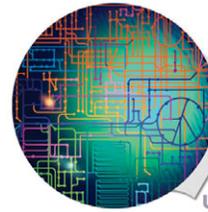


DOE Genomics:GTL

Merging Technologies, Computing, and Microbial Systems to Build a New Foundation for Energy and Environmental Solutions



DOE GENOMICS:GTL
ACCELERATING
DISCOVERY FOR ENERGY
AND ENVIRONMENT
OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

U.S. Department of Energy Office of Science

DOEgenomes.org

The remarkable successes of the Human Genome Project and spin-offs revealing the details of numerous genomes—from microbes to plants to mice—provide the richest resource in the history of biology. These achievements now empower scientists to address the ultimate goal of modern biology: to obtain a fundamental, comprehensive, and systematic understanding of life. This goal is founded, as is life itself, on the genome, whose genes encode the proteins that carry out most cellular activities via a labyrinth of pathways and networks that make the cell “come alive” (see figure below).

Catalyzing Systems Biology

The Department of Energy’s (DOE) Genomics:GTL (GTL) program* is combining high-throughput advanced technologies and computation with the information found in microbial genomes to establish a foundation for achieving an understanding of living systems (see “Microbes for DOE Missions,” p. 2). GTL is designed to help launch biology onto a new trajectory to comprehensively understand cellular processes in a realistic context. This level of exploration, known as systems biology, will empower scientists to pursue innovative approaches to discovery, transforming biology to a more quantitative and predictive science. The scientific goals of GTL target the fundamental processes of living systems by studying them on three levels:

1. Proteins and multicomponent molecular machines that form all of the cell’s structures and perform most of the cell’s work.
2. Gene regulatory networks and pathways that control cellular processes.
3. Microbial communities in which groups of cells conduct complex processes in nature.

*Formerly Genomes to Life.

Molecular machines perform chemical reactions, generate mechanical forces, transport metabolites and ions, and make possible every action of a biological system. A cell does not generate its entire repertoire of molecular machines at once. Genomic regulatory elements dictate the particular set produced according to the organism’s life strategy and in response to environmental cues, including other microbial populations in the larger ecological community.

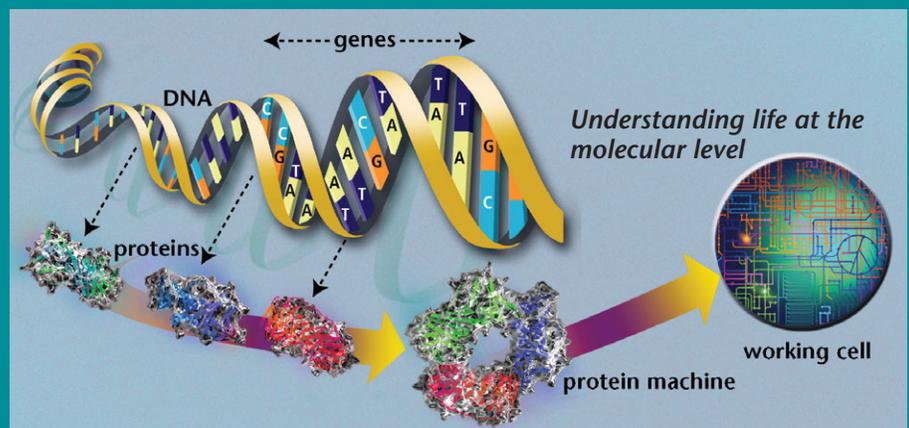
A comprehensive approach to understanding biological systems thus extends from individual cells to many cells functioning in communities. Such studies must encompass proteins, molecular machines, pathways, networks, cells, and, ultimately, their regulatory elements, cellular systems, and environments. This next generation of biology is viable only with vastly increased computational and informational capabilities to master the full complexities of biological systems.

Understanding life processes at the molecular level is a “national science priority.”

—OSTP-OMB FY 2005 budget guidance memo; see p. 6.

Catalyzing systems biology	1
Transforming biology with large-scale technologies and computing	2
Emerging technologies and computing for systems biology	3–4
Integrated user facilities democratizing access to systems biology resources	5
A growing mandate for molecular studies	6

From DNA Sequence to Living Systems



Genes are made up of DNA and contain the information used by other cellular components (e.g., RNA and ribosomes, not shown here) to create proteins. A working cell is tightly packed with tens of thousands of proteins and other molecules, often working together as multimolecular “machines” to perform essential cellular activities (see also cell figure, p. 5).

Transforming Biology

Just as DNA sequencing capability was completely inadequate at the beginning of the Human Genome Project (HGP), the quantity and complexity of data that must be collected and analyzed for systems biology research far exceed current capabilities and capacities. The HGP taught that aspects of biological research can be made high-throughput and cost-effective (see graph, p. 5). Collecting and using such data and reagents will require an organizational model that coordinates and integrates dozens of high-throughput technologies and approaches, some not yet refined or even developed. This is the central principle of GTL and indeed of all systems biology research.

Analysis of living systems will require a new generation of experimentation and the computational methods and capabilities to assimilate, understand, and model the data on the scale and complexity of real living systems. Computing must guide the research questions and interpretation at every step.

The knowledge base resulting from GTL will provide the entire research community with data, models, and simulations of gene expression, pathways, and network systems; molecular machines; and cell and community processes. These new capabilities and resources will transform the entire life sciences landscape, providing a new foundation for many biotechnology applications and inspiring revolutionary solutions to DOE mission needs.

Microbes for DOE Missions: Energy Security, Cleanup, Climate Change

Why Study Microbes?

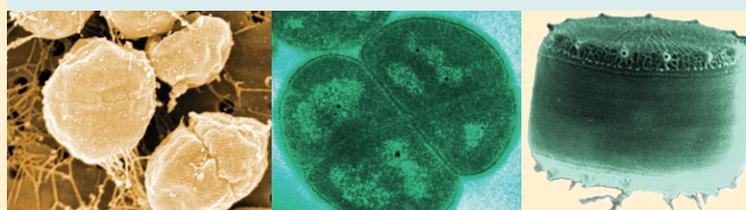
The ability of earth to sustain life is largely dependent on microbes. They are the foundation of the biosphere, controlling earth's biogeochemical cycles and affecting the productivity of the soil, quality of water, and global climate. An understanding of the roles microbes play in these dynamic and interconnected processes, however, has been limited by scientists' inability to grow them in the lab; estimates are that only about 1% of microbes are currently known.

Scientists are beginning to use DNA-based approaches to identify microbial genomes, genes, and functions present at a particular time in a specific environment or microenvironment. Such "ecogenomic" studies in the open seas, for example, revealed recently some 150 completely new types of bacteria and more than one million new genes (about 10 times the number of genes previously studied). In another new research path, nanoscale ecogenomic sensors will monitor microbial communities in the ocean and link their activities to environmental conditions. These investigations will yield critical information on ocean processes, including those affected by climate change.

These and other GTL projects are making microbial research the most exciting frontier in biology today. In addition to achieving a deeper and more profound understanding of how life works on this planet, future discoveries about the vast microbial world hold great promise for finding new sources of clean energy (e.g., hydrogen), generating more effective cleanup and industrial processes, and for other biological applications to DOE missions.

The Challenge

Microbes have become masters at living in almost every environment, harvesting energy in almost any form. Their sophisticated biochemical capabilities can be utilized for transforming wastes and organic matter, cycling nutrients, and, as part of the photosynthetic process, converting sunlight into energy and "fixing" (storing) CO₂ from the atmosphere. The analytical complexity involved in understanding these processes is enormous. Thousands of microbes have capabilities of interest. Moreover, each microbial genome contains thousands of genes capable of producing an even-greater number of proteins. These proteins combine to form innumerable molecular machines in myriad pathways and networks, many of which carry out biological processes useful for DOE missions (see "Potential Applications of GTL," p. 3).



Methanococcus jannaschii: Produces methane, an important energy source; contains enzymes that withstand high temperatures and pressures; possibly useful for industrial processes.

Deinococcus radiodurans: Survives extremely high levels of radiation and has great potential for radioactive waste cleanup.

Thalassiosira pseudonana: Ocean diatom that is a major participant in biological pumping of carbon to ocean depths and has potential for mitigating global climate change.

Establishing a Firm Foundation for Genomics:GTL

Numerous projects funded by the Office of Biological and Environmental Research (BER) over the past 5 years have established a strong foundation for GTL. These projects underscore the need for high-throughput biological research and novel computational approaches. They also are demonstrating the power of mass spectrometric analyses of whole microbial proteomes, developing new imaging methods, advancing the use of microarrays for expression analyses, exploring scalable ways to generate microbial proteins, and developing computational tools for second-generation genome analysis and annotation.

Several collaborative groups are integrating technologies and computational modeling to gain a systems understanding of specific microbes in their natural environments. For example, the *Shewanella* Federation, consisting of teams from academia,

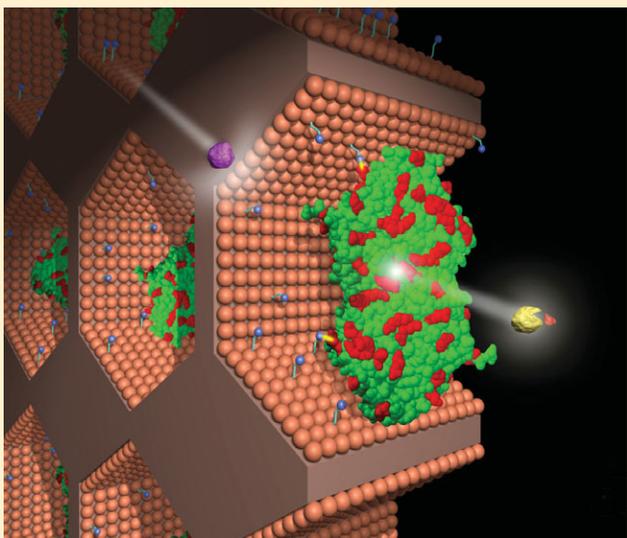
national laboratories, and other organizations, is making progress in preliminary proteome and expression analyses of this remarkably versatile organism that can immobilize toxic uranium from ground water. By focusing multiple technologies on a single organism, the federation is integrating diverse experimental results into a multidimensional perspective of the biology of this key microbe. Using ultrahigh-resolution mass spectrometry techniques, the group thus far has identified over 70% of the predicted proteome for *Shewanella* (as well as for other environmentally important microbes). These and other groundbreaking BER projects (e.g., on *Deinococcus radiodurans*) have elucidated the highest percentages of organism proteomes studied to date. These projects also have set the stage and identified the need for developing high-throughput user facilities accessible to the biological research community (see p. 5).

Office of Science—At the Forefront of the Biological Revolution

In 1986 the DOE Office of Science launched the Human Genome Project to understand, at the DNA level, the effects of energy production on human health. The HGP's innovative operational model proved highly successful, and benefits far exceeded the original goal. Today, DOE is poised to take the next vital steps—translating the genetic code in DNA into a new understanding of how life works and applying those biological processes to serve its challenging missions. Effective use of microbial and other biological systems and components will generate new biotechnological industries involving fuels, biochemical processing, nanomaterials, and broader environmental and biomedical applications.

The Office of Science has the capabilities and institutional traditions to bring the biological, physical, and

computing sciences together at the scale and complexity required for success. Its academic affiliations, national laboratories, and other resources include major facilities for DNA sequencing and molecular-structure characterization, the high-performance computing resources of the Office of Advanced Scientific Computing Research (OASCR), the expertise and infrastructure for technology development, and a tradition of productive multidisciplinary research essential for such an ambitious research program. In the effort to understand biological systems, these strong assets and GTL will complement and extend the capabilities and research efforts supported by the National Institutes of Health, National Science Foundation, other agencies and institutions, and industry.

Potential Applications of GTL

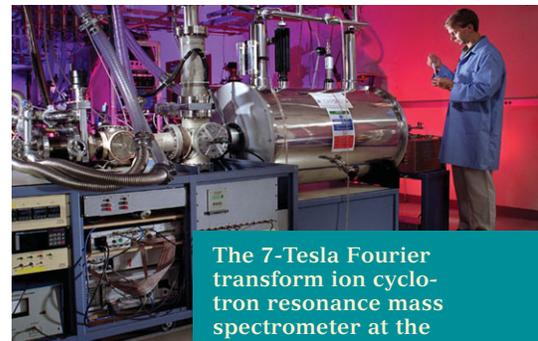
Learning about the inner workings of microbes and their diverse inventory of molecular machines can lead to the discovery of ways to isolate and use these components to develop synthetic nanostructures that carry out some of the functions of living cells. In this figure, the enzyme organophosphorus hydrolase (OPH) has been embedded in a synthetic nano-membrane (mesoporous silica) that enhances its activity and stability [*J. Am. Chem. Soc.* 124, 11242–43 (2002)]. The OPH transforms toxic substances (purple molecule at left of OPH) to harmless by-products (yellow and red molecules at right). Applications such as this could enable development of efficient enzyme-based ways to produce energy, remove or inactivate contaminants, and sequester carbon to mitigate global climate change. The knowledge gained from DOE genomics research also could be highly useful in food processing, pharmaceuticals, separations, and the production of industrial chemicals.

DOE Genomics:GTL Awards

GTL continues to build its R&D portfolio, with 45 awards totaling \$135.1M for FY 2002 through FY 2007. Projects were chosen to test the concept of systems biology and to demonstrate advanced technologies (see picture at right), computation, and potential high-throughput methods in areas having possible impact on DOE missions.

Seven major projects focus on isolating and characterizing protein machines, understanding complex biological communities, modeling cellular metabolic and regulatory processes, modeling carbon sequestration processes in marine microbes, and sequencing ocean microbial communities. Microbes studied in these and pilot projects have potential for bioremediating metals

and radionuclides, degrading organic pollutants, producing energy feedstocks including biomass and hydrogen, sequestering carbon, and demonstrating importance in ocean carbon cycling.



The 7-Tesla Fourier transform ion cyclotron resonance mass spectrometer at the William R. Wiley Environmental Molecular Sciences Laboratory. Mass spectrometry is the most sensitive method for identifying proteins.

Major Projects FY 2002–FY 2007

For more details, see doegenomestolife.org/research

- **Oak Ridge National Laboratory:** Developing technologies needed to identify and characterize the complete set of multiprotein complexes in microbes involved in the carbon cycle (important for climate) and cleaning up metals in contaminated soil.
- **Lawrence Berkeley National Laboratory:** Developing computational models that describe and predict the behavior of gene regulatory networks in microbes in response to environmental conditions found at DOE waste sites.
- **Sandia National Laboratories:** Developing experimental and computational methods to understand the proteins, protein-protein interactions, and gene regulatory networks in a marine microbe that plays a significant role in earth's carbon cycle.
- **University of Massachusetts, Amherst:** Developing computational models to predict the activity of natural communities of microbes having potential for uranium bioremediation and production of electricity.
- **Harvard Medical School:** Developing computational methods to understand the proteins, protein-protein interactions, gene regulatory networks, and community behavior of microbes active in the carbon cycle.
- **Shewanella Federation:** Characterizing and modeling the biology of the metabolically versatile bacterium *Shewanella oneidensis* MR-1 by a cross-institutional consortium. Has potential for bioremediating metals and radionuclides.

- **Institute for Biological Energy Alternatives:** Sequencing the collective genomes of microbial communities in the Sargasso Sea, a genomics approach to discovery of metabolic diversity; developing a synthetic minimal genome; and investigating photosynthesis and hydrogen production.

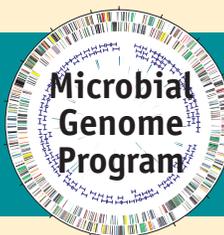
Collaborating Institutions

Argonne National Laboratory
 BIATECH
 Brigham and Women's Hospital
 Diversa Corporation
 Los Alamos National Laboratory
 Massachusetts General Hospital
 Massachusetts Institute of Technology
 Michigan State University
 Pacific Northwest National Laboratory
 National Center for Genome Resources

The Institute for Genomic Research
 The Molecular Science Institute
 University of California (Berkeley, San Diego, Santa Barbara)
 University of Illinois
 University of Michigan
 University of Missouri
 University of North Carolina
 University of Southern California
 University of Tennessee (Knoxville, Memphis)
 University of Utah
 University of Washington
 Woods Hole Oceanographic Institution

Microbes studied in GTL have had their genetic sequences determined under DOE's Microbial Genome Program.

www.ornl.gov/microbialgenomes

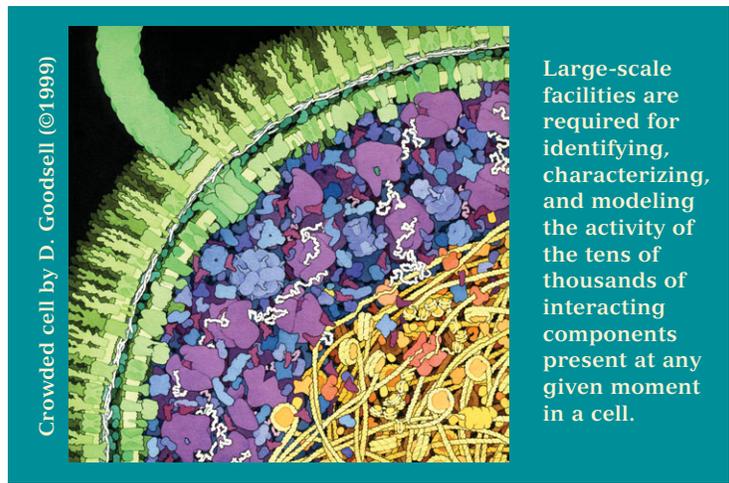
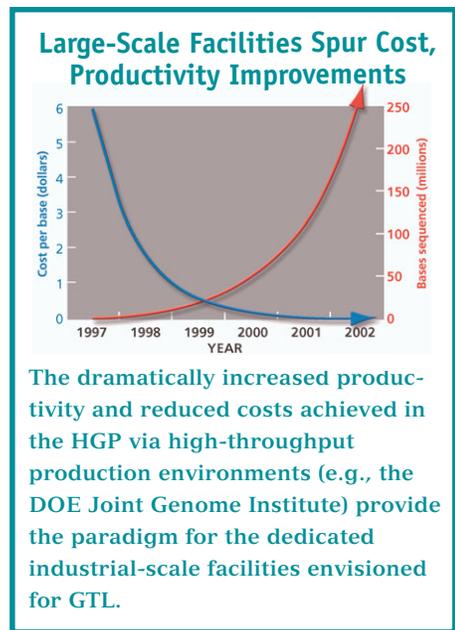


A Plan to Democratize Access to Systems Biology Resources

Analyzing whole microbial systems requires economies of scale. Traditionally, scientists have tried to understand the functions of individual proteins or small groups of proteins. In the new era of systems biology, researchers will study the behavior of the cell's entire working complements of proteins (proteomes), their regulatory pathways, and their interactions as they perform functions. These activities must be carried out on a scale that far exceeds today's capacities.

To meet this challenge, BER and OASCR have planned a set of four core research facilities. Building on each other, these facilities are intricately linked in their long-term goals, targets, technologies, capabilities, and capacities. They will provide scientists with an enduring comprehensive ability to understand and, ultimately, reap enormous benefit from the biochemical functionality of microbial systems. Making the most advanced technologies and computing resources available to

scientists in large and small laboratories will democratize access to the tools needed for systems biology. The facilities thus open new avenues of inquiry, fundamentally changing the course of biological research and greatly accelerating the pace of discovery. Hallmarks of these facilities include high-throughput advanced technologies, automation, and tools for data management and analysis, simulation, and an integrated knowledge base.



DOE's Plan for Genomics User Facilities

www.doegenomestolife.org/research/facilities.shtml

Facility for Production and Characterization of Proteins and Molecular Tags would use highly automated processes to mass-produce and characterize proteins directly from microbial genome data and create affinity reagents ("tags") to identify, capture, and monitor the proteins from living systems.

Facility for Characterization and Imaging of Molecular Machines would isolate, identify, and characterize thousands of molecular machines from microbes and develop the ability to image component proteins within complexes and to validate the presence of the complexes within cells.

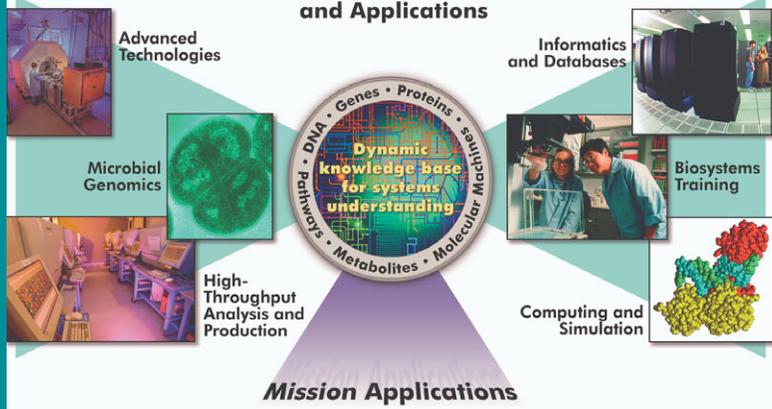
Facility for Whole Proteome Analysis would characterize the expressed proteomes of diverse microbes under different environmental conditions as an essential step toward determining the functions and interactions of individual proteins and sets of proteins.

Facility for Analysis and Modeling of Cellular Systems and Microbial Community Dynamics would combine advanced computational, analytical, and experimental capabilities for the integrated observation, measurement, and analysis of spatial and temporal variations in the structures and functions of cellular systems—from individual microbial cells to complex communities and multicellular organisms.

Hallmarks of User Facilities

Open Access to Data and Facilities

Technology Integration for Tomorrow's Biology and Applications



Mission Applications

These facilities would serve as focal points for the life sciences research community, providing a national venue to pursue multidisciplinary systems biology and promote cross-disciplinary education.

A Growing Mandate for Molecular Studies

OSTP, OMB: "National Science Priority"

Achieving a molecular-level understanding of life processes is a national science priority, according to the Office of Science and Technology Policy (OSTP) and Office of Management and Budget (OMB) *FY 2005 Interagency Research and Development Priorities*. This guidance reflects views found throughout much of the biological research community.

DOE Office of Science: Facilities "Extend the Frontiers of Science"

In November 2003, Genomics:GTL facilities were among the top priorities identified in the 20-year plan of DOE's Office of Science (*Facilities for the Future of Science: A Twenty-Year Outlook*, www.sc.doe.gov). "Investment in these facilities," noted Secretary of Energy Spencer Abraham, "will yield extraordinary scientific breakthroughs and vital societal and economic benefits."

The plan was developed through discussions with the White House Office of Science and Technology Policy, the Office of Management and Budget, and members of Congress, with the assistance of the U.S. scientific community as represented by Office of Science advisory committees. The DOE Office of Science supports basic research to serve the department's missions.

AAM: "Develop New Technologies"

Specific recommendations made by the American Academy of Microbiology (AAM) in its 2001 colloquium report, *Microbial Ecology and Genomics: A Crossroads of Opportunity*, include the following:

- "Develop new technologies including methods for measuring the activity of microorganisms (at the level of populations and single cells), approaches to cultivating currently uncultivable species, and methods for rapid determination of key physiological traits and activities.
- "Establish mechanisms to encourage the necessary instrument development.
- "Encourage instrumentation development through collaboration with device engineers, chemists, physicists, and computational scientists, since uncovering the diversity and activities of the microbial world is dependent on such advances.
- "Develop technology and analysis capability to study microbial communities and symbioses holistically, measuring system-wide expression patterns (mRNA and protein) and activity measurements at the level of populations and single cells."

GTL Program Development

Genomics:GTL (GTL) is a joint program of the Office of Biological and Environmental Research (BER) and the Office of Advanced Scientific Computing Research (OASCR) within the DOE Office of Science.

In the past 3 years, scientists from industry, national laboratories, and academia have participated in some 29 workshops to provide guidance in developing GTL. In April 2002, the BER Advisory Committee (BERAC) approved a strategic plan for advanced and high-throughput facilities to serve the program and the entire scientific community. BERAC voiced approval of the subsequent facilities plan in December 2002. Workshops are identifying capabilities and capacities needed for the facilities.

History: In 1999 the Office of Science charged BERAC with defining DOE's potential roles in post-HGP science. In response, DOE published the report, *Bringing the Genome to Life*, in August 2000. This report set forth recommendations that led to initiation of the Genomes to Life program and publishing of its *Genomes to Life* roadmap in April 2001. Funding from BER and OASCR grew from \$21.7M in FY 2002 to \$71.3M in FY 2004. The FY 2005 budget proposed in the President's Request to Congress is \$80M.



U.S. Department of Energy Office of Science

Marvin Frazier
Office of Biological and Environmental Research
(BER: SC-72)
301/903-5468, Fax: 301/903-8521
marvin.frazier@science.doe.gov

Gary Johnson
Office of Advanced Scientific Computing Research
(OASCR: SC-30)
301/903-5800, Fax: 301/903-7774
gary.johnson@science.doe.gov

Web site for this document:

- DOEgenomestolife.org/pubs/overview.pdf