

Stepping Up the Pace of Discovery: the Genomes to Life Program

Marvin Frazier, David Thomassen,
and Aristides Patrinos

*Office of Biological and Environmental Research
U.S. Department of Energy Office of Science*

Gary Johnson and Carl E. Oliver

*Office of Advanced Scientific Computing Research
U.S. Department of Energy Office of Science*

Edward Uberbacher

*Oak Ridge National Laboratory
UT-Battelle, LLC*

Abstract

Genomes to Life (GTL), the U.S. Department of Energy Office of Science's systems biology program, focuses on environmental microbiology. Over the next 10 to 20 years, GTL's key goal is to understand the life processes of thousands of microbes and microbial systems in their native environments. This focus demands that we address huge gaps in knowledge, technology, computing, data capture and analysis, and systems-level integration.

Distinguishing features include (1) strategies for unprecedented, comprehensive, and high-throughput data collection; (2) advanced computing, mathematics, algorithms, and data-management technologies; (3) a focus on potential microbial capabilities to help solve energy and environmental challenges; and (4) new research and management models that link production-scale systems biology facilities in an accessible environment.

This unprecedented opportunity to provide the scientific foundation for solving urgent problems in energy, global climate change, and environmental cleanup demands that we take bold steps to achieve a much faster, more efficient pace of biological discovery.

1. Introduction

The remarkable successes of the Human Genome Project and its spin-offs, by revealing the details of numerous genomes from microbes to plants to mammals, have created a revolution in biology that has no equal in the history of science. Our perspective is *forever changed*.

Genomics is now the starting point for studies in biology, making tractable for the first time a systematic and deep understanding of life's processes. This goal is founded, as is life itself, on the information in the genome, which encodes all of life's molecular machines together with all the control and logic for orchestrating their

deployment. It is this exquisite machinery, operating through a labyrinth of pathways and networks and chemistry and mechanics, that makes the cell and the organism *come alive*. To reach the fundamentally new systems level of understanding will compel substantial changes in the way biology is practiced—its goals, organizational structure, and investment strategy. The experience gained in sequencing genomes, including building the infrastructure for large-scale biology projects, both underpins and exemplifies a path forward to dramatically accelerate the pace of discovery.

While stunning in its impact, our foray into genomes has touched upon only the tiniest fraction of life on earth. In particular, the still immeasurably immense and largely untapped diversity of microbes and microbial communities presents a scientific treasure of vast practical and scientific importance. Comprising about 50% of the earth's biomass [1], microbes are in consequence the foundation of the biosphere, controlling earth's biogeochemical cycles and affecting the productivity of the soil, quality of water, and global climate [2, 3]. Indeed, the ability of this planet to sustain life is completely dependent on their ongoing activity. Understanding the microbial world, therefore, will be key to our energy and climate futures, to developing sustainable modes of living, to developing advanced industrial technologies, and, in no small part, to understanding how life on earth functions.

The diversity and range of the environmental adaptations of microbes mean that they long ago evolved solutions to many problems that we as scientists must now address. Microbes have become, for example, the uncontested masters at harvesting essentially every form of energy available (e.g., from sunlight or rocks). Their sophisticated biochemical capacities can be utilized for transforming wastes and organic matter, cycling nutrients, and, as part of the photosynthetic process, converting sunlight into energy and "fixing" (removing) CO₂ from the atmosphere.

2. Critical societal challenges

The connection between climate change and the microbial biosphere presents a revealing example of the issues at stake. For more than a century, humankind has continued to fundamentally reform the earth's biosphere, in large part by altering the chemical makeup of the atmosphere. The biosphere's response to that change is dominantly managed by microbes. To understand and potentially cope with this change, we need an intimate knowledge of ecosystems and how they are responding to climate and atmospheric change. And we do not have 100 years to accomplish this objective. The Climate Change Research Initiative declared that understanding the role of microbes in sustaining the biosphere is a key science goal [4]. Furthermore, a recent American Academy of Microbiology report calls for the development of new technologies to measure the activity of microorganisms [5], and reaching "a molecular-level understanding of life processes" is a designated national science priority [6].

Achieving a fundamental, whole-systems understanding of life is one of the most daunting challenges in the history of science. Furthermore, as the genome project has well illustrated, the speed of our progress toward the goal will be very greatly affected by how we institutionally approach the challenge. If we avoid an exclusive adherence to historical, pregenomic strategies in favor of a judicious mix of "old" and "new" strategies, we can attain a pace of discovery many times faster and less expensive. Our approach to the study of biology clearly must change radically if we are not to squander most of the opportunity now before us, and we must reap the understanding needed to support key biotechnology applications in energy production, climate stabilization, and environmental reclamation. For example, more than one technology will be required to displace much of our imported oil needs and reduce net atmospheric carbon emissions to zero. Biological solutions will be an important part of the mix, and new systems must be in common use at least within the next half-century (a process that, after the science is done, could take 30 to 40 years and billions of dollars in investments).

3. A DOE approach to the challenges: the Genomes to Life program

The systems biology revolution is proceeding along multiple pathways as different science agencies and the private sector have adopted strategies suited to their particular needs and cultures. Over the past 3 years, the DOE Office of Science has held 20 workshops involving some 500 scientists to advise the department on how it should contribute to the biology revolution and to determine the technological needs, potential applications,

and societal considerations involved. The result was the development of the DOE Genomes to Life (GTL) program [7]. A central focus of GTL is environmental microbiology, and its key goal is to achieve, over the next 10 to 20 years, a basic understanding of thousands of microbes and microbial systems in their native environments [8]. This focus demands that we address huge gaps in knowledge, technology, computing, data capture and analysis, and systems-level integration.

The Genomes to Life program has several distinguishing features, including strategies for unprecedented levels of comprehensive data collection using emerging high-throughput technologies; tightly coupled advanced computing, mathematics, algorithms, and data-management technologies; a unique focus on microbial organisms and systems possessing capabilities for possible solutions to energy and environmental challenges; and implementation of new research and management models that link user facilities dedicated to production-scale systems biology data generation and analysis with a teaming environment for a large community of individual investigators.

The DOE Office of Science has sequenced, or sequenced to high draft, about 80 microbial genomes, while other groups have sequenced another 400 or so [9]. Consistently, about half the genes found are of unknown function or have not been observed previously, suggesting that the number of essentially novel genes we eventually will encounter could range into the tens of millions. And we now appreciate more fully that genes, proteins, molecular machines, regulatory processes, and cells themselves do not work in isolation—emphatically telling us that we cannot continue to study them in isolation only. To derive the principles and mechanisms of life's processes, we must use a systems approach to dynamically analyze all these elements in cells, populations, and communities, and in environments in which they naturally live.

For example, we must image and functionally analyze the critical molecular machines within a microbe as it responds to environmental cues and signals from other microbes in the niche environment, which also must be analyzed. We must then be able to computationally encompass and model our findings constrained by everything we understand about the relevant biophysical principles, genomic information, cellular composition, environment, and life strategies to see the cell and the community in full functional "dissection" (see Fig. 1). An ultimate goal will be to develop predictive models and, among other things, gain enough insight to be able to write the *Bergey's Manual of Systematic Bacteriology* entry for a microbe using its genome only.

While this vision may seem impossible based on past technologies and practices, emerging and future technologies and mathematics and computing tools can

make it achievable if we move aggressively and effectively. Breakthroughs in mass spectroscopy (MS), nuclear magnetic resonance, microtechnologies, automation, imaging, and other tools and methods are starting to give us a systems view of microbial communities. To take on this challenge, we must deploy and manage these resources at a meaningful scale.

Perhaps the greatest challenge is to capture and control the torrent of new data and information that must and will be forthcoming over the next decades. All the technologies we are seeking to develop will generate massive data sets. The only feasible approach to their assimilation and systems-level interpretation is through a comprehensive data-management strategy coupled to new generations of modeling and simulation tools. Many different types of data must be integrated into data and metadata structures that capture great biological richness and range well into the petabyte (billion megabyte) scale, presenting enormous challenges in design, data mining, and visualization.

New tools for making this information useful to all biologists also will be required, including faster and more capable computer architectures tied together in faster networks linked into research laboratories. Substantial advances in mathematical, statistical, and algorithmic methods clearly will be needed to power the new infrastructure. A partnership between DOE's Office of

Biological and Environmental Research (BER) and Office of Advanced Scientific Computing Research (ASCR) has been formed to address this critical need by enlisting a large community of mathematical and computational scientists as partners and by developing a comprehensive computing and information infrastructure. The significance of this partnership goes well beyond mere formalism. It reflects the fact that biology is now becoming a computational science, in addition to retaining its roots in vigorous experiment and theory. Computing is an essential new ingredient key to the future successes of biology.

3.1. Big science vs small science: the need for a new model

The genome revolution presents the biology community with a complex, yet critical, challenge: preserve the creativity and entrepreneurial spirit of the single investigator in the face of increasingly sophisticated and costly resource requirements of leading-edge biological research. Already, only a minority of even large laboratories can afford to be adequately endowed. Individual investigators need the capabilities of big science, and big science needs access to the energy and creativity of individual investigators. The GTL user facilities strategy is designed to help build a bridge

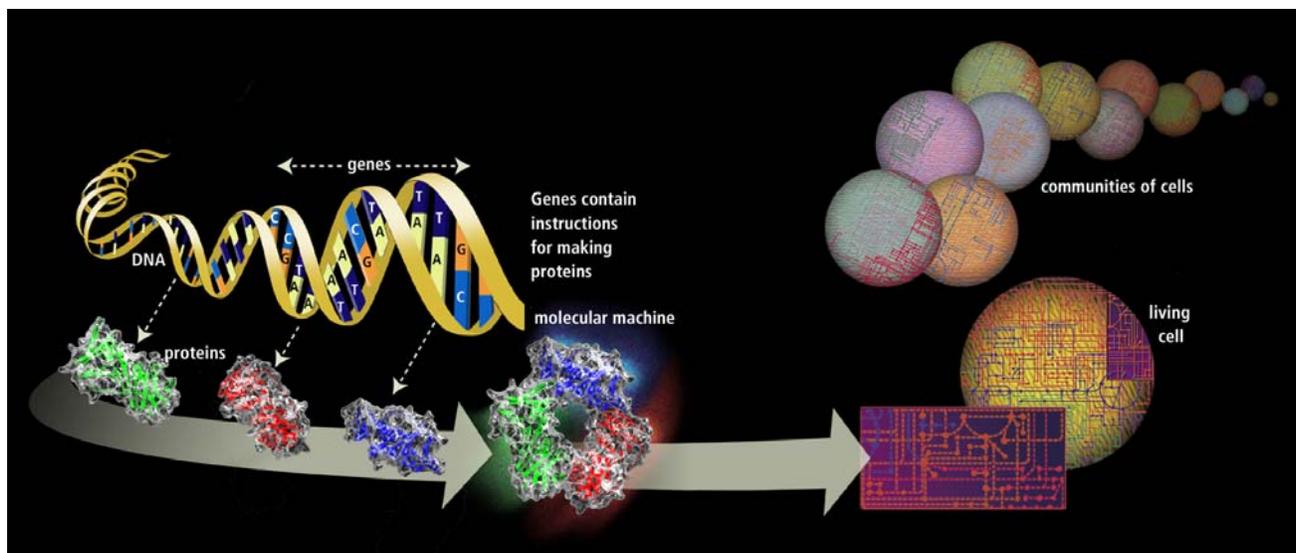


Fig. 1. A primer: From genomes to life. Cells contain DNA, the hereditary material of all living systems. The genome is an organism's complete set of DNA and is organized into chromosomes. DNA contains genes whose sequence specifies how and when to build proteins. Proteins perform most essential life functions, often working together and with other types of molecules as molecular machines. Molecular machines interact through complex, interconnected pathways and networks to make the cell come alive. Communities of cells range from associations of hundreds, perhaps trillions, of microbes (each a single cell) to a single multicellular organism's trillions of cells.

between large and small labs by seeking to make the most sophisticated and comprehensive technologies, materials, and information available to all scientists on an equal basis.

3.2. GTL user facilities planning

Lessons taken from the genome projects heavily influenced our thinking that facilities focused on providing critical data sets and capabilities for all scientists can greatly speed progress and offer unprecedented opportunities for discovery. Such facilities must be driven by *nearly unattainable* cost and production goals that force rapid technology evolution. We are confident that this approach can achieve the necessary dramatically increased productivity, improved data quality, and reduced costs, while providing unprecedented opportunities for discovery. Just as in the beginning of the Human Genome Project when DNA sequencing capability was completely inadequate, the quantity and complexity of data that must be collected and analyzed for systems biology research far exceed current capabilities and capacities. The nation's plan must accommodate that reality, and we must get started now.

A new set of biology user facilities is being planned to provide infrastructure that will enable cells to be studied at increasing levels of biological complexity [10]. In time, these four user facilities will furnish scientists with the means to understand the full functionality of living systems (see Fig. 2).

The first GTL facility will surmount a principal roadblock to whole-system analysis by producing and characterizing microbial proteins by high-throughput means. This facility also will generate the protein-tagging reagents needed to identify, track, quantify, control, capture, and image individual proteins and molecular machines in living systems.

A second facility will provide a comprehensive measure of the microbial proteome as well as views of the entire complement of individual proteins and associated metabolites generated as living microbes respond to varying environmental conditions—a precursor to a systems understanding of functionality.

A third facility will isolate, analyze, and model the multiprotein assemblies that are the critical functional units of all living systems. Virtually all processes in living cells are facilitated by protein interactions, including those of stable assemblies and protein machines, as well as through very transient or dynamic interactions. A comprehensive understanding of these events is essential.

A fourth facility builds on the capabilities of the first three to observe and identify the network of genes, molecular machines, metabolites, and regulatory pathways operating in microbes under a wide range of conditions. This information will enable scientists, for the first time, to predictively and comprehensively model functions of individual microbes and microbial communities.

4. Cross-cutting data, analysis, and computing infrastructure

The GTL program brings us to the verge of a revolution in systems biology not only by providing the high-throughput, high-quality experimental data required for systems biology, but also by developing the computational infrastructure, algorithms, and tools to fully exploit these data and extract knowledge and insights from them. Progress depends substantially on the emergence of a new mathematical, quantitative, predictive, and ultimately systems-level paradigm for the life sciences. Modeling complex biological systems demands new methods to treat the vastly disparate length and time scales of individual molecules, molecular complexes, metabolic and signaling pathways, functional subsystems, individual cells, and, ultimately, even interacting organisms and ecosystems. Meeting these demands will require computing environments of far greater complexity than those commonly used by biological researchers today.

The new paradigm is one in which biologists represent their most fundamental

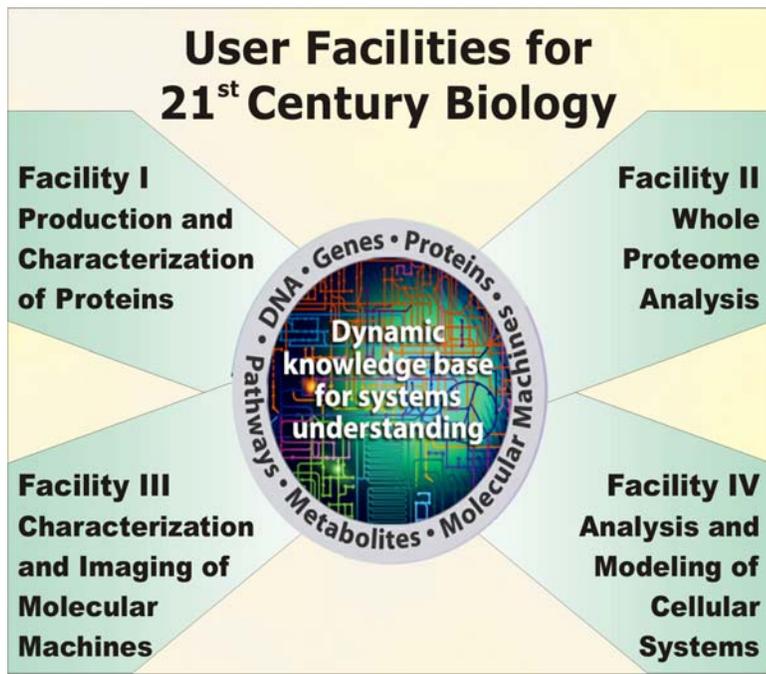


Fig. 2. Planned Genomes to Life user facilities.

knowledge of complex biological systems as mathematically based computer models. Such models will be used to capture and represent data, predict behavior, and generate hypotheses that can be tested by generating more data (see Fig. 3). Biologists will need the means to move data and knowledge back and forth between experiments and their computing environments on an everyday basis and make large-scale computing an integral part of their daily lives. To ask next-generation questions and do next-generation experiments, computing should be guiding the questions and the interpretation at every step. To accomplish this, the GTL program must invest in significant new infrastructure for data management, analysis, and modeling and for simulation. The major components of this infrastructure are described below.

4.1. Data analysis

Genomes to Life user facilities and projects will generate vast amounts of diverse and complex data that must be analyzed, integrated, and interpreted. Data analysis is key to systems biology, and analysis tools and tool frameworks are critical in allowing biologists to derive new insights. The complexity of data-generation modalities is much higher than in the genome era, and the amounts of data generated will be much larger than for the human genome. Achieving the necessary data-analysis

throughput will require significant research, advances in software tools, and better models to share analysis software and resources.

4.2. Modeling and simulation

To understand the astounding complexity of biological systems, extracting and developing general systems principles from systems data, modeling, and simulation are fundamental. Mathematical representations of complex biological systems are essential to the conceptual breakthroughs anticipated in Genomes to Life, where the complexity of systems in even the simplest microbes requires such representation. This mind-boggling complexity must be captured in the computer, and biology must be represented in mathematical ways that parameterize system complexity. New modeling and simulation methods for biological problems are needed to explain and understand biological phenomena. The scale and complexity of these developments will require a coordinated approach to developing algorithms and codes as well as formal methods for tool repositories that manage, support, and maintain codes. The compatibility of codes with hardware platforms and GTL data sources must be ensured.

A current challenge for modelers is to find meaningful ways to initiate modeling with incomplete (and possibly inaccurate) data. This is a challenge in which the applied-mathematics community has significant expertise and potentially can jump-start the modeling of biological systems. Simulating, predicting, and quantitating behavior are necessary to understand biological systems completely and generate further testable hypotheses. Genomes to Life goals will require simulation of heterogeneous biological systems over long time scales and include simulations of molecular complexes, pathways, networks, and, eventually, communities of organisms. Quantitation is needed to test our understanding and representation of systems.

4.3. Data management

High-quality production-scale experimentation will create massive amounts of diverse data and drive the development of large-scale data-intensive

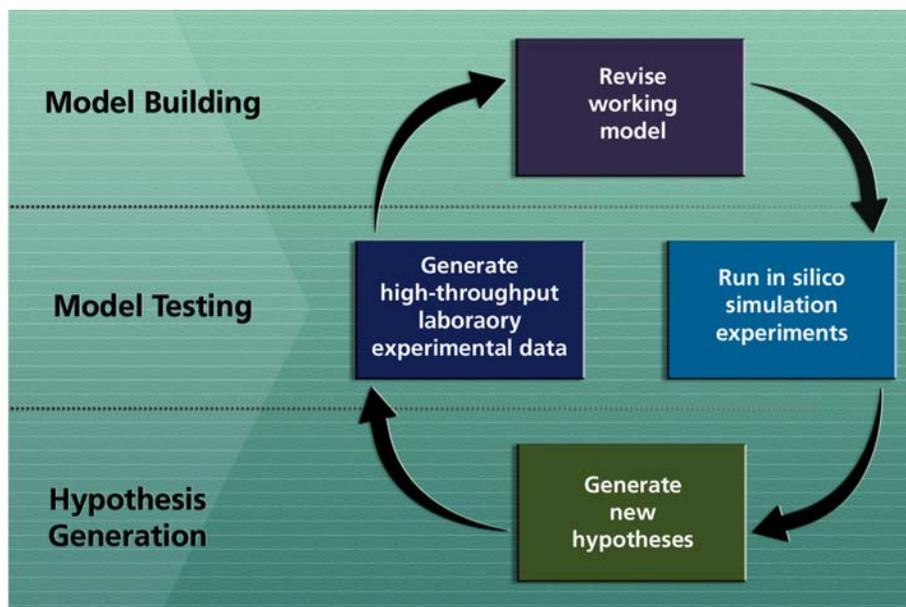


Fig. 3. The goal of systems biology is to create increasingly accurate mathematical models of life processes that enable prediction of cell behavior and, eventually, use of their capabilities. This approach is driven by the availability of whole genome sequences and data provided by high-throughput technologies and computation.

computation. To manage, order, understand, and extract inferences from GTL data will require development of a national data infrastructure. Such an infrastructure, consisting of linked core databases and attendant tools, would be one of the primary GTL fruits available to all biologists. While the data-management challenge is staggering because of the rich relationships in biological data, the immense scale of GTL data sets poses its own challenge. Multiterabyte biological data sets and multipetabyte data archives will be generated by high-throughput technologies. New hardware, software, and design approaches will ensure the availability of such unique and valuable data to the entire biological research community.

4.4. Computing environments and specialized hardware

Getting biologists to embrace computing as a tool is essential for the revolutionary progress in systems biology expected in GTL. The GTL computational infrastructure will utilize both local and distributed computing and experimental facilities. Computing grids link hardware, data, and software components into a single environment tailored to discipline-specific needs. Grids can be designed to provide “intelligent” user environments that permit biologists to easily configure complex analysis and to access specialized computing resources, data, and remote instruments without having expert knowledge of the technical details of the computing environment.

Additionally, many analysis, modeling, and simulation tasks will require the ultimate high-end computing architectures. For example, detailed knowledge of structure and molecular dynamics will be essential for understanding the behavior of molecular machines and protein interactions. Computation already plays a significant role in determining biological structure (e.g., homology modeling) as well as in understanding protein dynamics. As computational methods mature and more powerful computers become available, more fundamental computational approaches to the analysis of biological structures may become feasible. Many expected challenges in GTL will require next-generation computing environments to satisfy the need for biological simulation scale and complexity (see Fig. 4).

The unique set of GTL user facilities and computing resources will provide a national infrastructure for biology and will empower scientists to pursue whole new avenues of inquiry, fundamentally changing the course of biological research and greatly accelerating the pace of discovery. The user facilities also will promote cross-disciplinary education of a new generation of scientists trained with a systems biology perspective.

5. Early results

5.1. Pilot projects and awards

Numerous piloting activities funded by BER over the past 5 years have established a foundation for addressing these challenges and have underscored the need for advanced-technology user facilities accessible by the whole biological research community. Projects thus far include systems biology studies and the development of advanced technologies. These projects have demonstrated MS analysis of microbial proteomes, the development of new imaging modalities, small-scale production of microbial proteins, and the development of computational tools for first-generation genome analysis and annotation.

To go beyond technologies to experiments in systems biology, several federations of scientists are now integrating technologies and using modeling methods to gain an understanding of specific microbes in their natural environments. For example, the *Shewanella* federation, consisting of teams of scientists from academia, national laboratories, and private industry, is probing in detail this remarkably versatile organism capable of immobilizing toxic uranium in groundwater and rendering it biologically unavailable [11, 12]. The federation has made excellent progress in its preliminary proteome analyses, using a combination of MS methods to yield identification of 3862 uniquely expressed proteins (including many post-translational modifications) representing virtually every functional class [13].

Continuing this development, the first Genomes to Life project awards, made in July 2002, are focusing on GTL goals: isolating and characterizing molecular machines and understanding cellular metabolic and regulatory processes and complex biological communities [14].

5.2. Novel applications

The comprehensive understanding of cellular systems attained in systems biology programs such as GTL will generate revolutionary applications of biology. For instance, highly efficient single enzymes or entire metabolic pathways embedded on biomimetic membranes could produce fuels, remove or inactivate contaminants, sequester carbon to mitigate global climate change, process foods, produce chemicals, and achieve separations. One successful example of engineering a biological system for efficient use is represented in Fig. 5.

Genomes to Life Computing Roadmap

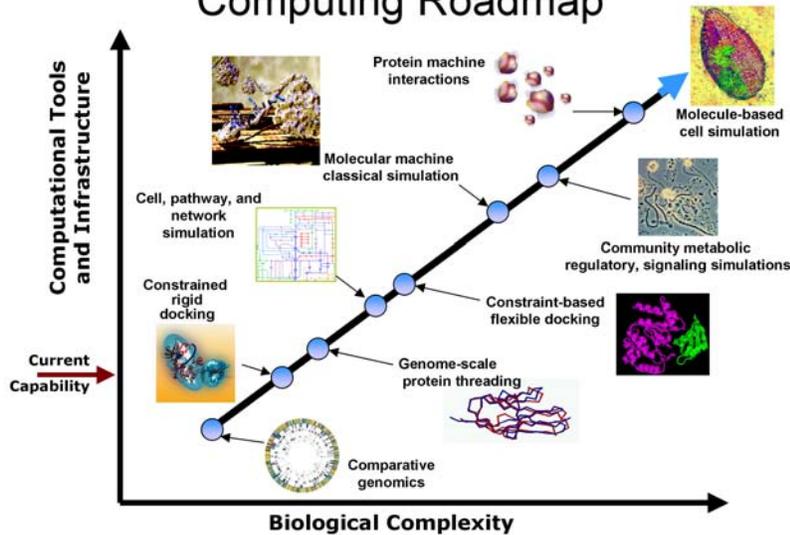


Fig. 4. Concept diagram schematically illustrates a path from basic genome data to a more detailed understanding of complex molecular and cellular systems and the need to develop new computational analysis, modeling, and simulation capabilities to meet this goal. The points on the plot are very approximate, depending on the specifics of problem abstraction and computational representation. Research is under way to create mathematics, algorithms, and computer architectures for understanding each level of biological complexity.

Key

Comparative genomics. Genome-scale sequence analysis and comparisons used to identify genes and make an initial estimate of gene function.

Constraint-based rigid docking. Computational process in which possible geometries that interact with macromolecules are explored and evaluated in energetic terms. *Constraints* can be used to limit the search for such possible interactions (e.g., knowing the surface interaction location on one protein). *Rigid* refers to molecules' not being allowed to change conformations during the computed interaction, thus simplifying the computation.

Genome-scale protein threading. Process whereby protein sequences derived from a newly sequenced genome are evaluated through an alignment procedure against a database of known protein structures or protein folds. If such a match is energetically favorable, the known structure with its alignment can provide an approximate three-dimensional structure for the new sequence.

Cell, pathway, and network simulation. Variety of mathematical, differential-equation, stochastic, and flux-based models for cellular metabolism and gene regulation, as well as models that combine multiple aspects of cell modeling and simulation.

Constraint-based flexible docking. Like *constraint-based rigid docking*, a computational process in which possible geometries that interact with macromolecules are explored and evaluated in energetic terms. Again, *constraints* can be used to limit the search. *Flexible*

refers to molecules' being allowed to change conformations during the computed interaction. Many interaction surfaces are disordered before the interaction partner is introduced. *Flexible docking*, although computationally very expensive, is needed to capture and computationally model such interactions.

Molecular machine classical simulation. Molecular dynamics simulation of a protein or modest protein complex with duration sufficient for observation of key functional elements. The classical molecular dynamics method applies nonquantum mechanical physical laws of energy and motion to molecular structures.

Protein machine interactions. Molecular dynamics simulation of a complex set of proteins or interactions among large assemblies of macromolecules.

Community metabolic, regulatory, and signaling simulations. Methods that apply modeling at the cellular level (see *Cell, pathway, and network simulation*, above) to large microbial communities of potentially different types and conditions. These processes are designed to capture dependencies among organisms in complex communities.

Molecule-based cell simulation. Refers to the long-range goal of simulating a cell at comprehensive molecular detail, in which all molecular types are present and stochastic interactions and events are tracked.

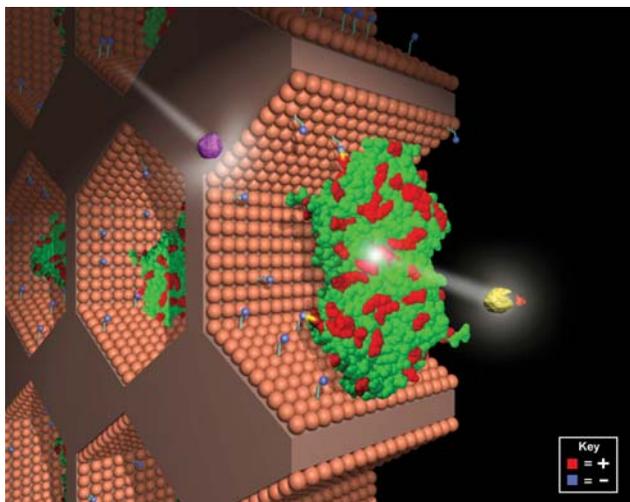


Fig. 5. A single enzyme (organophosphorus hydrolase, OPH), embedded in a nanomembrane, functions more efficiently and longer than enzymes in vivo to render pesticides inert [15]. The knowledge gained from Genomes to Life research could enable the development of nanostructures containing microbial enzymes or whole pathways of enzymes to meet DOE mission challenges in energy and environment as well as providing useful applications in food processing, separations, and the production of industrial chemicals and pharmaceuticals. [Image courtesy of M. Perkins, Pacific Northwest National Laboratory]

6. Summary

Knowledge is power—but only if you use it. The vast amount of information contained in the hundreds of sequenced genomes, and the thousands to come, offers an unprecedented opportunity for understanding complex biological systems. Exciting new avenues are opening toward solving some of our most urgent problems in healthcare, national security, agriculture, energy, the environment, and industry. Expediently addressing these challenges demands that we take bold steps now to achieve a new, much faster, and more efficient pace of biological discovery. We cannot afford otherwise.

References

- [1] W. B. Whitman, D. C. Coleman, and W. J. Wiebe, "Prokaryotes: The Unseen Majority," *Proc. Natl. Acad. Sci. U.S.A.* **95**, 1998, p. 6578.
- [2] J. Copley, News Feature: "All at Sea," *Nature* **415**, Feb. 7, 2002, p. 572.
- [3] D. K. Newman and Jillian F. Banfield, "Geomicrobiology: How Molecular-Scale Interactions Underpin Biogeochemical Systems," *Science* **296**, 2002, p. 1071.
- [4] Charles Kennel's breakout group report at the Climate Change Science Program's Planning Workshop for Scientists and Stakeholders, December 3–5, 2002, Washington, D.C. (<http://www.climate-science.gov/Library/workshop2002/closingsession/kennel-5dec2002.htm>).
- [5] American Society for Microbiology, *Microbial Ecology and Genomics: A Crossroads of Opportunity*, 2001, p. 7.
- [6] FY 2004 Interagency Research and Development Priorities, Office of Science and Technology Policy and Office of Management and Budget, 2002.
- [7] Genomes to Life: <http://DOEGenomesToLife.org>
- [8] *Genomes to Life: Accelerating Biological Discovery* (programmatic roadmap), U.S. Department of Energy, 2001.
- [9] <http://www.ornl.gov/microbialgenomes/organisms.html>
- [10] "User Facilities for 21st Century Systems Biology: Providing Critical Technologies for the Research Community," Presentation to the Biological and Environmental Research Advisory Committee, Dec. 3, 2002 (<http://DOEGenomesToLife.org/pubs/GTLFac34BERAC45.pdf>).
- [11] www.shewanella.org.
- [12] J. M. Tiedje, "Shewanella—the Environmentally Versatile Genome," *Nat. Biotechnol.* **20**, 2002, p. 1093.
- [13] Personal communication from Richard D. Smith, Pacific Northwest National Laboratory.
- [14] DOEGenomesToLife.org/research/index.html.
- [15] C. Lei, Y. Shin, J. Liu, and E. Ackerman, "Entrapping Enzyme in a Functionalized Nanoporous Support," *J. Am. Chem. Soc.* **124**(38), 2002, p. 11242–43.

Author contact information

Ari.Patrinis@science.doe.gov
 David.Thomassen@science.doe.gov
 Marvin.Frazier@science.doe.gov

Ed.Oliver@science.doe.gov
 Gary.Johnson@science.doe.gov

Uberbacherec@ornl.gov