Biology at the Exascale

Advances in computational hardware and algorithms that have transformed areas of physics and engineering have recently brought similar benefits to biology and biomedical research.

Contributors: Laura Wolf and Dr. Gail W. Pieper, Argonne National Laboratory

Biological sciences are undergoing a revolution. High-performance computing has accelerated the transition from hypothesis-driven to design-driven research at all scales, and computational simulation of biological systems is now driving the direction of biological experimentation and the generation of insights.

As recently as ten years ago, success in predicting how proteins assume their intricate three-dimensional forms was considered highly unlikely if there was no related protein of known structure. For those proteins whose sequence resembles a protein of known structure, the three-dimensional structure of the known protein can be used as a "template" to deduce the unknown protein structure. At the time, about 60 percent of protein sequences arising from the genome sequencing projects had no homologs of known structure.

In 2001, Rosetta, a computational technique developed by Dr. David Baker and colleagues at the Howard Hughes Medical Institute, successfully predicted the three-dimensional structure of a folded protein from its linear sequence of amino acids. (Baker now develops tools to enable researchers to test new protein scaffolds, examine additional structural hypothesis regarding determinants of binding, and ultimately design proteins that tightly bind endogenous cellular proteins.)

Two years later, a thirteen-year project to sequence the human genome was declared a success, making available to scientists worldwide the billions of letters of DNA to conduct postgenomic research, including annotating the human genome. (Today, with technology driving the sequencing and annotation of thousands of single-celled organisms, the majority of diverse microbial organisms will be completely sequenced by the time exascale computing arrives.)

In results also published in 2003, a research team led by the University of Chicago described the first molecular dynamics computation that accurately predicted a folded protein—a small, fast folding 36-residue alpha helical protein called the villin headpiece. (The villin headpiece, which folds on the microsecond timescale, remains the subject of computational studies today.)

In 2005, when Los Alamos National Laboratory researchers Dr. Kevin Sanbonmatsu and Dr. Chang-Shung Tung conducted the first million-atom simulation in biology, they noted that the first million-particle simulations in materials science and cosmology had been performed over a decade prior. (Today, computational physicists in both fields perform multibillion-particle simulations.) The all-atom biomolecular simulation of the ribosome is considered a high-performance computing milestone in the field of computational molecular biology—it is the most demanding in terms of compute, communication speed, and memory.

Exascale Computing Challenges

Fifteen years ago, petaflop computation was not possible. Today, simulation of biological processes is already pushing beyond the petascale class of computing systems coming online. Such simulations are now capable of delivering sustained performance approaching 10^{15} floating-point operations per second (petaflops) on large, memory-intensive applications.

In a series of three town hall meetings held in 2007 to engage the computational science community about the potential benefits of advanced computing, several global challenge problems were identified in the areas of energy, the environment, and basic science—all with significant opportunities to exploit computing at the exascale (10^{18}) and all with a common thread: biology.

In energy, scientists look to exascale to be able to attack problems in combustion, the solution of which could improve the efficient use of liquid fuels, whether from fossil sources or from renewable sources. First-principles computational design and optimization of catalysts will also become possible at the exascale, as will de novo design of biologically mediated pathways for energy conversion.

In the environment, scientists anticipate the need for exascale computing in climate modeling; integrated energy, economics, and environmental modeling; and multiscale modeling from molecules to ecosystems. Many biological processes of interest to the Office of Science are mediated by membrane-associated proteins, including the detoxification of organic waste products.

In biology, the challenges of modeling at multiple scales—from atomic, through genomic and cellular, to ecosystems—is driving the need for exascale computing and a new set of algorithms and approaches. For example, a computational approach to understanding cellular machines and their related genes, and biochemical pathways, referred to as “systems biology,” aims to develop validated capabilities for simulating cells as spatially extended mechanical and chemical systems in a way that accurately represents processes such as cell growth, metabolism, locomotion, and sensing. Modeling and simulation provide only a local view of each process, without interaction between modalities and scales. Exascale algorithms and software tools are needed to represent the various macroscopic subsystems and to enable a multiscale approach to biological modeling.

New Tools Needed for New Challenges

In 2005, Dr. Nobuyasu Koga and Dr. Shoji Takada of Kobe University carried out a folding-based molecular simulation but were limited by the capabilities of existing high-performance computing systems. Two years later, Dr. Benoit Roux, a computational scientist at Argonne National Laboratory and the University of Chicago, found that the time and energy scales of computing these underlying molecular processes (figure 1, p. 00) are just within reach of the Cray XT and IBM Blue Gene leadership-class computers (see sidebar “Using Leadership Facilities to Advance Biology,” p. 00).

Even with these leadership-class systems, however, the infrastructure needed to generate and analyze molecular data will require development of simulation management tools that encompass clustering, archiving, comparison, debugging, visualization, and communication—all of which must also address current computing bottlenecks that limit the scope of analysis. For example, researchers are limited by the current microsecond timescale for protein folding required by the huge number of intermolecular, interaction computations. Scientists also lack rigorous coarse-grained models that permit the scaling up of macromolecular pathways and supramolecular cellular processes. Similarly, systems biology methods lack the dynamic resolution needed for coupling genomic and other data in order to map cellular networks, to predict their functional states, and to control the time-varying responses of living cells. Nor can current analytic models adequately analyze the dynamics of complex living systems. Researchers have achieved impressive methodological advances that permit the modeling of the largest assemblies in the cell, but for short periods of time. And unfortunately, these simulations are unlikely to scale to the size of a single cell, even a small bacterium, for relevant times such as minutes or hours—even if researchers can employ computers capable of achieving 1,000 petaflops. New, scalable, high-performance computational tools are essential.

Seven Success Stories

In anticipation of exascale computing, and capitalizing on the capabilities of current leadership-class computers, researchers are conducting simulations previously believed infeasible.

Large-Scale Simulations of Cellulases. Dr. Jeremy Smith, a molecular biophysicist at Oak Ridge National Laboratory (ORNL), together with colleagues at the National Renewable Energy Laboratory in Colorado and Cornell University, are using ORNL's Jaguar supercomputer to model bacterial and fungal cellulases in action. Understanding how cellulases degrade cellulose is the key to increasing the efficiency and lowering the cost of ethanol production using sugar from cellulose in trees and other biomass. If the team can understand how the cellulase enzyme functions, how it recognizes cellulose strands, and how the chemistry is accomplished inside the enzyme, it may be able to determine what the rate-limiting steps are that might be genetically engineered to make cellulase more efficient at degrading cellulose into glucose. Large-scale molecular dynamic simulations generated by the supercomputer allow researchers to "watch" these simulated enzymes attack digital cellulose strands, transfer a strand's sugar molecules to the enzyme's catalytic zone, and chemically digest the sugar to provide the microbe with energy.

Building a Cognitive Computing Chip. Scientists from IBM Research and five university partners are leading an effort to understand the complex wiring system of the brain and to build a computer that can simulate and emulate the brain's abilities of sensation, perception, action, interaction and cognition while rivaling its low power consumption and compact size. Using the Dawn Blue Gene/P supercomputer at Lawrence Livermore National Laboratory with 147,456 processors and 144 terabytes of main memory, the team achieved a simulation with 1 billion spiking neurons and 10 trillion individual learning synapses. This is equivalent to 1,000 cognitive computing chips, each with 1 million neurons and 10 billion synapses, and exceeds the scale of cat cerebral cortex. The simulation ran 100 to 1,000 times slower than real-time. The team has also developed a new algorithm, BlueMatter, that exploits the Blue Gene supercomputing architecture to noninvasively measure and map the connections between all cortical and subcortical locations within the human brain using magnetic resonance diffusion-weighted imaging. Mapping the wiring diagram of the brain is crucial to untangling its vast communication network and understanding how it represents and processes information. Only recently has the technology increased sufficiently to match the density of neurons and synapses in real brains—around 10 billion to one square centimeter.

International Union of Physiological Sciences Physiome Project. The Physiome Project is a worldwide public domain effort to provide a computational framework for understanding human and other eukaryotic physiology. It aims to develop integrative models at all levels of biological organization, from genes to the whole organism, via gene regulatory networks, protein pathways, integrative cell function, and tissue and whole organ structure/function relations. Current projects include the development of ontologies to organize biological knowledge and access to databases; markup languages to encode models of biological structure and function in a standard format for sharing between different application programs and for reuse as components of more comprehensive models; databases of structure at the cell, tissue, and organ levels; software to render computational models of cell function. in 2 and 3D graphical form; and software for displaying and interacting with the organ models that will allow the user to move across all spatial scales.

Large-Scale Simulation of the Ribosome. Dr. Kevin Sanbonmatsu and Dr. Chang-Shung Tung of Los Alamos National Laboratory (LANL) have simulated the rate-limiting step in genetic decoding by the ribosome. The simulations used experimentally determined ribosome structures in different functional states as the initial and final conditions, making the simulations rigorously consistent with the experimental data. The calculations required approximately 1 million CPU-hours on 768 CPUs, or about 10% of the 13.88-teraflop LANL Q Machine. Previously, only static snapshot structures of the ribosome were available; limitations in time resolution and spatial resolution prevented experimental imaging of the ribosome in motion in atomic detail. The simulations on the Q Machine allowed the researchers to visualize the motion of transfer RNAs inside the ribosome occurring during decoding. The ribosome simulations helped to elucidate a crucial molecular mechanism for gene expression, which opens the door for simulations of other large molecular machines important for gene expression and drug design.

Large-Scale, Folding-Based Molecular Simulation. In 2005 Kobe University researchers Dr. Nobuyasu Koga and Dr. Shoji Takada published the results of their folding-based molecular simulations that revealed the mechanisms of the rotary motor F_1–ATPase. Biomolecular machines, such as the ribosome, transporter, and molecular motors, fulfill their function through large-amplitude conformational change. Molecular dynamics simulation is potentially powerful because it can provide full time-dependent structural information about biomolecular machines; but functional cycles of these systems typically take milliseconds or longer, which is far beyond the current reach of molecular simulations with all-atom standard force fields. Another approach is to use a coarse-grained molecular representation, thereby enabling the simulation orders of magnitude-longer time scales; however, coarse-graining drops some details from the model. Structural information before and after conformational change has been provided by x-ray crystallography and other methods for many cases; but these methods do not directly observe the molecular dynamics that connects two-end structures. These dynamical aspects can be observed directly by fluorescence and other time-resolved spectroscopy; however, the latter methods monitor local structure but do not give global structural information. The solution formulated by Drs. Koga and Takada was a computational framework, called a "switching Gō model," for simulating large-amplitude motion of biomolecular machines. For representing large-amplitude conformational dynamics, Gō models are suitable because they account for both small fluctuations around the native basin and large fluctuations that involve local unfolding. By combining all available experimental data with simulation results, the team identified the rotary motion of F_1–ATPase, which had long been under debate in the field. This work opens an avenue of simulating large-scale motion involved in dynamical function of large biomolecular complexes by folding-based model.

Large-Scale Simulation of Ion Channels. Voltage-gated ion channels, or Kv channels, are involved in the generation and spread of electrical signals in neurons, muscle, and other excitable cells. In order to open the gate of a channel, the electric field across the cellular membrane acts on specific charged amino acids that are strategically placed in the protein in a region called the voltage sensor. In humans, malfunction of these proteins, sometimes owing to the misbehavior of only a few atoms, can result in neurological diseases. A wealth of experimental data exists from a wide range of approaches, but its interpretation is complex. One must ultimately be able to visualize atom by atom how these tiny mechanical devices move and change their shape as a function of time while they perform. Dr. Roux and a team of researchers are using a tight integration of experiment, modeling, and simulation to gain insights into Kv channels (sidebar "Understanding of the Structure and Function of Ion Channels," p. 00). Their studies serve as a roadmap for simulating, visualizing, and elucidating the inner workings of these nanoscale molecular machines. Because these channels are functional electromechanical devices, they could be used in the design of artificial switches in various nanotechnologies. The practical applications of this work are significant. For example, the research in ion channel mechanisms may help identify strategies for treating cardiovascular disorders such as long-QT syndrome, which causes irregular heart rhythms and is associated with more than 3,000 sudden deaths each year in children and young adults in the United States. Moreover, the studies may help researchers find a way to switch or block the action of toxins—such as those emitted by scorpions and bees—that plug the ion channel pores in humans.

Protein Folding. In 2001, Howard Hughes Medical Institute investigator Dr. David A. Baker and his colleagues at the University of Washington, successfully predicted the three-dimensional structure of a folded protein from its linear sequence of amino acids. Key to the success was Rosetta, a computer algorithm for predicting protein folding. Experimental studies of protein folding by Baker's laboratory and many others had shown that each local segment of the chain flickers between a different subset of local conformations. Folding to the native structure occurs when the conformations adopted by the local segments and their relative orientations allow burial of the hydrophobic residues, pairing of the beta strands, and other low-energy features of native protein structures. In the Rosetta algorithm, the distribution of conformations observed for each short sequence segment in known protein structures is taken as an approximation of the set of local conformations that sequence segment would sample during folding. The program then searches for the combination of these local conformations that has the lowest overall energy. Dr. Baker has also set up a project, called Rosetta@home (http://boinc.bakerlab.org/rosetta/rah_about.phpto, to run the Rosetta program on unused computer resources. The intent is to accurately predicting and designing protein structures and protein complexes that may ultimately lead to finding cures for some major human diseases.

Exciting Applications Awaiting Exascale

Researchers have identified numerous exciting areas in biology for which exascale computing is required.

- **"Building the system" problems:** Rapid and high-fidelity assessment of metabolic and regulatory potential of thousands of cultured and sequenced prokaryotes of DOE-mission importance.
- **"Simulating the behavior" problems:** Predicting and simulating microbial behavior and response to changing environmental or process-related conditions—from simple to complex communities and ecosystems—spanning a range of spatial and temporal scales.
- **Reverse engineering of the brain:** Bottom-up models incorporating all available physiological detail in order to capture the biological function of the brain, predicting consequences of activation and of pharmacological or electrophysiological intervention.
- **Image-based phenotyping:** Segmentation of images that scale well to large amounts of data (e.g., a human) that could lead to personalized medicine.
- **Phylogenetics:** Phylogeny estimation, models of evolution, comparative biological methods, and population genetics, with particular focus on understanding horizontal gene transfer and the evolution of populations

- **Genome analysis and sequence analysis:** Genome assembly, genome and chromosome annotation, gene finding, alternative splicing, comparative genomics, multiple sequence alignment, sequence search and clustering, function prediction, motif discovery, and functional site recognition in protein, RNA and DNA sequences.

- **Structural bioinformatics:** Structure matching, structure prediction, analysis and comparison; methods and tools for docking; protein design and drug design
- **Systems biology:** Systems approaches to molecular biology, multiscale modeling, pathways, gene networks, large-scale development of models for many organisms and comparative modeling
- **Microbial ecology:** Engineering of stable microbial communities for practical applications, and understanding the carbon cycle through multiscale modeling of complex ecosystems

Further Reading
INCITE website: http://www.er.doe.gov/ascr/INCITE

SIDEBAR #2
Using Leadership Facilities to Advance Biology

The Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program established in 2003 and operated by the DOE's Office of Science, awards sizeable allocations (typically, millions of processor-hours per project) on America's premier leadership computing facility centers at Oak Ridge National Laboratory and Argonne National Laboratory. The program was conceived specifically to seek out computationally intensive, large-scale research projects with the potential to significantly advance key areas in science and engineering, and encourages proposals from universities, research institutions, and industry.

In 2009, the leadership computing facilities at Argonne and Oak Ridge national laboratories provided over 1 billion processor-hours to the INCITE program, reaching the expected milestone prior to the predicted date of 2010.

In the past two years, numerous INCITE awards have been allocated to researchers to advance the state of the art in the biological sciences.

2009 INCITE
Eight projects with strong biological components received INCITE grants in 2009.

1. Drs. Michael Heroux and Laura Frink of Sandia National Laboratories were awarded 1 million processor hours in 2009 to run unprecedented high-fidelity simulations of the interaction between antimicrobial peptides and the cell outer membrane; using recent modeling advances and advanced scalable solution algorithms. Understanding the detailed workings of the biophysics of these membranes may lead to breakthroughs in disease mechanisms and new treatments.

2. Dr. Ivaylo Ivanov of the University of California, San Diego, and Howard Hughes Medical Institute leads an interdisciplinary team from academia, two nonprofit research organizations, and two national laboratories, on a multiscale approach to modeling replisome assembly and function, specifically, the function of sliding clamps and clamp-loaders within the replisome; and the mechanisms of DNA repair enzymes. The replisome is a complex molecular machine that carries out replication of DNA. Allocated 2.6 million hours, this research has direct bearing on understanding the molecular basis of genetic integrity and the loss of this integrity in cancer and in degenerative diseases.

3. Human cells organize their functions through intricately shaped lipid membranes forming organelles (like the nucleus or endoplasmic reticulum), which facilitate cellular processes and characterize living cells. During the cell life cycle the curvature is sculpted by proteins that act on a nanometer scale, but through concerted action, produce shapes. Overall cellular shapes are induced by molecular events that naturally need to be concerted in a self-organized manner to produce cell-scale shapes. Dr. Klaus Shulten of the University of Illinois at Urbana-Champaign was allocated over 9.2 million computing hours to gain insight into the phenomenon.

4. Drs. Igor Tsigelny and Mark Miller of the University of California, San Diego research team were awarded 3 million processor-hours to develop a program package for modeling aggregation and membrane pore formation by unstructured proteins leading to neurodegenerative diseases and to devise new methods for stopping aggregation. The team also focused on computational engineering of microorganisms with molecular pores that will absorb specified ions and compounds. By simulating the various conformations sampled by protein in solution, the team investigated the process leading to the aggregation and creation of membrane-penetrating pores that allow selected ions to flow in and out of the cell—the key to stopping progression of diseases such as Parkinson and Alzheimer. The process of pore formation can also be modified for specific needs. For example, for environmental purposes the pores could be constructed in a way that they will permit flow to the cell of the selected ions only, for example, radionuclides.

5. University of Washington's Dr. David Baker received a renewal award of 12 million processor-hours to develop high-resolution structure prediction tools to build models of proteins with atomic-level accuracy and to computationally engineer both proteins and enzymes with new functions for applications ranging from basic research to therapeutics to bioremediation. Prediction of high-resolution protein structures from their amino acid sequences and the refinement of low-resolution models to high-resolution are long-standing problems in computational biology. The tools being developed will help experimentalists solve structures of biologically important proteins for which experimental x-ray phases are not available or arre hard to obtain.

6. Dr. Jeffrey Fox and his research team from Gene Network Sciences are simulating the complex, multiscale biological processes that control the heart's rhythm (figure ??). The goal is to gain insight into the underlying electrical mechanism for dangerous cardiac rhythm disorders and to determine the effects of interventions, such as drugs, that may prevent or exacerbate these potentially deadly arrhythmias. Recent advances in experimental technologies have allowed for more detailed characterizations of normal and abnormal cardiac electrical activity. The 21.4 million processor-hour renewal allows rapid testing of hypotheses for the initiation and maintenance of heart rhythm disturbances, specifically investigating ventricular fibrillation. The computer time will also be used to study how drug-induced modifications of molecular properties of heart cells can cause changes in tissue properties that might lead to another disorder called Torsades de Pointes.

7. Cell membrane-associated proteins play an essential role in controlling the bidirectional flow of material and information, and as such, they are truly "molecular machines" able to accomplish complex tasks. Large-scale gating motions, occurring on a relatively slow time-scale, are essential for the function of many important membrane proteins such as transporters and channels. Dr. Benoit Roux of Argonne National Laboratory and the University of Chicago received a renewal

award of 45 million processor-hours to further his studies of how these molecular protein-machines are able to carry out their function.

8. Efficient production of ethanol via hydrolysis of cellulose into sugars is a major energy policy goal. Plant cell wall lignocellulosic biomass is a complex material composed of crystalline cellulose microfibrils laminated with hemicellulose, pectin, and lignin polymers. ORNL's Jeremy Smith performs highly parallelized multi-length-scale computer simulations to help understand the physical causes of resistance of plant cell walls to hydrolysis—the major technological challenge in the development of viable cellulosic bioethanol. The solution to this challenge may be the improvement of pretreatments or the design of improved feedstock plants (or both).

Smith's simulations are part of a larger effort to integrate the power and capabilities of the neutron scattering and high-performance computing at ORNL to derive information on lignocellulosic degradation at an unprecedented level of detail.

2008 INCITE
Two projects in chemical science and physical chemistry had strong biologic components.

1. Dr. Christopher Mundy and colleagues at the Pacific Northwest National Laboratory and IBM Research-Zurich were awarded 1.5 million processor hours to apply the efficient sampling methods used with density functional-based interaction potentials, to generate full elucidation of complex chemical processes. Mundy's team is gaining a better understanding of chemical reactions in solutions and at interfaces, especially as related to hydrogen storage and catalysis. This research will help establish the future protocol for the application of HPC to present and future grand challenges in the chemical sciences.

2. Drs. Giulia Galli (UC, San Diego) Jeffrey Grossman (UC, Berkeley) and Eric Schwegler (LLNL) are using INCITE cycles to investigate water in confined states by carrying out ab initio simulations for water confined between hydrophilic and hydrophobic surfaces; and by studying the influence of dimensionality reduction and surface chemistry on the properties of the confined fluid (figure ??). The grand challenge is to define a computational paradigm to simulate water flow and transport at the nanoscale which can be applied to both materials science problems (e.g., water in zeolites) and problems of biological interest (e.g., water in contact with amino acids and proteins). While the properties of the bulk fluid are relatively well characterized, much less is known about water confined at the nanometer scale, where conventional experimental probes (neutron diffraction and X-ray scattering) are difficult to use.

SIDEBAR #2
Understanding of the Structure and Function of Ion Channels

Dr. Benoit Roux (University of Chicago and Argonne National Laboratory) is developing theoretical and computational methods to advance scientists' understanding of the structure, dynamics and function of biological macromolecular systems at the atomic level. A 2007 Innovative and Novel Computational Impact on Theory and Experiment (INCITE) award allowed Dr. Roux and his team to extend their work on Kv channels in mammalian cells to more complex membrane proteins. Their study produced exciting results, confirming the hypothesis that the electric field controlling the voltage gating is focused over a particular area, rather than spread throughout the whole thickness of the cellular membrane. As a result, Dr. Roux received a 2008 INCITE award to continue this research on the Blue Gene/P at Argonne and the Cray at Oak Ridge.

Similar voltage-driven molecular motions, occurring on a relatively slow time-scale, are also essential for the function of many important membrane proteins such as transporters, pumps, and channels. But although Kv channels are among the most well characterized molecular machines experimentally, many outstanding issues remain before scientists fully understand the gating transition of a K+ channel. One such issue focuses on the open and closed conformations of the channel. The gating charge tells how those conformations are energetically coupled to the transmembrane potential.

As part of the 2008 award, Roux and his colleagues refined models of the open and closed state, running large-scale simulations. The next challenge will be the conformational pathway for the open and closed gating transition of the channel. Advanced and novel strategies will be essential here in determining the reaction pathway by describing the transition process through a chain of states.