.net

Colella, Phillip
Lawrence Berkeley National Laboratory
PColella@lbl.gov

Colvin, Michael
University of California, Merced
mcolvin@ucmerced.edu

Cottingham, Robert
Oak Ridge National Laboratory
cottinghamrw@ornl.gov

Couch, Jennifer
NCI, NIH
couchj@mail.nih.gov

D'haeseleer, Patrik
LLNL
patrikd@llnl.gov

DeJongh, Matt
Hope College
dejongh@hope.edu

DeLong, Edward
MIT
delong@mit.edu

Demir, Semahat
NSF
sdemir@nsf.gov

Drell, Daniel
Office of Biological and Environmental Research
daniel.drell@science.doe.gov

Duan, Yong
UC Davis
duan@ucdavis.edu

Edwards, Rob
Argonne National Laboratory
redwards@mcs.anl.gov

Ellisman, Mark
UCSD
mark@ncmir.ucsd.edu

Emonet, Thierry
Yale University
thierry.emonet@yale.edu

Faeder, James
University of Pittsburgh
faeder@pitt.edu

Feng, Wu
Virginia Tech
feng@cs.vt.edu

Fink, Wolfgang
Caltech
wolfgang.fink@jpl.nasa.gov

Foster, Ian
Argonne/University of Chicago
foster@anl.gov

Galas, David
Institute for Systems Biology
dgalas@systemsbiology.org

Garrity, George
Michigan State University
garrity@msu.edu

George, John
Los Alamos National Laboratory
jsg@lanl.gov

Germain, Robert
IBM T.J. Watson Research Center
rgermain@us.ibm.com

Gibson, Garth
Carnegie Mellon University/Computer Science Dept.
GARTH@CS.CMU.EDU

Giles, Roscoe
Boston University
roscoe@bu.edu

Gilna, Paul
Calit2/UCSD
pgilna@ucsd.edu

Godzik, Adam
Burnham Institute for Medical Research
adam@burnham.org

Grama, Ananth
Purdue University
ayg@cs.purdue.edu

Gregurick, Susan
DOE/Office of Biological and Environmental Science
susan.gregurick@science.doe.gov

Grethe, Jeffrey
University of California, San Diego
Jeffrey.S.Grethe@alumni.usc.edu

Henry, Christopher
Argonne National Laboratory
chenry@mcs.anl.gov

Hey, Tony
Microsoft
carolepo@microsoft.com

Hill, Sean
Blue Brain Project
sean.hill@epfl.ch

Hines, Michael
Yale University
michael.hines@yale.edu

Johnson, Chris
SCI Institute
crj@sci.utah.edu

Jouline (Zhulin), Igor
University of Tennessee - Oak Ridge National Laboratory
ijouline@utk.edu

Karniadakis, George
Brown University
gk@dam.brown.edu

Kostova- Vassilevska, Tanya
National Science Foundation
tvassile@nsf.gov

Kusnezov, Dimitri
National Nuclear Security Administration
Dimitri.Kusnezov@nnsa.doe.gov

Lang, Sam
Argonne National Lab
slang@mcs.anl.gov

Lawrence, Albert
University of California, San Diego
albert.rick.lawrence@gmail.com

Le Gros, Mark
Lawrence Berkeley National Laboratory
malegros@lbl.gov

Livny, Miron
University of Wisconsin
miron@cs.wisc.edu

Loriaux, Paul
Univ of CA San Diego
ploriaux@ucsd.edu

Lusk, Ewing (Rusty)
Argonne National Laboratory
lusk@mcs.anl.gov

Luthey-Schulten, Zaida
University of Illinois
schulten@scs.uiuc.edu

Maccabe, Arthur
Oak Ridge National Laboratory
maccabeab@ornl.gov

Maranas, Costas
Penn State
costas@psu.edu

May, Elebeoba
Sandia National Laboratories
eemay@sandia.gov

McCue, Lee Ann
Pacific Northwest National Lab
leeann.mccue@pnl.gov

Messina, Paul
Argonne National Laboratory
messina@mcs.anl.gov

Meyer, Folker
Argonne National Laboratory
folker@anl.gov

Mitchell, Julie
University of Wisconsin - Madison
jcmitchell@wisc.edu

Morss, Helaine Sue
ANL/DOE-ASCR
smorss@anl.gov

Nealson, Kenneth
Univ. Southern California
knealson@usc.edu

Nichols, Jeff
Oak Ridge National Laboratory
nicholsja@ornl.gov

Nowell, Lucy
U.S. Department of Energy
lucy.nowell@science.doe.gov

Oehmen, Chris
Pacific Northwest National Laboratory
christopher.oehmen@pnl.gov

Olsen, Gary
University of Illinois at Urbana-Champaign
gary@life.illinois.edu

Papadopoulos, Philip
UCSD
philip.papadopoulos@gmail.com

Papka, Michael
Argonne National Laboratory
papka@anl.gov

Polz, Martin
Massachusetts Institute of Technology
mpolz@mit.edu

Reed, Jennifer
UW-Madison
reed@engr.wisc.edu

Riley, Katherine
Argonne National Laboratory
riley@alcf.anl.gov

Roux, Benoit
University of Chicago
roux@uchicago.edu

Russell, Thomas
National Science Foundation
trussell@nsf.gov

Samatova, Nagiza
North Carolina State University / ORNL
samatovan@ornl.gov

Sanbonmatsu, Kevin
Los Alamos National Laboratory
kys@lanl.gov

Schlick, Tamar
NYU
schlick@nyu.edu

Smith, Jeremy
Oak Ridge National Laboratory
smithjc@ornl.gov

Stevens, Rick
Argonne National Laboratory/U. of Chicago
stevens@anl.gov

Stiles, Joel
Pittsburgh Supercomputing Center
stiles@psc.edu

Tasdizen, Tolga
University of Utah
tolga@sci.utah.edu

Thakur, Rajeev
Argonne National Laboratory
thakur@mcs.anl.gov

Thelen, Michael
Lawrence Livermore National Laboratory
mthelen@llnl.gov

Thiele, Ines
Center for Systems Biology, U. of Iceland
ithiele@ucsd.edu

Thomassen, David
DOE/Biological and Environmental Research
david.thomassen@science.doe.gov

Tyson, John
Virginia Tech
tyson@vt.edu

Tzonev, Svilen
DOE Joint Genome Institute
stzonev@lbl.gov

White, Andy
Los Alamos National Laboratory
abw@lanl.gov

Woloschak, Gayle
Northwestern University
g-woloschak@northwestern.edu

Wooley, John C.
University of California, San Diego
jwooley@ucsd.edu