DOE GENOMICS:GTL
SYSTEMS BIOLOGY
FOR ENERGY AND
ENVIRONMENT

OFFICE OF SCIENCE
U.S. DEPARTMENT OF ENERGY

This document

has been updated.

New file is available at

http://doegenomestolife.org/pubs.shtml

November 2005

DOE/SC-0036

GENOMES to LIFE

ACCELERATING BIOLOGICAL DISCOVERY

Program proposed by the
Office of Biological and Environmental Research
and
Office of Advanced Scientific Computing Research
of the
U.S. Department of Energy
April 2001

Note: Archival document. Some aspects of the Genomes to Life (GTL) program have changed since publication of this April 2001 roadmap. This document provides a historical perspective on the development of the program and reflects early thinking on its broadest possibilities and impacts. The refined and current focus of GTL is on microbes relevant to DOE needs in energy production, environmental cleanup, and carbon sequestration. For a current overview of the program, see "Genomes to Life: Realizing the Potential of the Genome Revolution" and Funding Announcements.

Prepared by the Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, Tennessee

DOEGenomesToLife.org

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

Having the complete DNA sequences of genomes for organisms ranging from humans to mice to microbes now brings us to perhaps the greatest scientific frontier ever. The aspiration of the biology for the 21st century is to build from the foundation of whole-genome sequences a new, comprehensive, and profound understanding of complex living systems.

This objective can be achieved only by joining revolutionary technologies for systems-level and computational biology. A central goal of the Genomes to Life program introduced in these pages is to establish, within a decade, a national infrastructure to transform the tremendous outpouring of data and concepts into a new computationally based biology. The U.S. Department of Energy's (DOE) offices of Biological and Environmental Research (BER) and Advanced Scientific Computing Research have formed a strategic alliance to meet this grand challenge.

Genomic and advanced technological resources provide an opportunity for DOE to more effectively address its broad mission needs—produce energy, sequester excess atmospheric carbon that contributes to global warming, clean up environments contaminated from weapons production, and protect people from energy byproducts such as radiation and from the threat of bioterrorism. Until now, solutions have focused on physical and engineering strategies, but many of these missions have a basis, and possibly a solution, in the biological world.

Microbes, for example, make up most of the earth's biomass, have evolved for some 3.7 billion years, and have been found in virtually every environment. The diversity and range of their adaptations mean that microbes long ago "solved" many problems for which scientists are still actively seeking solutions. Their capabilities will offer an astonishingly diverse set of biological tools.

In this booklet, we offer a roadmap for these new explorations in "systems biology." The 10-year program aims to use DNA sequences from microbes and higher organisms, including humans, as starting points for systematically tackling questions about critical life processes. Success in this quest will require joining powerful new biological, mathematical, computational, engineering, and physical concepts, approaches, and technologies and using the capabilities of other federal agencies as well.

DOE facilities and research supported at its national laboratories and in academic institutions played key enabling and scientific roles in the genomics revolution. We are again poised to make important contributions to the next revolution in biology. We are grateful to the many scientists who contributed to the development of this new program—they are the pioneers who will help lead the way.

This roadmap was prepared under the auspices of BER in response to recommendations set forth in the BER advisory subcommittee's report "Bringing the Genome to Life" (August 2000). The subcommittee, chaired by Ray Gesteland (University of Utah), acted in response to a letter from the director of the DOE Office of Science (November 1999).

AAN Patrinos

Aristides A. N. Patrinos, Associate Director
Office of Biological and Environmental Research
U.S. Department of Energy
ari.patrinos@science.doe.gov

Carl Edward Oliver

Edward Oliver, Associate Director
Office of Advanced Scientific Computing Research
U.S. Department of Energy
ed.oliver@science.doe.gov

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

Genomes to Life Contents

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

GENOMES to LIFE

ACCELERATING BIOLOGICAL DISCOVERY

Executive Summary

Built on the continuing successes of international genome-sequencing projects, the Genomes to Life program will take the logical next step: a quest to understand the composition and function of the biochemical networks and pathways that carry out the essential processes of living organisms. This roadmap sets forth an aggressive 10-year plan designed to exploit high-throughput genomic strategies and centered around four major goals:

- Identify and characterize the molecular machines of life—the multiprotein complexes that execute cellular functions and govern cell form.
- Characterize gene regulatory networks.
- Characterize the functional repertoire of complex microbial communities in their natural environments at the molecular level.
- Develop the computational methods and capabilities to advance understanding of complex biological systems and predict their behavior.

The Genomes to Life program reflects the fundamental change now occurring in the way biologists think about biology, a perspective that is a logical and compelling product of the Human Genome Project. The new program will build on the Human Genome Project, both by exploiting its data and by extending its paradigm of comprehensive, whole-genome biology to the next level. This approach ultimately will enable an integrated and predictive understanding of biological systems—an understanding that will offer insights into how both microbial and human cells respond to environmental changes. The applications of this next level of understanding will be revolutionary.

The current state-of-the-art instrumentation and computation enable and encourage the immediate establishment of this ambitious and far-reaching program. However, concurrent technology development will be needed to reach all goals within the next decade. Substantial efforts will be devoted, for example, to improving technologies for characterizing proteins and protein complexes, localizing them in cells and tissues, carrying out high-throughput functional assays of complete cellular protein inventories, and sequencing and analyzing microbial DNA taken from natural environments.

The Genomes to Life program complements and augments the DOE Microbial Cell Project, launched in FY 2001. The goal of this established project is to collect, analyze, and integrate data on individual microbes in an effort to understand how cellular components function together to create living systems, particularly those with capabilities of interest to the DOE.

DOE is strongly positioned to make major contributions to the scientific advances promised by the biology of the 21st century. Strengths of DOE's national laboratories include major facilities for DNA sequencing and molecular structure characterization, high-performance computing resources, the expertise and infrastructure for technology development, and a legacy of productive multidisciplinary research essential for such an ambitious and complex program. In the effort to understand biological systems, these assets and the Genomes to Life program will complement and fundamentally enable the capabilities and efforts of the National Institutes of Health, the National Science Foundation, and other agencies and institutions around the world.

GENOMES to LIFE

ACCELERATING
BIOLOGICAL
DISCOVERY

Introduction

The remarkable successes of the Human Genome Project, whose origins can be traced to a Department of Energy (DOE) initiative launched in 1986, provide the richest intellectual resource in the history of biology. In June 2000 in two national capitals, the draft sequence of the human genome was announced as complete. Further, the Human Genome Project has generated the enabling tools—and the scientific will—to produce whole-genome catalogs for many microbes, the plant *Arabidopsis thaliana*, the fruit fly *Drosophila melanogaster*, the roundworm *Caenorhabditis elegans*, and soon the pufferfish *Fugu rubripes*.

Genomes are made of DNA—entwined strands of molecules known as nucleic acids that store the information each organism needs to grow, develop, and function. Obtaining the DNA sequence of the entire human genome, along with those of scores of microbes and other organisms, stands as one of the greatest achievements of the 20th century. And yet, these complete genome sequences, the “recipes for life,” serve merely as a foundation for the biology of the 21st century, the departure point for an effort aimed at the most far-reaching of all biological goals:

> *Achieve a fundamental, comprehensive,*
> *and systematic understanding of life.*

The Genomes to Life program within DOE’s Office of Science will be an important part of this effort. Jointly implemented by the offices of Biological and Environmental Research (BER) and Advanced Scientific Computing Research (ASCR), the program aims to develop the knowledge base and the national infrastructure for computational biology in addition to achieving the goals outlined in this document.

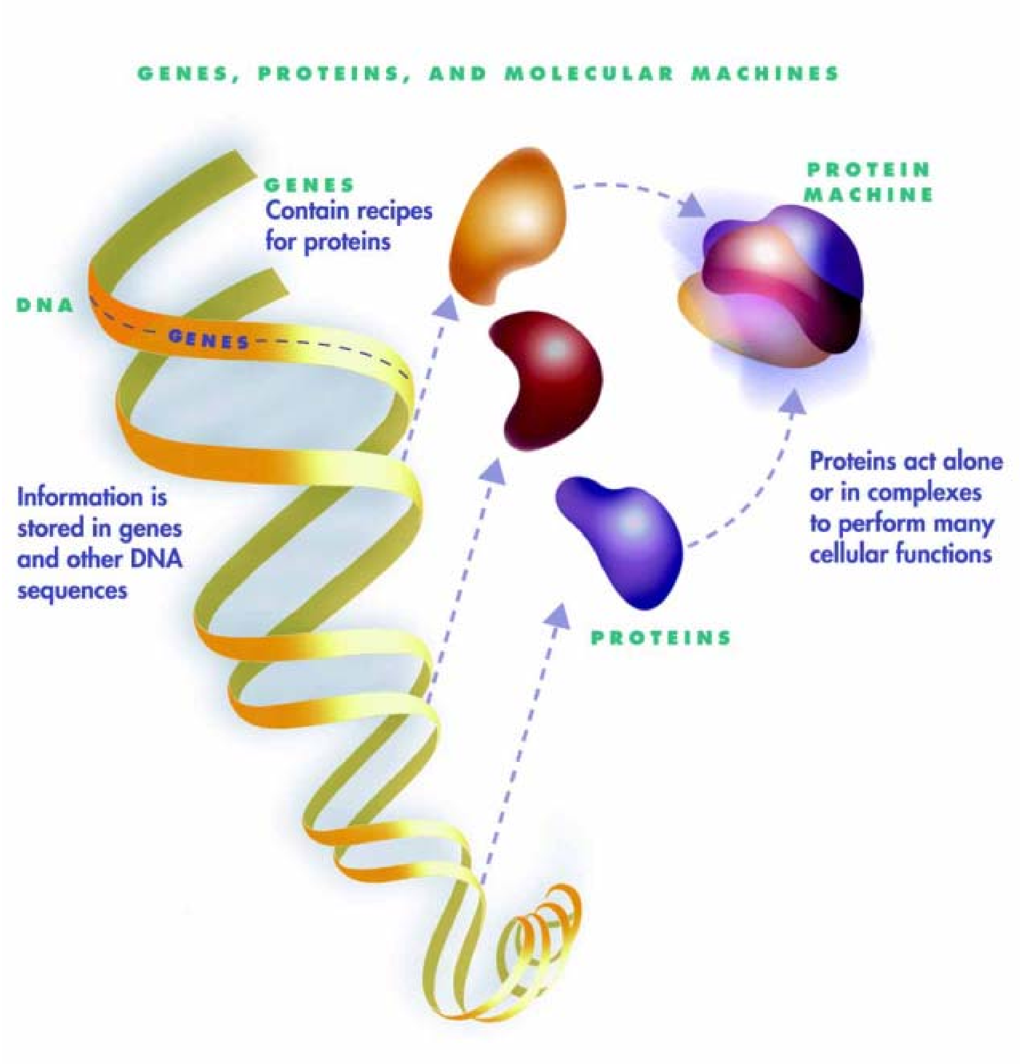
By developing a fundamental understanding of living systems, the Genomes to Life program responds directly to DOE’s missions. The results will help the department develop new sources of energy, mitigate the long-term impacts of climate change, clean up the environment, reduce the threat of biological terrorism, and protect people from adverse effects of exposure to environmental toxins and radiation (see sidebar, page 9).

Web site for program:
DOEGenomesToLife.org

Beyond the Sequences

Genomes are “brought to life” by being read out or “expressed” according to a complex set of directions embedded in the DNA sequence. The products of expression are proteins that do essentially all the work of the cell: they build cellular structures, digest nutrients, execute other metabolic functions, and mediate much of the information flow within a cell and among cellular communities. To accomplish these tasks, proteins typically work together with other proteins or nucleic acids as multicomponent “molecular machines”—structures that fit together and function in highly specific, lock-and-key ways (see figure below and a more comprehensive explanation and pictorial on pp. 14–17).

To understand how genomes are brought to life in both simple and complex organisms, biologists face two immediate challenges. First, they must characterize the full repertoire of molecular machines employed by living systems. And second, they must understand how the operations of these machines are orchestrated to give life to both single cells and complex multicellular organisms. Genomes to Life addresses these challenges with four goals:


GENES, PROTEINS, AND MOLECULAR MACHINES
GENES
Contain recipes for proteins
PROTEIN MACHINE
DNA
GENES
Information is stored in genes and other DNA sequences
Proteins act alone or in complexes to perform many cellular functions
PROTEINS

1. Identify and characterize the molecular machines of life—the multiprotein complexes that execute cellular functions and govern cell form.
2. Characterize gene regulatory networks.
3. Characterize the functional repertoire of complex microbial communities in their natural environments at the molecular level.
4. Develop the computational methods and capabilities to advance understanding of complex biological systems and predict their behavior.

Challenges of Complexity

While the benefits of this new understanding are apparent, the path forward is formidable. Biological systems, through evolution, have achieved levels of intricacy and subtlety that dwarf the complexity of the 20th century's most sophisticated engineering feats. Genomes contain thousands of genes, many of which make multiple proteins; many genes regulate other genes either directly or indirectly through dynamic and often-complex regulatory pathways.

The challenge presented by this complexity cannot be met using a traditional single-gene, single-protein approach. Instead, new methods must be built on the technical and conceptual foundation laid down by large-scale genome sequencing.

This natural complexity sets the stage for the first of two challenges: the need to collect enormous amounts of data about genomes, especially expressed genomes, and, ultimately, about the specific groups of biological molecules and protein machines expressed and assembled in different cell types and under varying conditions. The body of data produced by genome projects represents the first step in this data-gathering process, but—and this is a key element of Genomes to Life—the necessary additional data cannot be obtained efficiently with current technology. DOE is poised, as it was in the Human Genome Project, to make key contributions in technology development.

However, no amount of additional information will in itself yield the understanding sought. There remains a second, much deeper, complexity challenge—that of deriving underlying theoretical and mathematical principles for biology. Just as modern integrated circuits have become so complex that they cannot be designed or tested without the aid of extremely sophisticated computer simulation and modeling tools, so too is the case with biological systems. They are too intricate to study without advanced computational tools for managing and integrating the data into mechanistic models that describe how cells work. Herein lies a second key element of Genomes to Life—high-performance computing linked to biology via the detailed knowledge of protein structures and interactions. This linkage creates the ability to generate computational-experimental cycles that will provide the framework for systems biology.

Economies of Nature

With contributions from DOE and other agencies and organizations, there are signs of hope that the complexity challenges can be met. Although a model currently cannot be developed that describes in detail how even the simplest cell functions, various attributes of the cell can be modeled—metabolic processes, for example. In some cases, investigators have used models to successfully predict growth rates, metabolic byproducts, and consequences of DNA deletions. Furthermore, nature has provided two simplifying principles that encourage optimism for tackling the complexity challenges.

First, just as individual proteins use a finite number of rules to take on their final three-dimensional forms, so too it seems that life's molecular machines are finite in number, as proteins associate in precise ways to carry out crucial functions.

Second, once a successful protein machine arises by evolution, it tends to be preserved, subtly modified and optimized, and then reused as variations on an enduring theme across organisms and species. As a result, an inventory of the kinds of protein-containing complexes and the larger networks in which they are embedded—especially those involved in such basic cell functions as metabolism and structure—should prove to be small enough to permit practical study.

The Microbial Cell Project: A First Step

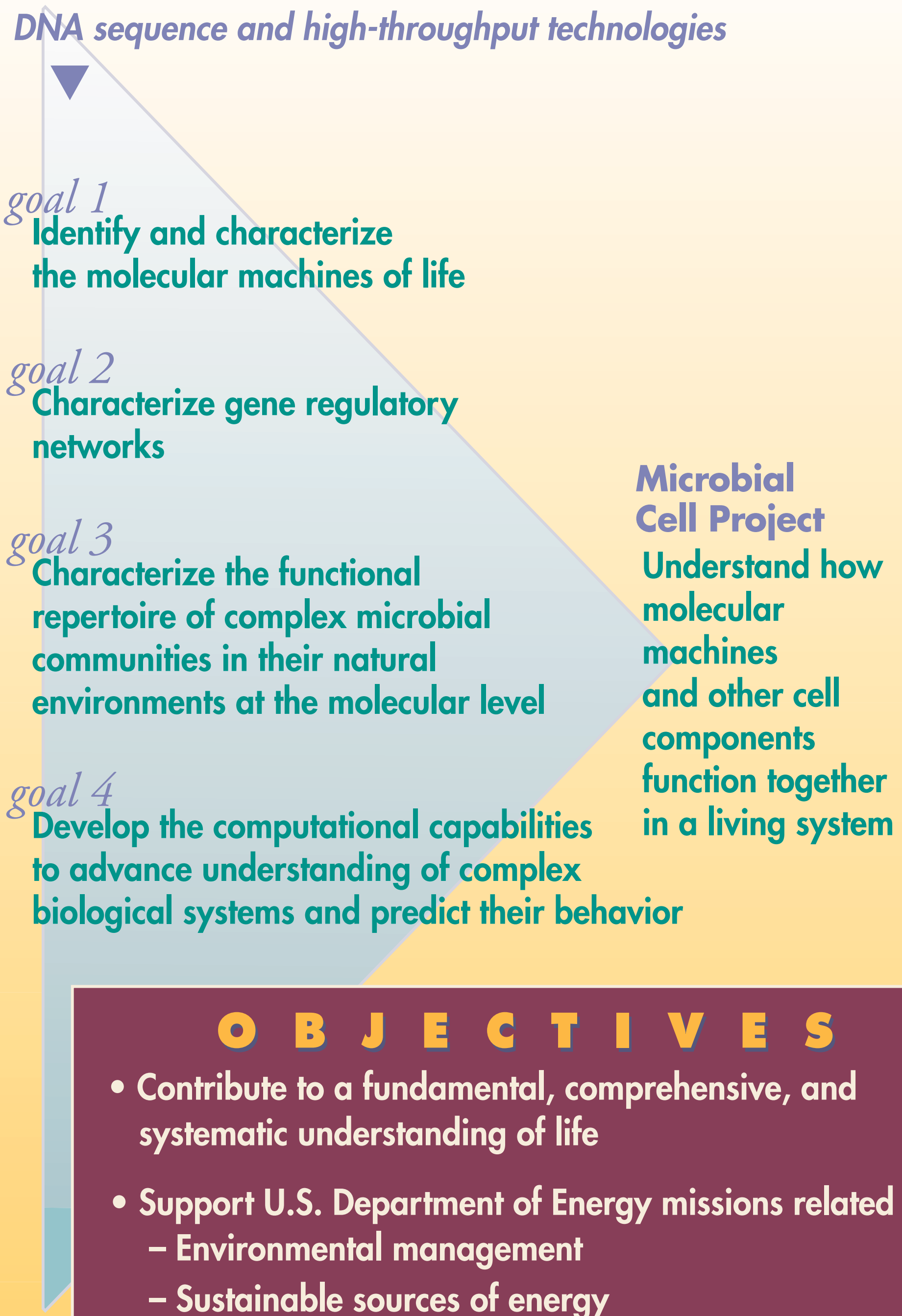
BER's Microbial Cell Project, initiated in FY 2001, provides both a key unifying component for Genomes to Life and an ultimate test of the genome-based understanding of living systems. The work itself is part of the federal Microbe Project, a multiagency effort that coordinates interdisciplinary teams to devise integrated approaches for characterizing the structure and function of a prokaryotic cell.

Genomes to Life and the Microbial Cell Project are designed to be complementary and to build on each other's successes. Genomes to Life emphasizes broad, genome-wide strategies to collect and analyze data across biological systems; the Microbial Cell Project focuses on collecting, analyzing, and integrating data about individual microbes. Eventually, the project will emphasize all of the molecular machines that work together to give life to a simple microbe.

Genomes to Life will develop maps of the complex regulatory networks that control these molecular machines in representative microbes and in higher, more complex organisms. The Microbial Cell Project will define the global interactions among proteins and other biomolecules that together form specific functional networks in microbes. This characterization will include information on the dynamic behavior

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

Program Goals



DNA sequence and high-throughput technologies

goal 1
Identify and characterize the molecular machines of life

goal 2
Characterize gene regulatory networks

goal 3
Characterize the functional repertoire of complex microbial communities in their natural environments at the molecular level

goal 4
Develop the computational capabilities to advance understanding of complex biological systems and predict their behavior

Microbial Cell Project
Understand how molecular machines and other cell components function together in a living system

OBJECTIVES

- Contribute to a fundamental, comprehensive, and systematic understanding of life
- Support U.S. Department of Energy missions related to
  - Environmental management
  - Sustainable sources of energy
  - Atmospheric and climate stability
  - Worker protection and human susceptibility

of the various molecules as the molecular machines perform their functions and on the distribution, localization, movement, and temporal variations of molecules and machines inside individual microbes.

The success of both programs will depend on a close coupling of computation with biological research. Genomes to Life will develop the broad computational and modeling infrastructure needed to simulate and predict the biology of individual microbes, microbial communities, and even humans. The Microbial Cell Project will focus on developing and using computational systems to simulate specific functional pathways and regulatory networks in individual microbes or, at a lower level of resolution, entire microbial communities.

Program Management

Understanding the fundamental processes of complex living systems—the grand challenge of the New Biology—is a task of major proportions that calls upon the capabilities and resources of the public and private sectors. The level of coordination and management required will be even greater than that of the collaborative Human Genome Project.

The success of the program thus rests heavily on a highly interactive and communicative environment. Program managers will meet with key stakeholders in a series of workshops, scientific society symposia, and other exchanges on scientific topics to guide program development. These activities will provide the interchange of ideas necessary to establish priorities for scientific directions, coordination, and investments. DOE will use established advisory mechanisms (e.g., the Biological and Environmental Research Advisory Committee and Advanced Scientific Computing Research Advisory Committee) and peer-review procedures to evaluate the program's strategic planning, its scientific progress, and its appropriate investment in future technologies, both at the program and investigator levels.

Genomes to Life research complements that of ongoing DOE programs in bioremediation, carbon sequestration, chemical and biological weapons nonproliferation, low-dose radiation research, and molecular imaging. Similarly, the major thrusts of the new program are distinct from and synergistic with the roles assumed by other federal agencies in understanding biological systems. Agencies involved in these programs include the National Institutes of Health, National Science Foundation, and Department of Agriculture. Interactions among all these programs offer exceptional opportunities for advances. For more details, see the list of programs and their Web sites on p. 70.

Biological Solutions for DOE Missions

The broad missions of DOE include producing energy, sequestering excess atmospheric carbon affecting global climate, cleaning up environments contaminated by weapons production, reducing the threat of chemical and biological warfare, and protecting people from radiation (an energy byproduct) and other environmental insults and stresses. Each of these missions has a basis, and possibly a solution, in the biological world, as described below.

Human Susceptibility

DOE has a need to protect its workers and the public from the effects of energy production and from low levels of weapons-related materials at DOE waste sites and those still in use at its laboratories. Because of their genetic makeup, some individuals may have a much greater health risk from exposures to these materials. A detailed understanding of how basic metabolic and regulatory pathways respond to the environment may offer new insights to help clarify the biological mechanisms responsible for adverse human responses and to provide tools that could be used to identify individuals at risk.

Chemical and Biological National Security

The DOE mission in chemical and biological national security is to develop, demonstrate, and deliver technologies and systems to improve the nation's ability to prepare for and respond to chemical or biological attacks. Genomes to Life research can support this effort in detection, therapeutics, and forensics. For example, improved knowledge about protein-protein interactions and molecular machines in microbes could lead to the development of sensors that detect chemical and biological agents, improved vaccines and treatment options, and strategies to enable strain identification.

Carbon Cycle and Sequestration

A strategy that could be used to counter greenhouse-gas buildup (an influence on global climate) is to alter natural biological cycles to store extra carbon in the terrestrial biomass, soils, and the biomass that sinks to ocean depths. This approach will be tied to the metabolism and activities of communities of microbes. Research into their enzymes, regulation, and environments will lead to new ways to store and monitor carbon.

Bioremediation

For more than 50 years, the United States has been creating a vast network of facilities for research, development, and testing of nuclear materials. As a result, subsurface contamination by radionuclides or metals has been documented at over 7000 discrete sites across the DOE complex, and physical treatments often are difficult and prohibitively expensive. Genomes to Life research is expected to provide knowledge about using natural populations of microbes to degrade or immobilize contaminants and accelerate the development of new, less costly strategies for cleaning up DOE waste sites.

Renewable and Alternative Energy Sources

A longstanding mission for DOE is to develop renewable energy from vegetation and alternative energy sources such as hydrogen. Renewable energy from plants requires the design of plants with biomass that can be transformed efficiently to fuels; however, a limiting factor noted in developing such plants is the lack of understanding about their metabolic pathways. Knowledge of biochemical pathways may lead to more efficient strategies for converting biomass to fuels. Similarly, an ability to harness these pathways in hydrogen-producing microbes could one day provide an alternate energy source.

Biological explorations taking place at the whole-systems level require scientists trained in new interdisciplinary areas: life scientists who use bioinformatics, modeling, technologies, and techniques from the physical sciences; and physical and computer scientists and engineers with an understanding of the life sciences. The Genomes to Life program thus will emphasize highly integrated approaches by interdisciplinary teams of scientists. It will draw on large-scale and multidisciplinary capabilities of the national laboratories as well as those of the academic and commercial communities.

Lessons learned from the Human Genome Project will guide the ongoing management and coordination of the Genomes to Life program. They include the importance of high-throughput approaches for collecting and analyzing biological data and the critical need for computational tools to manage, integrate, and interpret the resulting information.

Ethical, Legal, and Social Issues

The resources and profound insights expected from this new program will raise ethical, legal, and social issues (called ELSI). Advances will provide scientists with powerful tools to better predict and ultimately manipulate the biology of cells, tissues, organs, and eventually whole organisms. This, in turn, will confer abilities to alter microbial cell sensitivity to environmental signals and to use microbes to change their own local environments. One result may be new powers to design living systems that promote beneficial environmental processes in waste cleanup, carbon sequestration, energy production, and biotechnology, to name a few. Unless thoroughly understood and wisely used, however, activities such as these also have the potential to harm the environment.

The new knowledge also may enable prediction of individual human responses to environmental exposures, information that may be used by some to discriminate against others in employment or health insurance. Intellectual property issues may arise as well over the applications and commercialization of resources and data.

The Genomes to Life program will offer relevant scientific insights that can inform these dialogues. To maximize the benefits of these advances while anticipating and minimizing risks, collaborations and other cross-disciplinary interactions between scientists and nonscientists will be encouraged. As with the Human Genome Project, every effort will be made to understand the social implications of scientific progress and to promote education and effective policy development.

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

Technology Needs

The Human Genome Project taught that evolutionary improvement in existing technologies (e.g., DNA sequencing) can have a revolutionary impact on science. The systems approach taken by the Genomes to Life program dictates that existing technologies (some of which are described in Appendix A) must evolve to a high-throughput capability. In addition, revolutionary technologies need to be developed, incorporating new modes of robotics and automation as well as advanced information and computing technologies. The following is a list of some key high-throughput technologies.

DNA, RNA, Protein, Protein Machine, and Functional Analyses and Imaging

- High-throughput identification of the components of protein complexes; mass spectrometry, new chip-based analyses, and capture assays
- Parallel, comparative, high-throughput identification of DNA fragments among microbial communities and for community characterization
- Whole-cell imaging; novel imaging technologies, including magnetic resonance optical, confocal, soft X-ray, and electron microscopy; and new approaches for in vivo mapping of spatial proximity
- New technologies for mapping contact surfaces between proteins involved in complexes or molecular machines (e.g., FRET and neutron scattering)
- Functional assays; development of novel technologies and approaches for defining the functions of genes from uncultured microorganisms

Sampling and Sample Production

- Approaches for recovering RNA and high-molecular-weight DNA from environmental samples and for isolating single cells of uncultured microorganisms
- Advances in separation techniques, including new techniques to capture targeted proteins, and high-affinity ligands for all gene products
- Improved approaches for studying proteins that are hard to crystallize (e.g., membrane proteins)

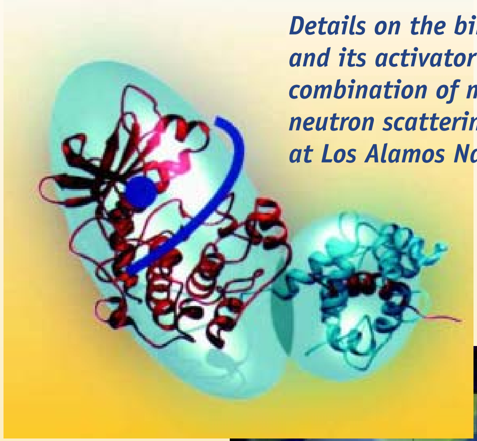
Informatics, Modeling, and Simulation

- Algorithms for genome assembly and annotation and for bioinformatics to measure protein expression and interactions
- Standardized formats, databases, and visualization methods for complex biological data sets, including expression profiles and protein-protein interaction data
- Molecular modeling methods for long-timescale, low-energy macromolecular interactions and for prediction of chemical reaction paths in enzyme active sites
- Methods for automated collection and integration of biological data for cell-level metabolic network analysis or pathway modeling; improved methods for simulation, analysis, and visualization of complex biological pathways; and methods for prediction of emergent functional capabilities of microbial communities

DOE Strengths and Capabilities

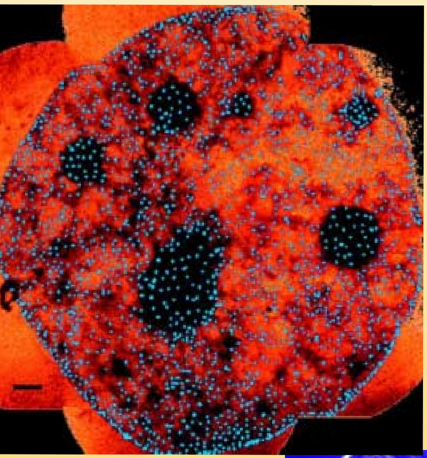
High-throughput methods, advanced computational and imaging resources, and multidisciplinary collaborations are essential elements for using the information contained in DNA sequences as a foundation for expanding knowledge of how living systems function—the focus of the Genomes to Life program. DOE research capabilities that will contribute to the success of this program include those outlined below. Descriptions of some technologies pictured here and a listing of BER- and ASCR-supported facilities appear in the appendices, starting on p. 53.

- High-throughput DNA sequencing at the Joint Genome Institute
- High-performance computing infrastructure and resources based in the Office of Advanced Scientific Computing Research
- Facilities and resources such as DOE synchrotron and neutron sources, Environmental Molecular Sciences Laboratory, mass spectrometers, nuclear magnetic resonance spectrometers, high-resolution electron and soft X-ray microscopes, and the Mouse Genetics Research Facility
- Tools developed for medical imaging programs to localize and visualize molecular machines at work in cells
- Knowledge, capabilities, and resources in the Biological and Environmental Research Program's Microbial and Human Genome Programs and in structural biology, proteomics and model organism research
- Tools and resources in the Nanotechnology Initiative of the Office of Basic Energy Sciences



Details on the binding and dynamics of CAM kinase II and its activator calmodulin were revealed using a combination of mutagenesis, crystallography, NMR, neutron scattering, and computational technologies at Los Alamos National Laboratory.

Production Sequencing Facility at DOE's Joint Genome Institute



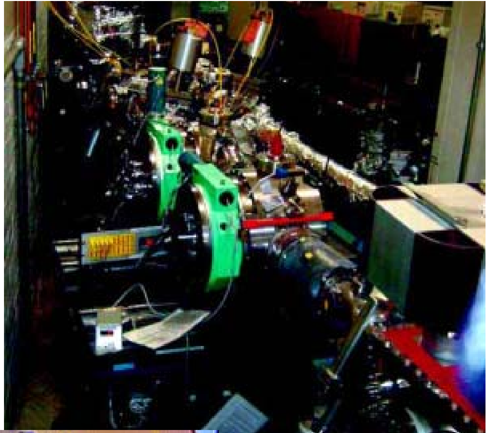
This image of a human mammary cell was produced using soft X-ray microscopy at Lawrence Berkeley National Laboratory. The blue dots label proteins of the nuclear pore complex, through which molecules enter and exit the nucleus.

IBM SP supercomputer at Oak Ridge National Laboratory

Next-generation DNA sequencing technology from University of California, Berkeley

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

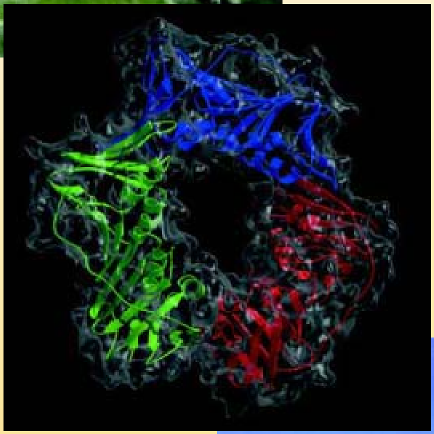
Site plan of the Spallation Neutron Source being built at Oak Ridge National Laboratory in collaboration with Argonne National Laboratory, Brookhaven National Laboratory, Lawrence Berkeley National Laboratory, and Los Alamos National Laboratory



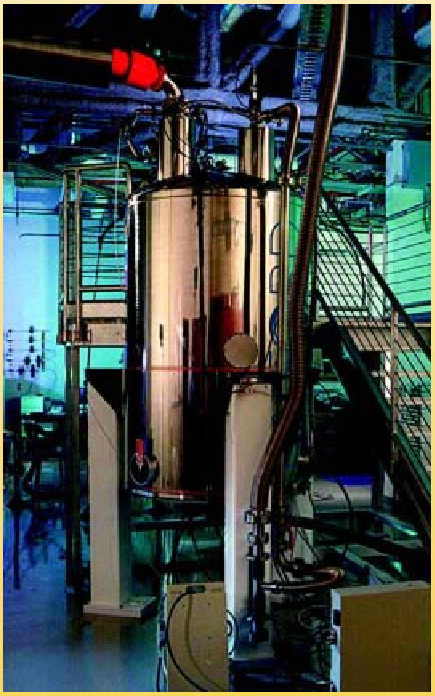
A beamline at the National Synchrotron Light Source at Brookhaven National Laboratory

Mass spectrometer in the Environmental Molecular Sciences Laboratory at Pacific Northwest National Laboratory

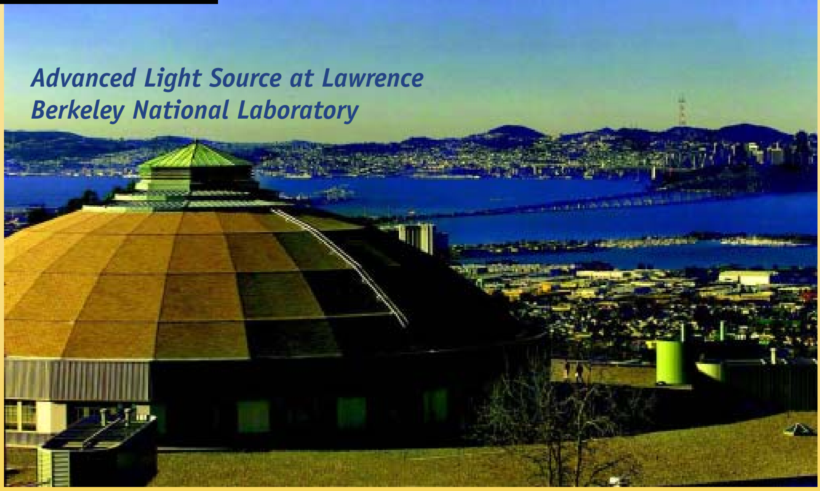
Advanced Photon Source at Argonne National Laboratory



The role of the Rad checkpoint complex was inferred from the 3-D structure predicted by comparative modeling at Lawrence Livermore National Laboratory. The Rad complex delays cell division to allow time for DNA repair to take place.



Environmental Molecular Sciences Laboratory's 800-MHz nuclear magnetic resonance spectrometer at Pacific Northwest National Laboratory

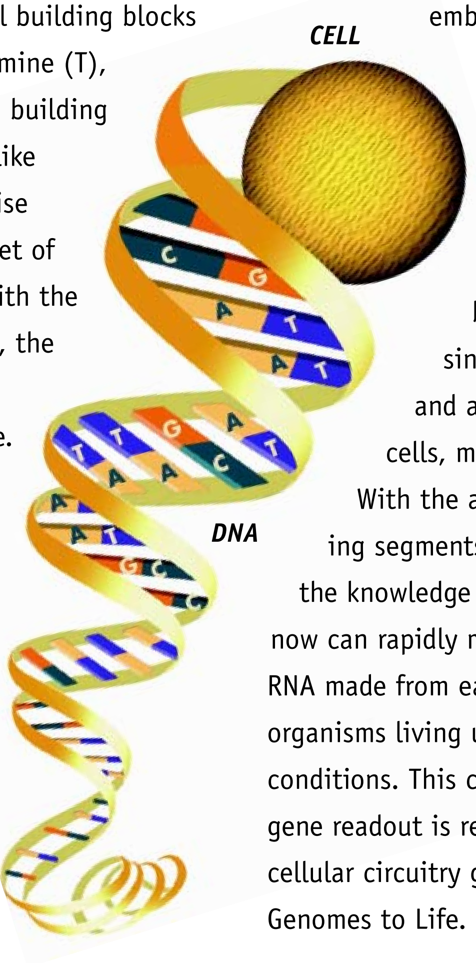


Advanced Light Source at Lawrence Berkeley National Laboratory

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

Genomes to Life: A Primer

All organisms face three information challenges, and all life on earth, from invisible microbes to the largest plants and most exotic animals, uses the same fundamental biochemical strategies to meet these challenges. First, the organism must encode and store, within each cell, all the instructions needed to build, operate, maintain, and reproduce itself and to respond to varied environmental conditions. DNA, the biochemical solution to this coding and storage problem, is made up of four chemical building blocks (nucleotide bases): adenine (A), thymine (T), cytosine (C), and guanine (G). These building blocks are organized in long chains like chemically linked beads, whose precise order spells out the organism's full set of genetic instructions—its genome. With the advent of whole-genome sequencing, the assembly and study of the entire instruction set have become possible.

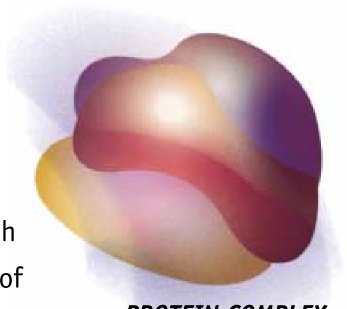


But the information stored in DNA is "lifeless" by itself, just as a recipe in a book is not a delectable dessert, nor a musical score a majestic symphonic performance. In the same way, the DNA sequence must be "expressed" to give life to a cell or organism. Furthermore, the sequence alone does not automatically provide understanding of how each segment contributes to the whole cell or organism. The overarching aim for Genomes to Life is to understand how the information in DNA spells out a living cell or organism.

The second information challenge is to read out the genome's instructions in the proper order, time, and amount for each gene product. The biochemical answer begins with the selective readout (transcription) of each functional segment of DNA sequence (gene) in the form of RNA, which is a close chemical relative of DNA. The set of RNA transcripts generated for a cell is called its transcriptome. RNA, in turn, is the direct molecular instruction for a specific protein's synthesis, accomplished by the cell in a process known as translation. Selective gene readout in the chemical form of RNA, therefore, can govern the identity and quantity of proteins, which are the cell's workhorse molecules and the ultimate physical embodiment of the information encoded in the DNA. The constellation of proteins in a cell is called its proteome.

Cells are the fundamental working units of living systems. The range of life's complexity varies from invisible bacteria that carry out all functions as single-celled organisms to complex plants and animals containing millions or trillions of cells, many with highly specialized functions. With the availability of gene microarrays containing segments of many different genes coupled with the knowledge of entire genome sequences, scientists now can rapidly monitor the identities and amounts of RNA made from each of thousands of genes in cells and organisms living under hundreds or thousands of varied conditions. This capability may provide insight into how gene readout is regulated. Mapping and modeling the cellular circuitry governing this process is a major goal for Genomes to Life.

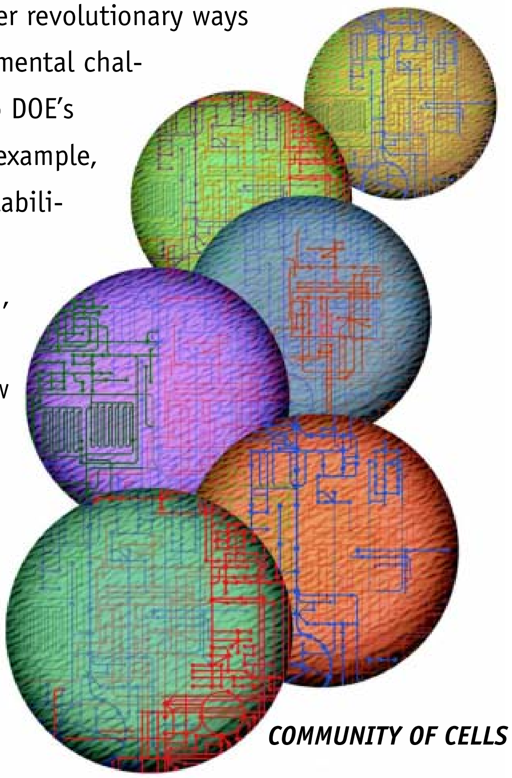
Like DNA and RNA, proteins are synthesized like "beads on a string" but with 20 different kinds of beads (amino acids) rather than the 4 of RNA or DNA. Chemical properties that distinguish different amino acids ultimately cause the protein chains to fold up into specific three-dimensional structures. It is the proteins that meet the third and greatest information challenge—which is to act out the instructions encoded in DNA. Although DNA and RNA are information rich, they are chemically simple and homogeneous. Proteins, by contrast, are chemically complex and diverse, properties that enable them to do so many different jobs. Proteins are “where the action is” in living systems. They are motors, pumps, chemical catalysts, detectors, signals and signalers, conveyers, structural units, gateway keepers, dismantlers, assemblers, and garbage handlers. They regulate cell replication, survival, and even death. Recent progress in whole-genome DNA sequencing and in areas of protein-structure determination have brought investigators to the point of knowing the composition of most proteins from model organisms, but the challenge is to know how proteins give cells their capabilities, structure, and higher-order properties.



PROTEIN COMPLEX

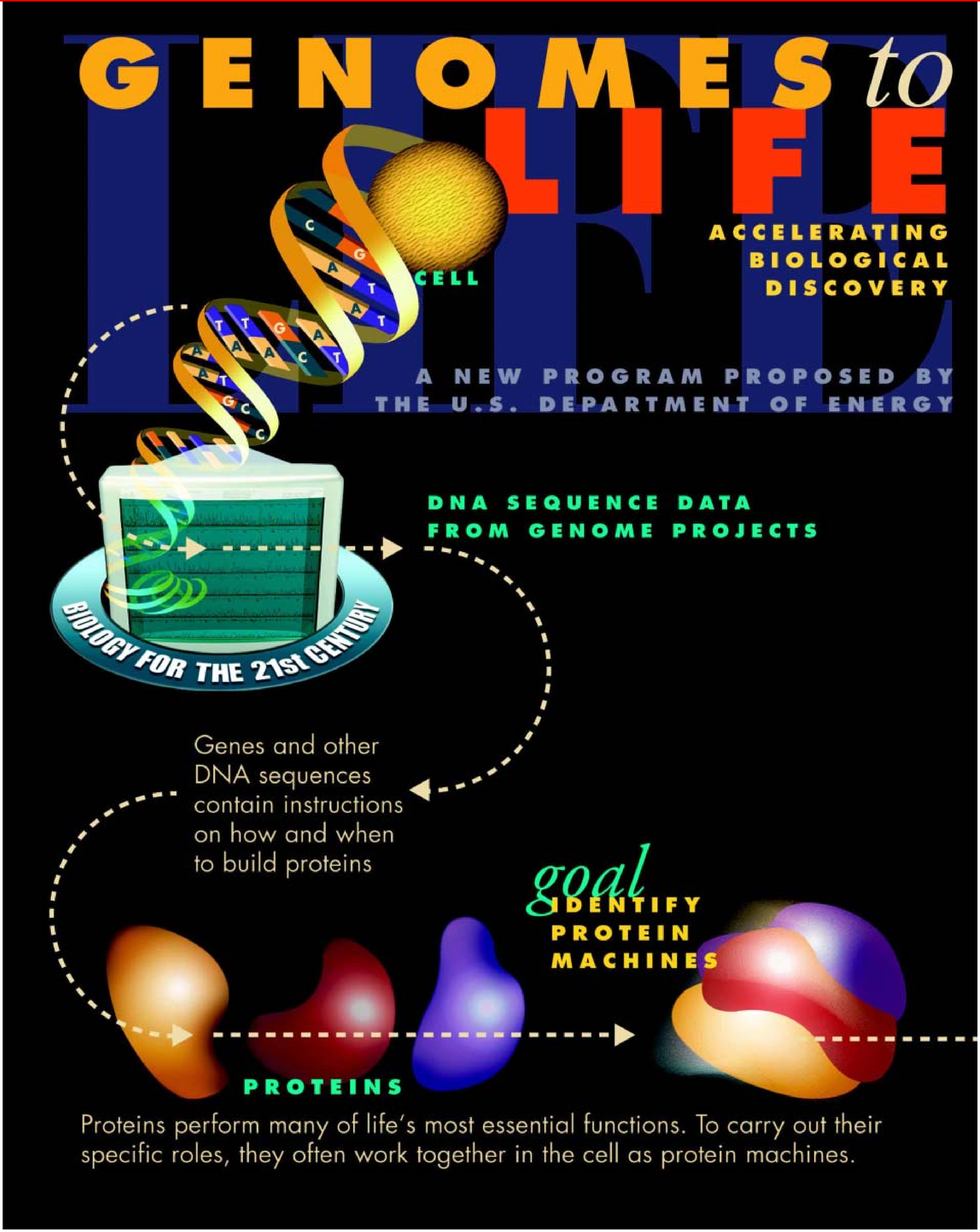
Proteins rarely solo. More often, they work by assembling into larger multiprotein complexes, some of which have the characteristics of rather complicated protein “machines.” These machines, in turn, execute such major functions as protein synthesis and degradation, cell-to-cell signaling, and a host of other operations. The properties of each kind of protein, which cause it to assemble with others into machines and to execute very specific and critical reactions in the cell, are the direct consequence of the protein’s amino acid sequence that dictates its final folded structure. That is, a protein’s chemistry and behavior are specified by the gene sequence and by the number and identities of other proteins made in the same cell at the same time and with which they associate and react. A major focus for Genomes to Life—and its first goal—is to learn the repertoire of protein complexes and machines needed to make different kinds of microbes and cell types function. These machines shift and change in composition, making their dynamics a further focus.

Cells do not solo very often, either. Although microbes are single-cell organisms, they typically live in communities composed of more than one kind of microbe—often many different kinds. Genomes to Life seeks to understand the properties of these cell communities by first learning about the “community” genome and relating it to the community’s capabilities to perform processes vital to DOE mission goals. Considering that life is found in virtually every environmental niche from arctic tundra to parched deserts to boiling sea vents on the deepest ocean floor, the global genetic “catalog” encoding all of life’s amazingly diverse capabilities must be astonishing, yet very few details are known. The recently discovered *Prochlorococcus* bacteria, for example, are now thought to be among earth’s major photosynthetic organisms, using carbon to produce life-sustaining oxygen. Scientists believe that harnessing the capabilities of these and other bacteria may offer revolutionary ways to solve environmental challenges related to DOE’s missions in, for example, global climate stabilization through carbon reduction, toxic-waste cleanup, and new and efficient energy sources. (See depiction of Genomes to Life program on the next page.)



COMMUNITY OF CELLS

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."



GENOMES to LIFE

ACCELERATING BIOLOGICAL DISCOVERY

A NEW PROGRAM PROPOSED BY THE U.S. DEPARTMENT OF ENERGY

DNA SEQUENCE DATA FROM GENOME PROJECTS

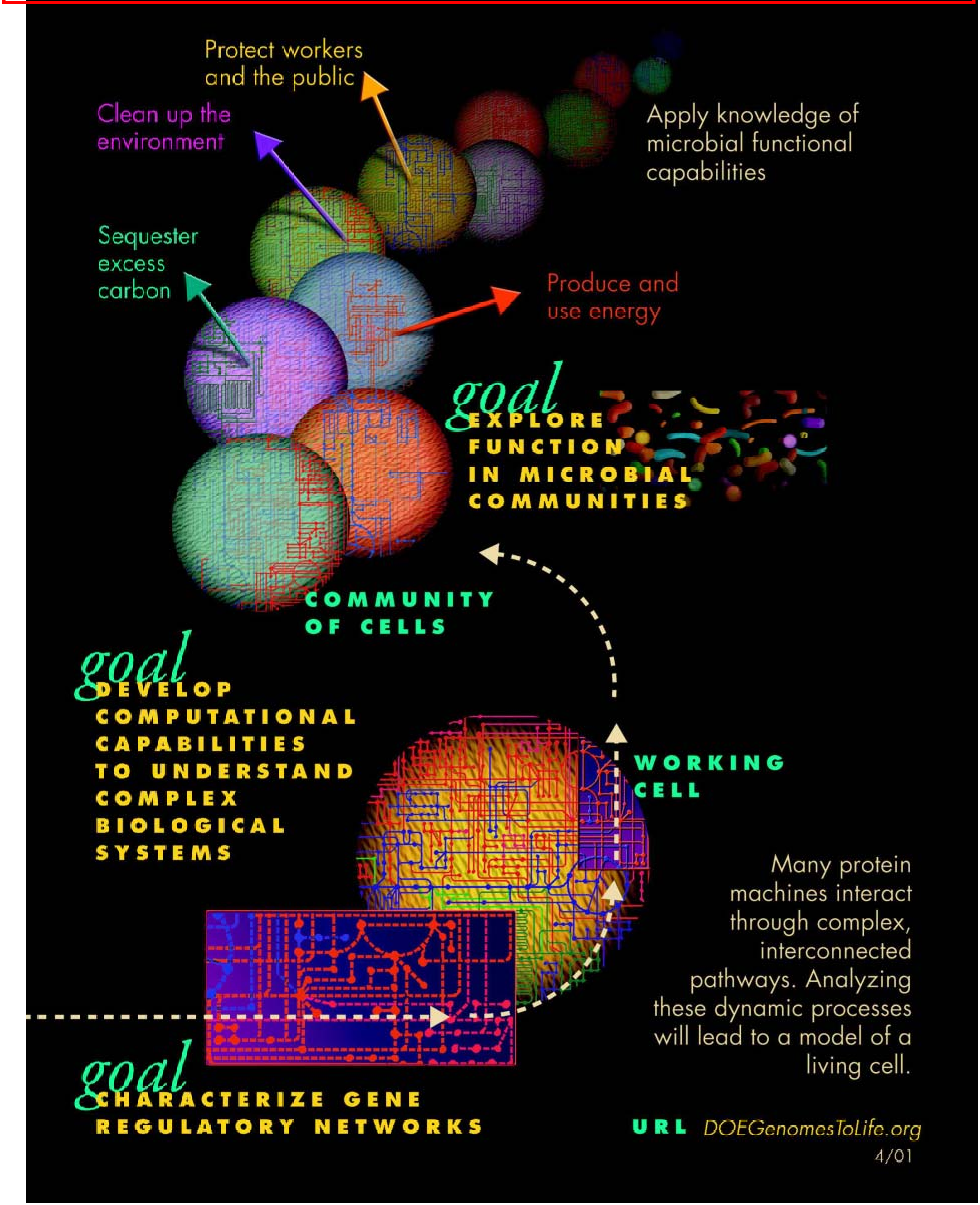
Genes and other DNA sequences contain instructions on how and when to build proteins

goal IDENTIFY PROTEIN MACHINES

PROTEINS

Proteins perform many of life's most essential functions. To carry out their specific roles, they often work together in the cell as protein machines.

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."



Protect workers and the public

Clean up the environment

Apply knowledge of microbial functional capabilities

Sequester excess carbon

Produce and use energy

goal EXPLORE FUNCTION IN MICROBIAL COMMUNITIES

COMMUNITY OF CELLS

goal DEVELOP COMPUTATIONAL CAPABILITIES TO UNDERSTAND COMPLEX BIOLOGICAL SYSTEMS

WORKING CELL

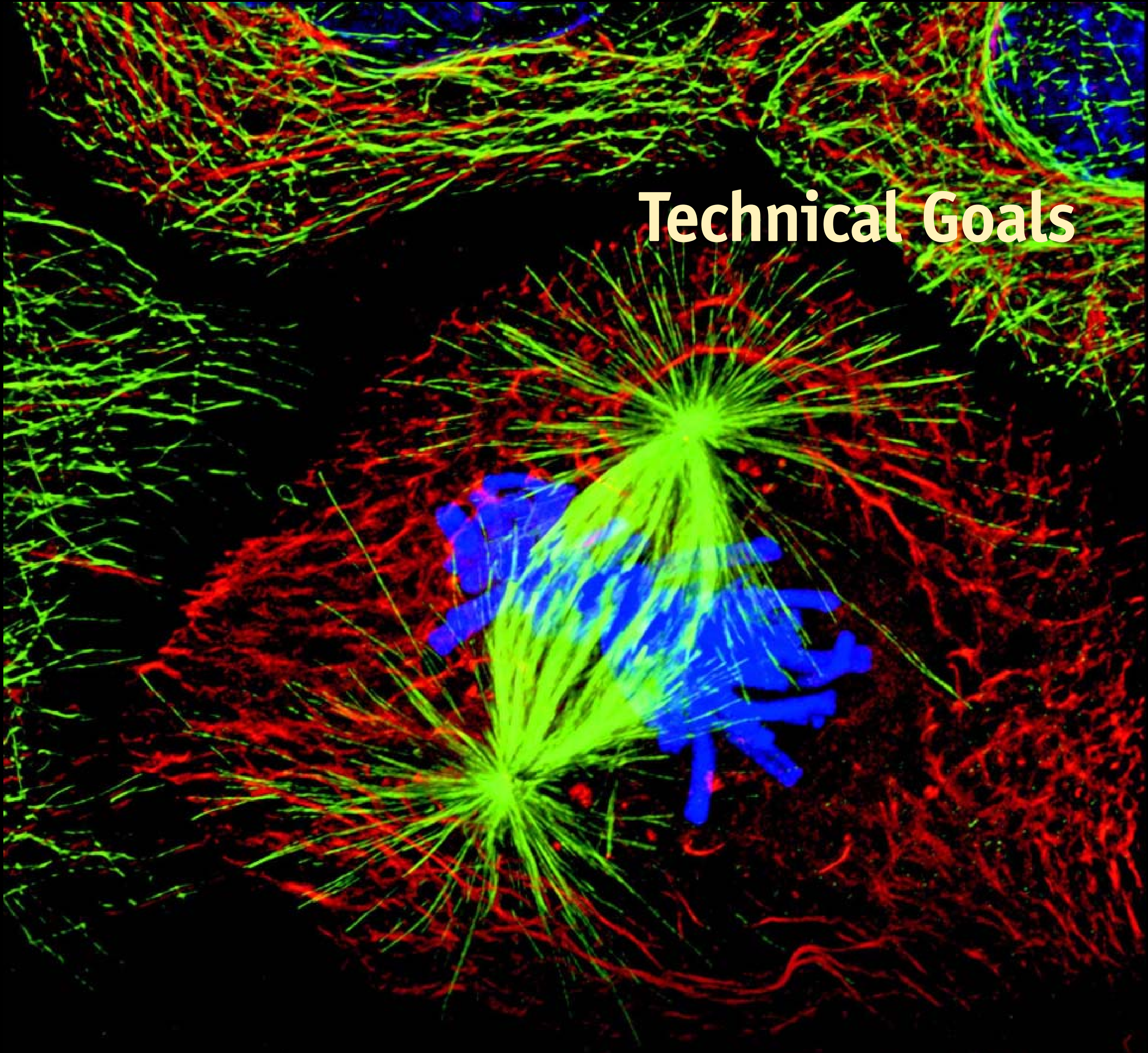
Many protein machines interact through complex, interconnected pathways. Analyzing these dynamic processes will lead to a model of a living cell.

goal CHARACTERIZE GENE REGULATORY NETWORKS

URL *DOEGenomesToLife.org*

4/01

Technical Goals



Cells are the basic working units of all living systems. They are biological "factories," performing and integrating thousands of discrete and highly specialized processes, often through the use of molecular "machines" composed of assemblies of different proteins and other molecules.

In the image above (obtained using a fluorescent-light microscope), a dividing cell is seen separating its duplicated chromosomes (blue) for equal distribution into two new cells. Spindle poles serve as centers from which microtubules (green) grow outward. The growing ends of microtubules interact with special structures on the chromosomes and generate a force to move them.

Enabling this activity is a cadre of tiny protein motors mobilized by the cell to move components into their proper positions, efficiently laying down the structure for each new cell. Dynein, a protein machine involved in cell division, is pictured and described further in the sidebar on p. 23.

Similarities in the composition and function of molecular machines have been observed across species. In Genomes to Life, identifying protein machines in a range of organisms and linking them to specific cellular functions of interest is a first step toward understanding the essential processes of living organisms. Achieving these goals will require the scrutiny afforded by new mathematical and computational tools and concepts.

[Cell image from Conly Rieder and Alexey Khodjakov, Wadsworth Center, Albany, New York (Rieder@Wadsworth.org)]

Goal 1

Identify and Characterize the Molecular Machines of Life—the Multiprotein Complexes That Execute Cellular Functions and Govern Cell Form

Background and Strategy

Proteins are the chemically active products of virtually all genes. Highly dynamic and shifting in amount, modification state, higher-order association, and subcellular localization, proteins carry out the primary functions of a cell in response to intracellular and extracellular signals. Most proteins do not act alone but instead are organized into multiprotein complexes that carry out activities needed for metabolic activity, communication, growth, and structure. The first goal of Genomes to Life is to systematically identify, characterize, and begin to understand these "machines of life." This will provide the essential knowledge base and set the stage for linking proteome dynamics and architecture to cellular and organismic function.

Identifying and characterizing multiprotein complexes on a genome-wide scale will require new tools and research strategies designed to increase throughput, reliability, and sensitivity. The dynamic nature of the proteome calls for methods to monitor, for any organism, the entire proteome's conditional state accurately and sensitively from thousands of samples. This will require greater completeness, resolution, and sensitivity than has been possible in the past using conventional imaging and gel-based technologies. Also, new tools characterizing these complexes must be developed to bridge the current size and resolution gap between single proteins suitable for high-resolution X-ray crystallographic study and the very large protein assemblies and cellular ultrastructures amenable to electron microscopy.

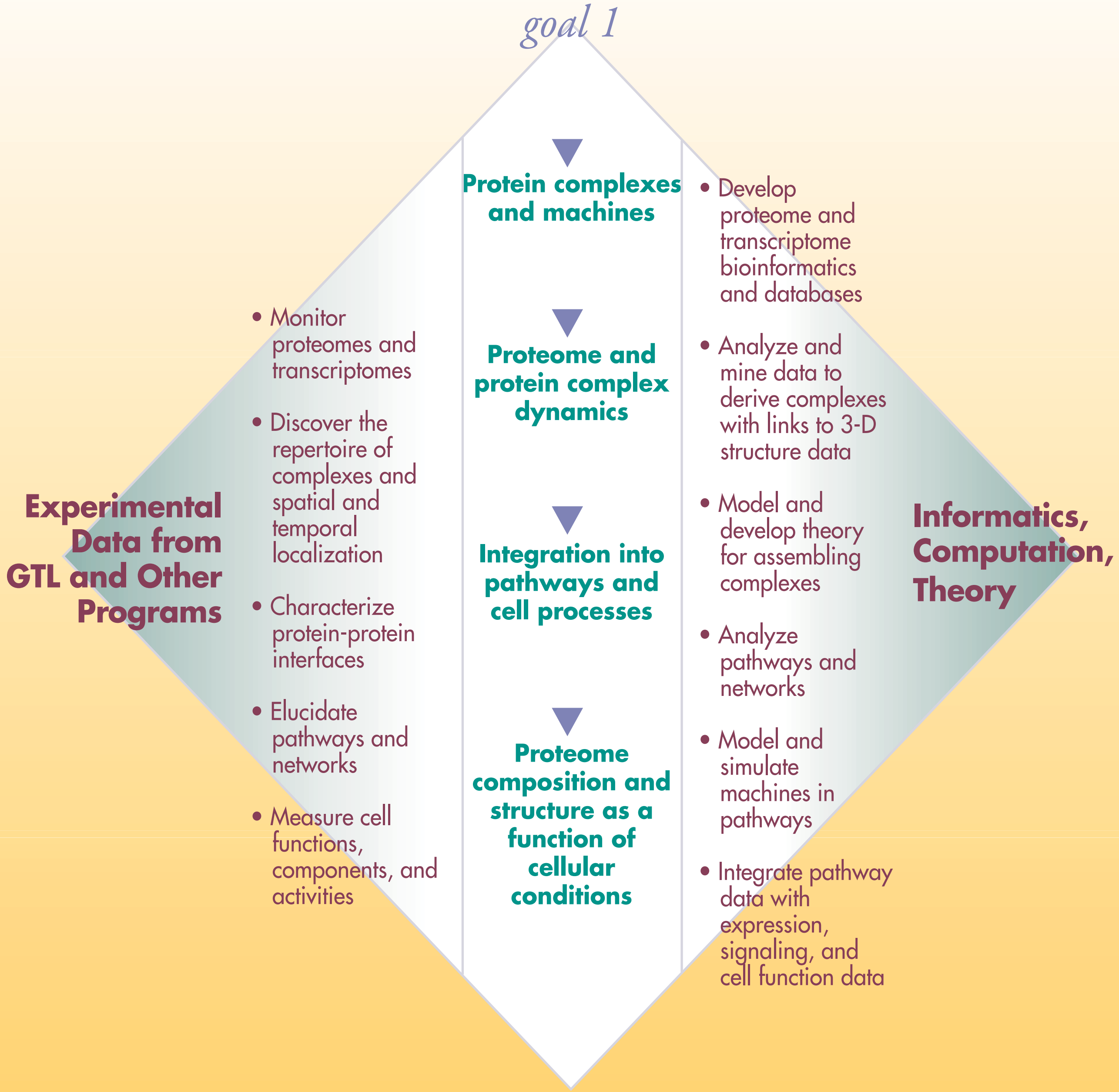
The importance of post-transcriptional and post-translational regulation is recognized by placing emphasis on direct knowledge of proteins and their higher-order associations in contrast to less direct inferences about proteome composition drawn from RNA measurements.

Providing a comprehensive view of proteome organization and dynamics promises to be a singularly important watershed of whole-genome biology for the coming decade because it will enable, inform, and improve virtually all other molecular and cellular investigations. Assembling a comprehensive view of the proteome and its multiprotein machines is also the best and most efficient way to address DOE-specific challenges in environmental and human biology. Having this picture would effectively leverage what is known about

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

GENOMES to LIFE

IDENTIFY AND CHARACTERIZE THE MOLECULAR MACHINES OF LIFE



goal 1

Protein complexes and machines

Proteome and protein complex dynamics

Integration into pathways and cell processes

Proteome composition and structure as a function of cellular conditions

Experimental Data from GTL and Other Programs

- Monitor proteomes and transcriptomes
- Discover the repertoire of complexes and spatial and temporal localization
- Characterize protein-protein interfaces
- Elucidate pathways and networks
- Measure cell functions, components, and activities

Informatics, Computation, Theory

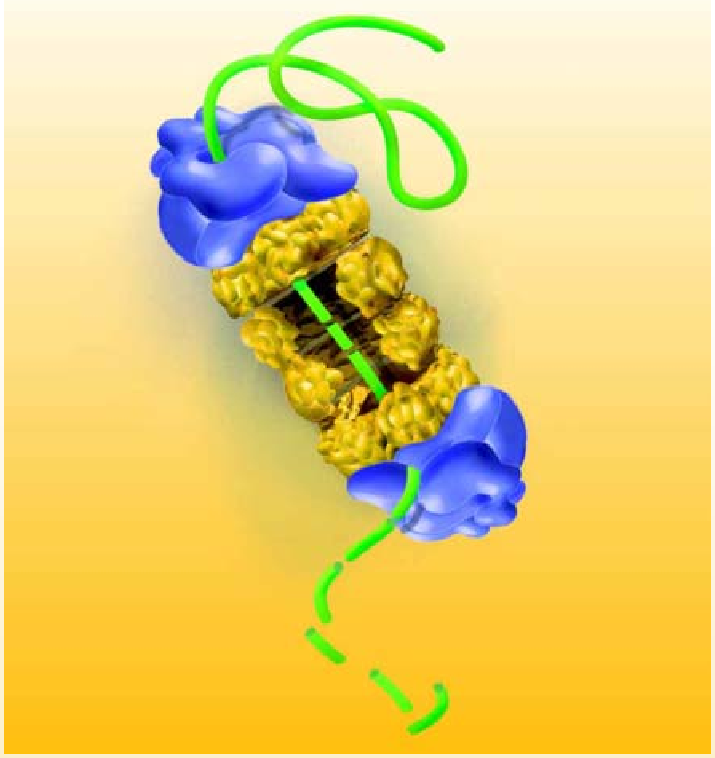
- Develop proteome and transcriptome bioinformatics and databases
- Analyze and mine data to derive complexes with links to 3-D structure data
- Model and develop theory for assembling complexes
- Analyze pathways and networks
- Model and simulate machines in pathways
- Integrate pathway data with expression, signaling, and cell function data

The Machines of Life

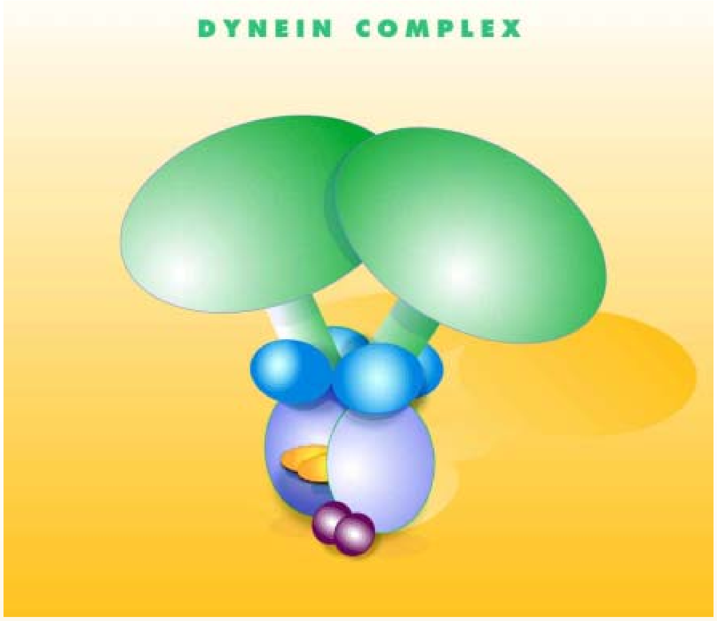
Proteins carry out almost all of life's essential processes, often working in highly complex multicomponent assemblies that sometimes include other types of macromolecules such as DNA or RNA. These machines typically work together in functional networks called "pathways" that underlie the dynamic life of a cell as they execute important metabolic functions, mediate information flow within and among cells, and build cellular structures.

Breaking down unneeded proteins—a task equal in importance to synthesizing new proteins—is accomplished by the orderly action of several multiprotein complexes. At the heart of this process is a multiprotein complex called the proteasome. This is a fundamental kind of machine that has been highly conserved during evolution. Some form of it is found in organisms ranging from simple bacteria to humans. These machines of destruction consist of a tunnel-like core with a cap at either or both ends (see figure at right). The core is formed by four stacked rings surrounding a central channel that acts as a degradation chamber. The caps recognize and bind to proteins targeted by the cell for destruction, then use chemical energy to unfold the proteins and inject them into the central core, where they are broken into pieces. Specific proteins are targeted to enter the proteasome by the action of yet another class of machine called E3 ligases. One type of E3 ligase machine is called the SCF complex. It consists of at least five kinds of proteins, and its job is to target and feed specific substrate proteins to the proteasome. In organisms ranging from yeast to human, this class of machines is responsible for chemically marking specific proteins for destruction by attaching to them yet another protein called ubiquitin that functions as a destruction tag. By using this multimachine pathway, cells can regulate and execute the highly specific elimination of a few kinds of proteins, while leaving the others intact. Such specific regulation of protein degradation gives cells a way to regulate major dynamic changes and cellular "decisions." A magnificent example is the central role of the proteasome pathway in causing a cell to proceed with the decision to replicate itself. In yeast cells a critical trigger for cell replication is degradation of Sic1, which is a protein that inhibits the chemical activity of CDC28. CDC28 is a cyclin-dependent kinase protein, and proteins like it are key regulators of the cell-division cycle in organisms ranging from yeast to human. After eliminating the biochemical Sic1 "brake" due to the action of SCF and the proteasome, the kinase is then free to trigger progress toward DNA replication and associated events of cell replication.

PROTEASOME



A very different kind of multiprotein complex forms in cells to transduce one important group of molecular signals from outside the cell. The immediate effect of this



signal transduction is to change gene expression in the cell receiving the signal, which in turn causes the cell to behave differently after receiving the signal. For example, in yeast the transduction of one such signal causes the cell to cease dividing and begin the process of mating. At the heart of this signal transduction pathway are three different proteins of the MAPkinase family that act upon each other in series in response to the signal. Recent discoveries show that they do not do this by “swimming” about the cell until they bump into each other entirely by chance, as was once thought. Instead the three kinases come together into an orderly structure by collectively binding to a protein “scaffold.” In the pathway that mediates mating in yeast, this scaffold is a protein called STE5. The pathway cannot function properly without forming the STE5-mediated complex. Many different variations on the three-member MAPK cascade theme are used in all higher organisms to transduce a diverse set of specific signals. It is therefore likely that there is a correspondingly large family of multiprotein complexes of this basic machine type, and scientists would be greatly aided by knowing the composition of all such machines. A more-general implication is that other pathways featuring a series of sequential reactions may also exist in the cell as ordered multiprotein complexes, and this in turn could powerfully affect the way the pathways function, whether and if they engage in chemical “crosstalk,” and how they are regulated. This is another critical reason for learning the composition of these machines.

An entirely different class of molecular machines functions as motors, converting chemical energy into mechanical motion, both linear and rotary. The dynein motor (see figure at left), a cellular complex believed to be composed of 12 distinct protein parts, performs fundamental transportation tasks critical to the cell; defects in its structure can prove fatal. This machine converts chemical energy stored in an ATP molecule into mechanical energy that moves material though the cell along slender filaments called microtubules. One of the dynein motor’s most important functions occurs during cell division, when it helps move chromosomes into proper position, as seen in the image on p. 19.

Multiprotein machines, such as the ones described above, tend to be quite highly conserved in overall composition and function during evolution albeit often as variant forms on a basic theme. As a result, one can expect that much of what is discovered in one organism will be applicable to others. Once a class of multiprotein machines is defined, variations on the theme can be plumbed to illuminate critical details for pathways of high relevance to DOE. No one yet knows the number of different basic types of such machines, but extrapolation from discoveries of the past few years, coupled with new knowledge of total protein diversity that comes from having whole-genome DNA sequences, suggests that the number is in the range of several thousand types. This suggests that the goal of characterizing the basic repertoire of machine types in the Genomes to Life program is an audacious but tractable undertaking.

whole-genome sequences and individual protein structures and functions. The new knowledge platform produced will aid scientists working on diverse problems by providing insights into major questions such as how and why proteins assemble into these machines and how specific pathways interact with each other to form a coherent system. It also will serve other parts of the Genomes to Life program focused on regulatory networks (Goal 2) and cell pathways and metabolic interactions in microbial communities (Goal 3) and will provide a central foundation for the ultimate problem of how an entire cell works (Microbial Cell Project). Further, it will provide key components of the integrated, highly annotated database containing data from all parts of the program and furnish input into robust computational tools developed for modeling and mining the data (Goal 4).

Specific Aims

Global characterization of multiprotein machines, and the essential associated technology and bioinformatics development, is the logical and proximal focal point in the initial years of the Genomes to Life program, as outlined in the first three specific aims below. Charting a course toward the longer-term goal of linking proteome architecture and dynamics to cellular and organismic function is the focus of the fourth specific aim.

Aim 1. Discover and define the repertoire of cellular protein complexes and machines in a comprehensive manner

A target for Genomes to Life is to be able within 5 years to measure and characterize thousands of protein machines per year. This capability would enable scientists to generate a draft map of the protein machinery of a typical microbe of interest to DOE in a year. Data-collection goals for the Genomes to Life program within the next decade include mapping (to a saturation of 80% or more) the protein machines of a model microbe and a model eukaryote, along with several additional microbes having biochemical pathways of particular interest to DOE missions. Currently, only a few of the most stable and common protein machines are well characterized, but data suggest that hundreds, if not thousands, of other machines operate together to carry out cellular functions. Many important associations may be less stable, less abundant, and more dynamic, and these will require new methods to fully probe their composition.

Aim 2. (a) Localize protein components within a multiprotein complex, and localize multiprotein machines within structural and functional compartments of the cell. (b) Determine the cellular and subcellular localization and co-localization of protein complexes, including their conditional and temporal variations. (c) Define physical relationships among protein complexes, and integrate this information with candidate functions. (d) Begin to develop high-throughput methods to characterize the protein-protein interfaces within and between complexes

A key challenge is to learn how multiprotein machines act and interact to regulate or execute cell functions. Data on coincident expression and cellular or subcellular localization can powerfully constrain possible functions for a given multiprotein machine. By coupling localization and co-localization information with genetic and biochemical data from diverse sources, the contributions of specific complexes to the survival and behavior of a cell can be postulated and ultimately tested. High-throughput implementation of new and existing technologies will be needed to achieve these goals.

An important aspect of understanding the assembly, stability, and function of protein machines is to characterize protein-protein proximity and interfaces within complexes and between interacting complexes in a high-throughput manner. When coupled with other information about the structure and interrelationships among proteins, this characterization will provide a comprehensive database for understanding spatial and temporal hierarchies in the assembly of protein complexes and for understanding how assembly and disassembly of these machines are organized and controlled.

Aim 3. Correlate information about multiprotein machines with structural information generated in the NIH Protein Structure Initiative to better understand the geometry, organization, and function of protein machines

Data from biochemistry, genetics, and molecular biology, together with other low-resolution information about protein machine structure, will be correlated with higher-resolution data from protein X-ray crystallography and nuclear magnetic resonance studies. Developing functional atomic-resolution models for representative molecular machines may lead to a better set of principles for understanding these machines.

Aim 4. Develop principles, theory, and predictive models for the structure, function, assembly, and disassembly of multiprotein complexes. Test the predictions of these models in experimental systems

Success in achieving Aims 1–3 will set the stage for causally linking proteome composition, architecture, and dynamics with cellular and organismic function. The ultimate test for a correct and useful understanding of causality in any system is the capacity to correctly predict how the system will change when perturbed by new external or internal stimuli, including in this instance genetic changes. A long-term aim of Goal 1 is to develop the theoretical infrastructure and the knowledge base needed to achieve this for the proteome at the level of multiprotein machines and the pathways and structures they comprise. This will require coupling increasingly sophisticated models with experimental tests of predictions from the models.

Computation Needs

The identification and characterization of the multiprotein machines of life involve substantial computational demands, ranging from sophisticated data analysis to atomic-scale simulations of protein interactions. Meeting these needs will require the development of new algorithms and databases and the use of very high performance computers. Following are the initial specific computational aims to be addressed for Goal 1:

- Improve bioinformatics methods needed to deconvolute mass spectroscopic protein-expression data, including isotopic labeling ratios, and to handle massive amounts of protein chip expression data.
- Adapt and develop databases and analysis tools for integrating experimental data on protein complexes measured with different methods under varied conditions.
- Develop algorithms for integration of diverse biological databases including transcriptome and proteome measurements, as well as functional and structural annotations of protein-sequence data.
- Develop modeling capabilities for simulating multiprotein machines and for predicting the behavior of protein machines in cell networks and pathways.

Some Federal Programs Complementary to Genomes to Life

Characterizing the composition and functions of multiprotein complexes and of the entire proteome would not be possible without the new knowledge of whole-genome DNA sequences and genome-wide comprehensive research strategies. Nor would it be possible without the success of research in other DOE programs and in programs across other agencies of the federal government. Genomes to Life will leverage and complement structural biology, genomics, proteomics, and cell biology research being conducted in other programs, including those outlined below. For more details on other related programs, see URL list on p. 70.

- Human Genome Programs at DOE and the National Institutes of Health (NIH): Help define the entire repertoire of human and model-organism genes and variations that encode the individual components of these molecular machines and instructions for their assembly and operation. (www.ornl.gov/hgmis, www.nhgri.nih.gov)
- NIH Protein Structure Initiative: Understand protein structural families, structural folds, and the relationship of structure and function by developing and using new methods and technologies for high-throughput protein-structure determination. (www.nigms.nih.gov/funding/psi.html)
- DOE Experimental and Computational Structural Biology Program: Develop and use novel approaches to understand the function of proteins and protein complexes relevant to the recognition and repair of DNA damage and the bioremediation of environmental contamination by metals and radionuclides. (www.science.doe.gov/ober/msd_struct_bio.html)
- Biocomplexity in the Environment Initiative of the National Science Foundation: Understand the nature and dynamics of complex interdependencies in systems ranging from individual cells to ecosystems to earth systems. (www.nsf.gov/home/crssprgm/be/start.htm)
- DOE microbial genomics research: Characterize microbes of potential use to DOE at the genomic and functional levels. This includes the Microbial Genome, the Carbon Sequestration, and the Natural and Accelerated Bioremediation Research programs. (www.ornl.gov/microbialgenomes, cdiac2.esd.ornl.gov, www.lbl.gov/NABIR)

Goal 2

Characterize Gene Regulatory Networks

Background and Strategy

Gene regulatory networks govern which genes are expressed in a cell at any given time, how much product is made from each one, and the cell's responses to diverse environmental cues and intracellular signals (see figure, p. 31). Among the myriad cellular outputs from such networks are the metabolic capabilities of microbes and the responses of cells to environmental stresses, toxins, and low doses of radiation, all topics of central importance to DOE bioscience missions. Because gene regulatory networks (GRNs) are so central to understanding and manipulating cells, they are critical to DOE's biology missions. For microbial systems, Genomes to Life will widen the scope of network analysis to encompass comprehensive mapping of all regulatory networks, including the "circuitry" that operates without altering gene expression.

Major objectives for Goal 2 are to discover the architecture, dynamics, and function of regulatory networks; make useful computational models of them; and learn how to adapt and design them. Because the theory and modeling of regulatory networks represent the core of a new discipline, Genomes to Life will also emphasize the recruitment and education of a cadre of regulatory biologists who specialize in the computational modeling and theory of regulatory networks that are intimately coupled with cycles of experimental testing and verification.

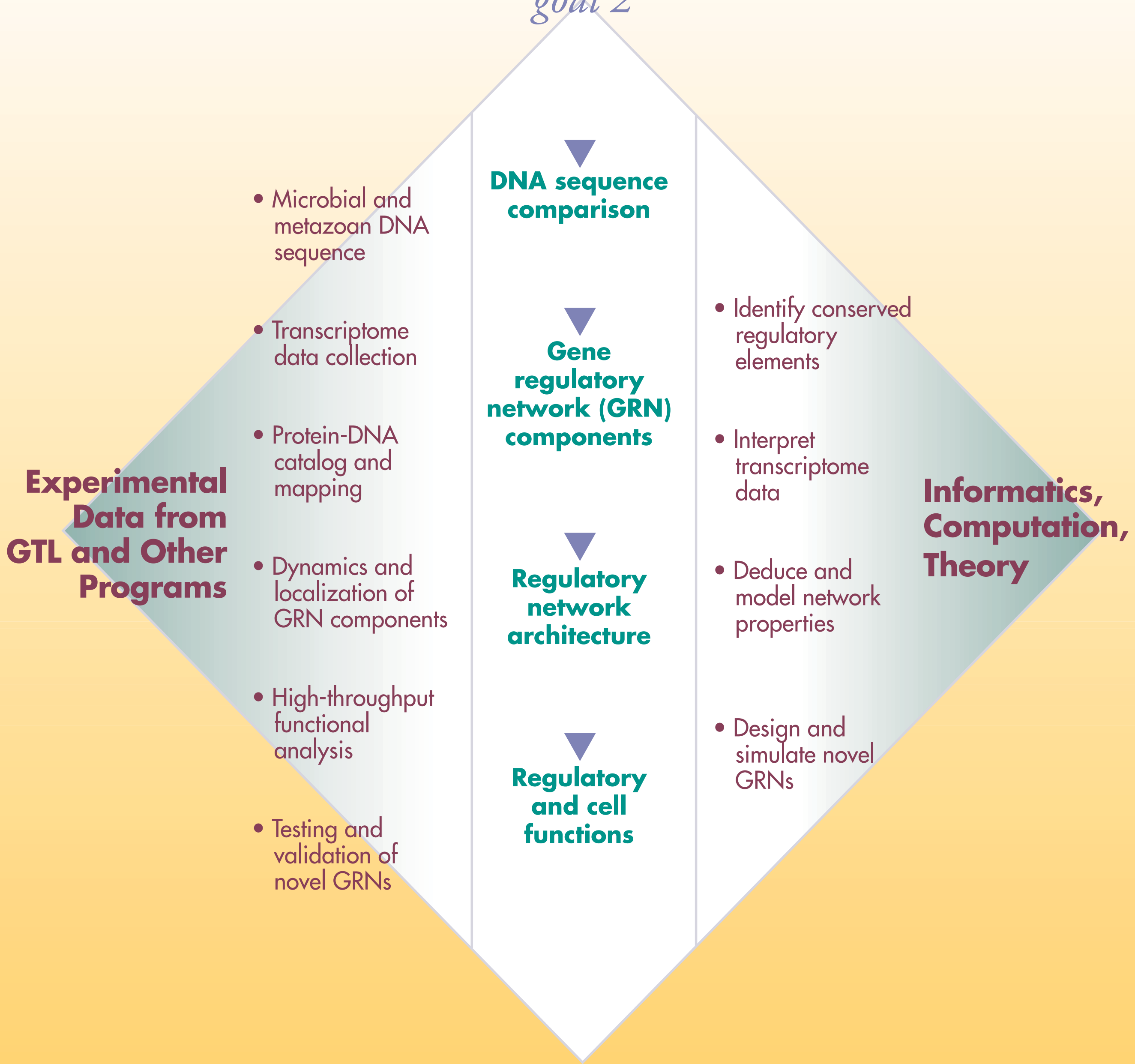
Within the network discovery portion of Goal 2, one activity is to map related networks at multiple nodes across phylogeny based on comparison of genome sequences. Knowledge of comparative network structure and function is likely to produce insights into fundamental issues in biology, in addition to providing essential information for later phases of Genomes to Life. One such basic question has emerged from the Human Genome Project: How can a multicellular organism as complex as a human, with all its cell and tissue types and functions, use only 2 or 3 times as many genes (about 30,000) as the simple worm and 5 to 10 times as many as a single-cell microbe? Much of the answer may be in the regulatory network architecture and complexity. The cis-acting regulatory apparatus (see sidebar, pp. 32–33) and, by implication, the gene regulatory networks of which it is a critical part are said to be at the nexus between evolution and development [C. H. Yuh et al., *Science* **279**(5358), 1896–1902 (1998)]. Complex body forms may therefore have emerged during evolution due mainly to the appearance

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

GENOMES *to* LIFE

CHARACTERIZE GENE REGULATORY NETWORKS

goal 2



Experimental Data from GTL and Other Programs

- Microbial and metazoan DNA sequence
- Transcriptome data collection
- Protein-DNA catalog and mapping
- Dynamics and localization of GRN components
- High-throughput functional analysis
- Testing and validation of novel GRNs

▼ **DNA sequence comparison**

▼ **Gene regulatory network (GRN) components**

▼ **Regulatory network architecture**

▼ **Regulatory and cell functions**

- Identify conserved regulatory elements
- Interpret transcriptome data
- Deduce and model network properties
- Design and simulate novel GRNs

Informatics, Computation, Theory

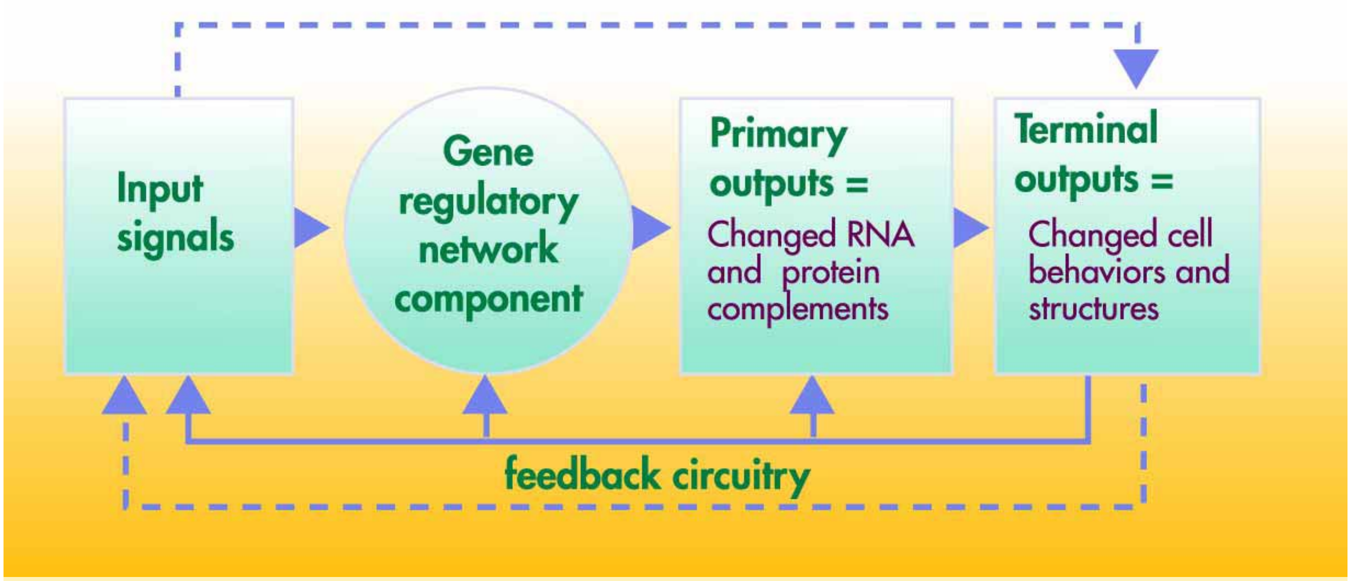
of more complicated and varied GRNs capable of controlling exquisitely complex combinatorial patterns of gene expression while the repertoire of genes itself is rather modestly expanded from much simpler organisms. If proved correct, an intriguing extension of this idea is that changes in the “wiring” of such networks might also dominate the functionally important differences between, for example, humans and our nearest animal relatives, the chimpanzees. Tracing the evolution of regulatory networks rather than individual genes will yield the information needed to probe these possibilities.

Specific Aims

Aim 1. Develop the capability within the next 10 years to comprehensively map microbial and metazoan regulatory circuitries. Use this capability to construct detailed GRN maps for specific subgenomic networks positioned across multiple species and to build comprehensive regulatory circuitry maps at the whole-genome level for selected microbes

Initial tasks for Aim 1 will be to identify and map core gene regulatory network components. In metazoa a major focus will be to identify cis-acting regulatory sequences in the genome and the regulatory proteins that interact with them. Integral to this effort is the task of relating the regulatory apparatus to the groups of target genes they regulate and to whatever is known about the function of those target genes. To map GRNs, several core technologies and approaches will likely be applicable to both microbial and eukaryotic systems, although pilot studies are needed to further define the best approach to use in genomes of varying sizes and structures. One such promising strategy is to use comparative genomics to initiate large-scale GRN component identification, focusing on candidate cis-regulatory sequences and the regulatory proteins with which they interact. Results from comparative sequence analysis would then be integrated with data from other key technologies such as large-scale gene-expression analysis, comprehensive loss-of-function and gain-of-function genetic analyses, and measures of in vivo protein-DNA interactions, and proteome status, among others.

Other critical elements in network mapping will come from activities encompassed by Goal 1 or by specific adaptation of technology developed in that work to regulatory network components. This includes learning the composition of multiprotein complexes that assemble on DNA to regulate gene expression; learning the composition and regulatory actions of protein machinery that govern


Input signals
Gene regulatory network component
Primary outputs = Changed RNA and protein complements
Terminal outputs = Changed cell behaviors and structures
feedback circuitry

Gross anatomy of a minimal gene regulatory network (GRN) embedded in a regulatory network. A regulatory network can be viewed as a cellular input-output device. At minimum, a gene regulatory network typically contains the following components: (1) an input signal reception and transduction system that mediates intra- and extracellular cues (left box; often, more than one signal impinges on a given target gene); (2) a "core component" complex composed of trans-acting regulatory proteins and cognate cis-acting DNA sequences (circle; functionally similar components may be associated with multiple target genes, resulting in similar gene-expression patterns); and (3) primary molecular outputs from target genes, which are RNA and protein (box to right of circle). The net effects are changes in cell phenotype and function (right box). Direct and indirect feedbacks typically are important. More realistic networks often feature multiple tiers of regulation, with first-tier gene products regulating expression of another group of genes, and so on. Beyond GRN boundaries are signaling responses and feedbacks, such as those that drive bacterial chemotaxis, which do not involve regulation of gene expression but instead act directly on proteins and protein machine assemblies (dashed arrows). Some regulatory networks have no embedded GRN component.

post-transcriptional and post-translational regulation, and determining subcellular localization of regulatory proteins and how that localization changes as a function of circuit dynamics.

Vigorous application of a comprehensive genome-wide approach to network mapping in selected microbes has the potential to yield the first complete dissection of the regulatory networks that run a living cell. The most experimentally advanced microbial systems already offer a uniquely powerful combination of approaches that are variously performed in vivo, in vitro, and in silico on a comprehensive genome-wide scale. The result is integrative knowledge of the subsystems and systems contained in bacterial gene regulatory circuitry [M. T. Laub et al., *Science* **290**, 2144–48 (2000) and Goal 2 sidebar, p. 33]. Moreover, recent discoveries clearly show that regulatory networks in both microbes and metazoa employ many mechanisms distinct from both transcription and translation. Examples include active control of protein turnover, dynamic localization of regulatory and structural

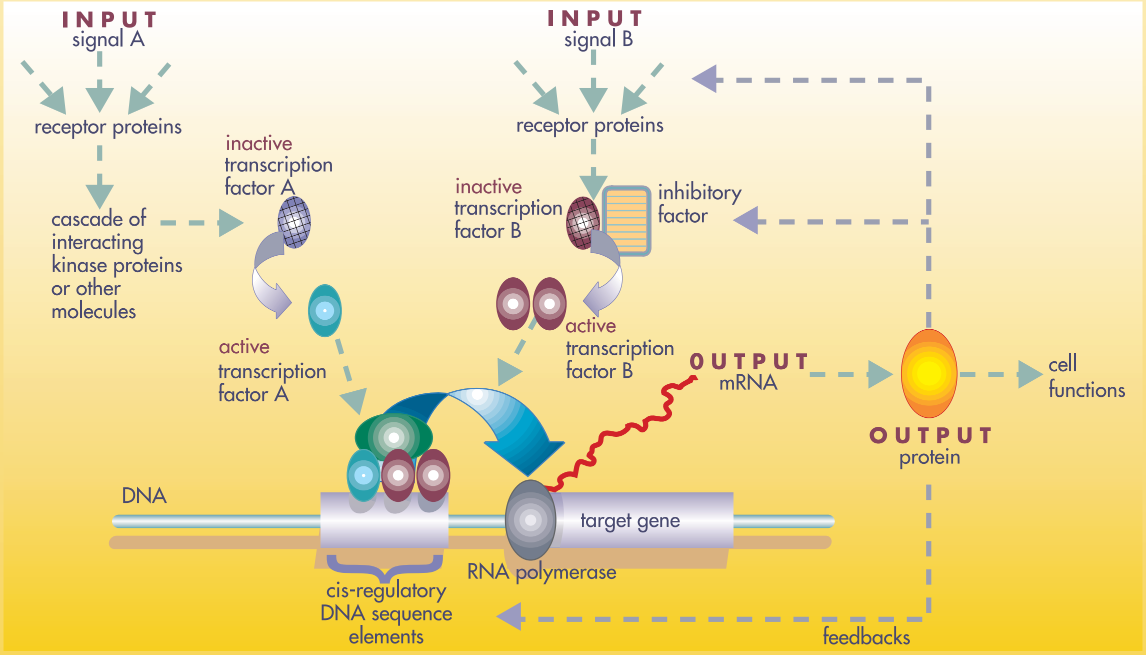
Gene Regulatory Networks

Gene regulatory networks (GRNs) are the on-off switches and rheostats of a cell operating at the gene level. They dynamically orchestrate the level of expression for each gene in the genome by controlling whether and how vigorously that gene will be transcribed into RNA. Each RNA transcript then functions as the template for synthesis of a specific protein by the process of translation. A simple GRN would consist of one or more input signaling pathways, regulatory proteins that integrate the input signals, several target genes (in bacteria a target operon), and the RNA and proteins produced from those target genes. In addition, such networks often include dynamic feedback loops that provide for further regulation of network architecture and output. As indicated in the schematic below, input signaling pathways transduce intracellular and/or extracellular signals to a group of regulatory proteins called transcription factors. Transcription factors activated by the signals then interact, either directly or indirectly, with DNA sequences belonging to the specific genes they regulate. The factors also interact with each other to form multiprotein complexes bound to the DNA.

GRNs act as analog biochemical computers to specify the identity and level of expression of groups of target genes. Central to this computation are DNA recognition sequences with which transcription factors associate. Every gene has its own novel “cis-acting” sequence elements. They vary greatly in complexity from one gene to another and from generally simpler structures in bacteria to more complex structures in multicellular organisms. When active transcription factors associate with the cis-elements of their cognate target genes, they can function to specifically repress (down-regulate) or induce (up-regulate) synthesis of the corresponding RNA. The immediate molecular output of a gene regulatory network is the constellation of RNAs and proteins encoded by network target genes. The resulting cellular readouts are changes in the structure, metabolic capacity, or behavior of the cell mediated by new expression of up-regulated proteins and elimination of down-regulated proteins.

GRNs are remarkably diverse in their structure, but several basic properties are illustrated in the figure below. In this example, two different signals impinge on a single target gene where the cis-regulatory elements provide for an integrated output in response to the two inputs. Signal molecule A triggers the conversion of inactive transcription factor A (green oval) into an active form that binds directly to the target gene’s cis-regulatory sequence. The process for signal B is more complex. Signal B triggers the separation of inactive B (red oval) from an inhibitory factor (yellow rectangle). B is then free to form an active complex that binds to the active A transcription factor on the cis-regulatory sequence. The net output is expression of the target gene at a level determined by the action of factors A and B. In this way, cis-regulatory DNA sequences, together

A GENE REGULATORY NETWORK


INPUT
signal A
receptor proteins
cascade of interacting kinase proteins or other molecules
inactive transcription factor A
active transcription factor A
INPUT
signal B
receptor proteins
inactive transcription factor B
inhibitory factor
active transcription factor B
OUTPUT
mRNA
OUTPUT
protein
cell functions
DNA
target gene
RNA polymerase
cis-regulatory DNA sequence elements
feedbacks

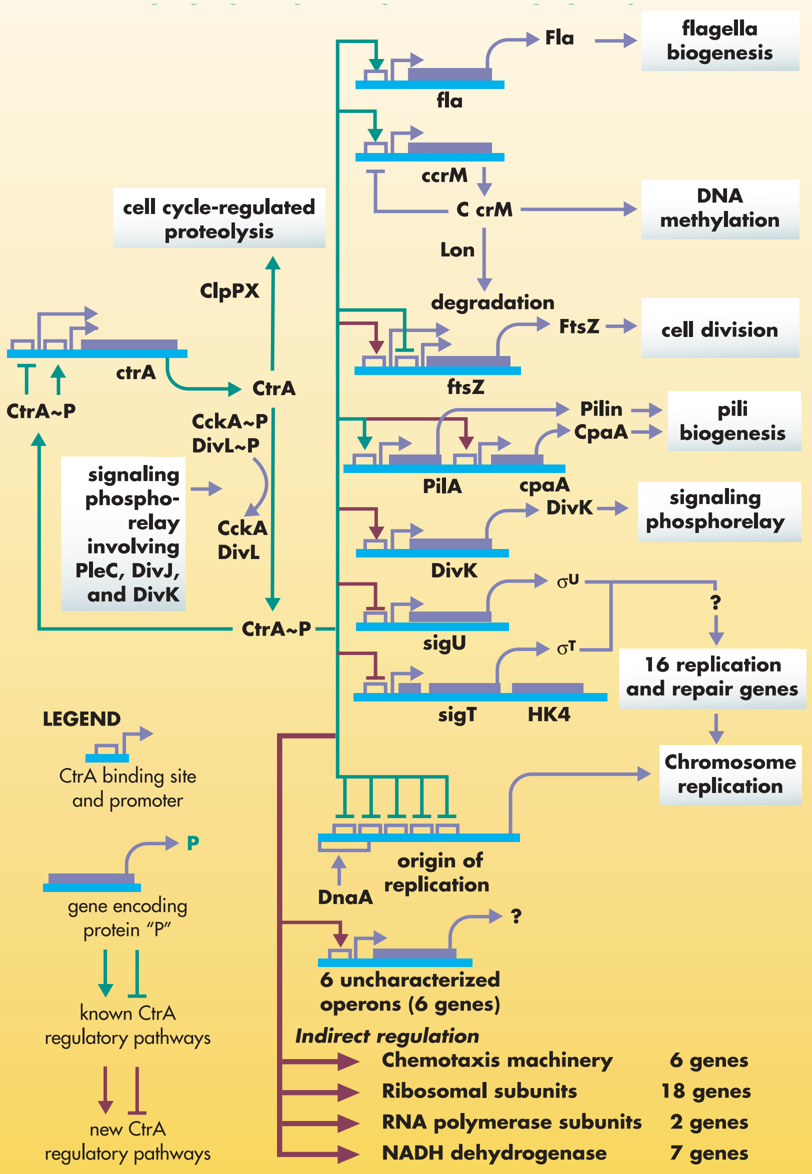
with the proteins that assemble on them, integrate information from multiple signaling inputs to produce an appropriately regulated readout. A more realistic network might contain multiple target genes regulated by signal A alone, others by signal B alone, and still others by the pair of A and B.

Co-regulated target genes often code for proteins that act together to build a specific cell structure or to effect a concerted change in cell function. For example, genes encoding components of the multiprotein proteasome machine (see Goal 1 sidebar, pp. 22–23) are co-regulated at the RNA level. This was shown by microarray gene chip analyses in yeast cells, and each gene was found to possess a similar cis-regulatory DNA sequence that mediates binding of a particular transcription factor. Similarly, a bacterium may respond to a shortage of its preferred energy source by activating expression of genes whose protein products function in a biochemical pathway that allows it to use a different, more abundant source of energy.

Some genes are regulated by a single input mechanism, but, especially in higher organisms, a gene often responds to information from multiple signals via the activity of diverse transcription factors. For example, in human fibroblast cells responding to a “growth factor” impinging on cell surface signal receptors, a platoon of “immediate early” genes is upregulated as the first step in the complex process of cell proliferation. Some of the same genes, though not all of them, can be activated in brain cells by the distinct stimulus of seizure. This network also illustrates that GRNs can be multitiered. The first signal (growth factor) initiates expression of “immediate early” target genes, which include transcription factors such as *c-Fos* and *c-Jun*. These transcription factors then cause a second group of target genes, called “early genes,” to be expressed. Among these early genes are other transcription factors such as *c-Myc*. They regulate expression of yet another group of genes of a “delayed early” group. In this way a multitiered GRN cascade can be constructed, replete with feedbacks and crosstalk to other networks.

A major gene regulatory network in the bacterium *Caulobacter* is now beginning to be mapped in a comprehensive manner based on genome-wide expression analyses coupled with genetic methods [M. T. Laub et al., *Science* **290**, 2144–48 (2000)]. *Caulobacter* has about 3000 genes, of which almost 20% were found to be differentially expressed during the cell division cycle. Of the 553 responding genes, 38 are likely to be direct targets of a sequence-specific DNA-binding protein called CtrA and another 144 are indirectly regulated by CtrA. A first-pass connectivity map of the CtrA gene regulatory network derived from this study is summarized in this figure. Green indicates previously known relationships, while red indicates relationships that emerged from this global gene-expression study performed using microarray technology.

A *CAULOBACTER* CELL DIVISION GRN


Fla
flagella biogenesis
fla
ccrM
C crM
DNA methylation
cell cycle-regulated proteolysis
Lon
ClpPX
degradation
FtsZ
cell division
ctrA
CtrA
ftsZ
CtrA~P
CckA~P
DivL~P
Pilin
CpaA
pili biogenesis
PilA
cpaA
signaling phospho-relay involving PleC, DivJ, and DivK
CckA
DivL
DivK
signaling phosphorelay
DivK
σU
?
CtrA~P
sigU
σT
16 replication and repair genes
sigT
HK4
LEGEND
CtrA binding site and promoter
Chromosome replication
P
gene encoding protein “P”
origin of replication
DnaA
?
6 uncharacterized operons (6 genes)
known CtrA regulatory pathways
Indirect regulation
Chemotaxis machinery 6 genes
Ribosomal subunits 18 genes
RNA polymerase subunits 2 genes
NADH dehydrogenase 7 genes
new CtrA regulatory pathways

proteins, and complex phospho-transfer pathways. In addition, the cell membrane appears to be an integral component of essential cell signaling processes. Including nontranscriptional systems is therefore critical for a full understanding of regulatory circuitry in all organisms. Genomes to Life will first capitalize on the relative simplicity of microbes to extend network analysis to include all regulatory circuitry.

Aim 2. Verify regulatory circuit architecture and connect regulatory network properties with their biological outputs

Genomes to Life's ambitious next step will be to map higher-order connectivity between these circuits and tie them to cellular functions and phenotypes. This will begin with experimental verification of network composition and architecture generated by Aim 1. The most effective, efficient, and scalable methods for doing this for both microbes and more complex creatures will be explored in the early years of Genomes to Life. On the output side of the network equation, Genomes to Life will seek to assign cellular function and phenotype to each subsystem, and then to link subsystems to learn higher-order connectivity. This is an enormous challenge, the dimension of which will become apparent only when we gain a much more comprehensive view than we now have. Genomes to Life will therefore draw on the data and resources from this and other DOE programs as well as those generated by the National Institutes of Health and the National Science Foundation. Functional annotation in Genomes to Life will require large bioinformatics and computational components, some aspects of which will make demands considerably greater than that provided by whole-genome sequence assembly of the human genome (see Goal 4, p. 44).

Aim 3. Develop a theoretical framework and associated set of computational modeling tools to predict the dynamic behavior of natural or designed regulatory networks

Developing a comprehensive view of the basic architecture of eukaryotic gene regulatory or microbial regulatory networks will be the culmination of a discovery process that began 40 years ago with the *Escherichia coli* Lac operon. Although that accomplishment will be a grand one, even a complete wiring diagram will not reveal how any regulatory network really works, nor will it provide a solid basis for using them, for modifying them in useful ways, or for designing new ones. To master the complexities of regulatory switches, oscillators, and more complex functions will require a predictive theoretical framework and computational horsepower of teraflop speed. To meet this challenge, Genomes to Life will seek to nurture and accelerate emerging capabilities that include new concepts combined with relevant ideas from engineering, applied mathematics, and other disciplines.

Aim 4. Learn to modify natural networks and design novel ones for DOE mission purposes

A major motivation for mapping regulatory networks and then developing predictive computational models and simulators is to ultimately learn how to prudently use such networks to develop biological solutions to important problems such as bioremediation. This is a long-term goal that will require results from Aims 1–3 above to bring network modification and design onto a firm footing. However, even in the early years of Genomes to Life, developing the capability to design increasingly complex networks with useful control properties will be an important activity.

Computation Needs

The task of elucidating and adapting the circuitry of gene regulatory networks will be dependent on computational methods to identify and characterize regulatory sequences and computer models of the regulatory networks. The computational research tasks for Goal 2 involve developing methods to do the following:

- Extract regulatory elements, including operon and regulon sequences, using sequence-level comparative genomics.
- Simulate regulatory networks using both nondynamical models of regulatory capabilities and dynamical models of regulatory kinetics.
- Predict the behavior of modified or redesigned gene regulatory networks.

Goal 3

Characterize the Functional Repertoire of Complex Microbial Communities in Their Natural Environments at the Molecular Level

Background and Strategy

Several of DOE's most distinguishing missions—energy security, environmental stewardship, science, and technology development—are linked directly to gaining a better understanding of the functions of the microbial communities inhabiting the planet. These communities catalyze such crucial environmental processes as the recycling of carbon, nitrogen, and many trace nutrients. Most notably for DOE missions, some microbial communities catalyze transformations of contaminants from toxic to benign forms and thus might be managed to accomplish remediation in situ. Other communities catalyze the transformations of reduced and oxidized forms of carbon and thereby contribute to the global carbon balance between atmospheric and sequestered carbon.

Microorganisms are the largest reservoir of genetic and biochemical diversity on earth. They have been evolving for around 3.7 billion years to colonize virtually every environment, often thriving under extremes of nutrient concentration, pH, salinity, pressure, and temperature. In the past several decades, new methods for examining microbial communities have revealed that uncultured microbes make up more than 99% of many natural microbial communities. Because of the uncultured status of these microbes and the historical reliance on culture methods for study, scientists today have almost no knowledge about the ecology, physiology, diversity, and biochemistry of earth's greatest fraction of life. Recent advances, however, have demonstrated that DNA of sufficient quality to enable sequencing of relatively long segments can be isolated directly from environmental samples. This genomic information is a tremendous resource for examining the function of microbial communities.

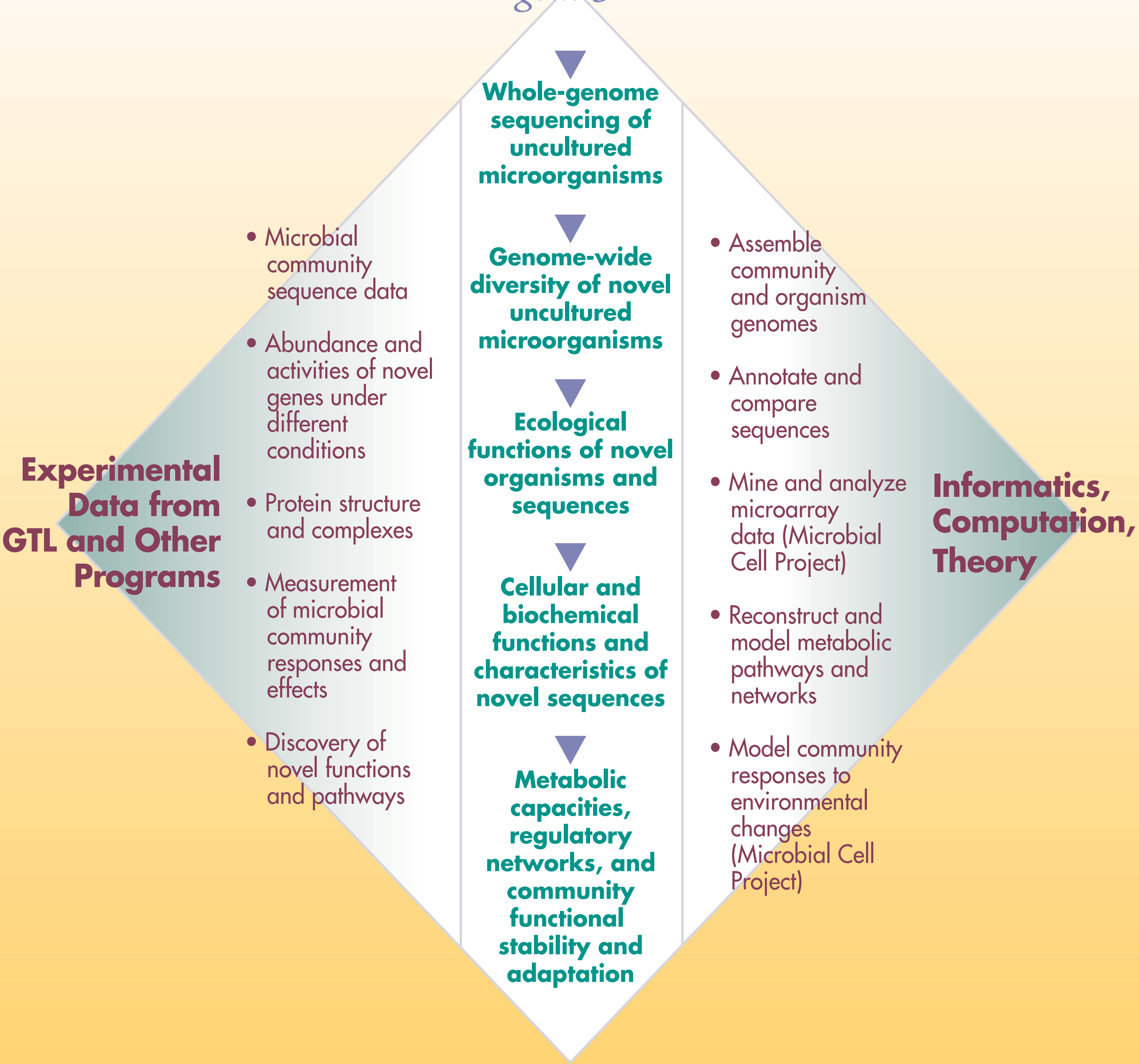
The overall objective of this goal in the Genomes to Life program is to dramatically extend current scientific and technical understanding of the genetic diversity and metabolic capabilities of microbial communities in the environment, especially those related to remediation, biogeochemical cycles, climate changes, and energy production. The program will focus on defining the repertoire of metabolic capabilities as embodied in the collective community's genomic sequence. Determining and annotating such sequences to infer the presence of protein complexes and regulatory networks responsible for function will give

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

GENOMES to LIFE

CHARACTERIZE THE FUNCTIONAL REPERTOIRE OF COMPLEX MICROBIAL COMMUNITIES IN THEIR NATURAL ENVIRONMENTS AT THE MOLECULAR LEVEL

goal 3



Whole-genome sequencing of uncultured microorganisms

Genome-wide diversity of novel uncultured microorganisms

Ecological functions of novel organisms and sequences

Cellular and biochemical functions and characteristics of novel sequences

Metabolic capacities, regulatory networks, and community functional stability and adaptation

Experimental Data from GTL and Other Programs

- Microbial community sequence data
- Abundance and activities of novel genes under different conditions
- Protein structure and complexes
- Measurement of microbial community responses and effects
- Discovery of novel functions and pathways

Informatics, Computation, Theory

- Assemble community and organism genomes
- Annotate and compare sequences
- Mine and analyze microarray data (Microbial Cell Project)
- Reconstruct and model metabolic pathways and networks
- Model community responses to environmental changes (Microbial Cell Project)

investigators a first glimpse at the community's metabolic capabilities, including those of its uncultured members.

One of the unifying themes observed in biology is the harvesting of energy through metabolic networks that link electron donors with acceptors. Although the genetic diversity in microbes is great, the number of known strategies for capturing energy from the environment is relatively small; this leads to the hypothesis that the diversity in functions carried out by microbial communities is much less than the sum of all represented genomes. Such conservation of capacities makes tractable the goal of describing major protein machines and regulatory networks that power microbial community functions. Scientific insights provided by Goals 1 and 2, together with this goal's focus on obtaining, assembling, and understanding genomic sequence data from microbial communities, will permit the testing of this hypothesis and the application of the derived scientific understanding to DOE missions.

Key technologies needed to achieve this goal include the following:

- New approaches for recovering RNA and high-molecular-weight DNA from environmental samples
- New approaches for isolating single cells of uncultured microorganisms.
- New parallel comparative approaches that allow unique microbial community DNA fragments to be identified and the community to be characterized in automated high-throughput ways.
- Novel technologies and approaches for defining the functions of genes from uncultured microorganisms.
- Advanced methods for community genome sequence assembly, genome comparison, microarray data analysis, and data management.

Specific Aims

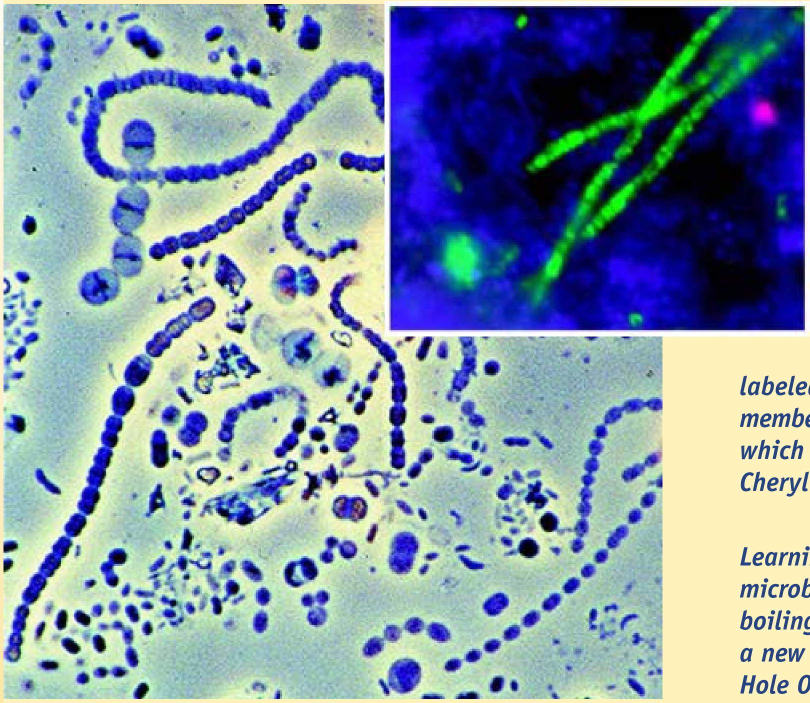
Aim 1. Determine whole-genome sequences of dominant uncultured microorganisms

Recent advances make possible the proposal to obtain whole-genome sequence information from uncultured microorganisms. Until investigators learn to culture these microorganisms, predictions of genome-specified protein complexes and regulatory networks and the functions they engender are the most powerful tools available for understanding the metabolic capabilities and ecological roles of uncultured microbes. In this aim, the following objectives will be addressed:

Exploring the Functions of Microbes and Their Communities

Microbes represent the greatest reservoir of genetic and biochemical diversity on the planet. They drive the chemistry of life, do much of the biogeochemical cycling that keeps the world habitable, and even affect the global climate. Over billions of years, microbes have developed a wealth of functions that enable their survival in virtually every environmental niche, often where no other life forms exist. Knowledge about the metabolic and regulatory pathways of microbes and their communities will provide the foundation to begin understanding and using their remarkable capabilities, especially those related to environmental remediation, biogeochemical cycles, climate changes, and energy production.

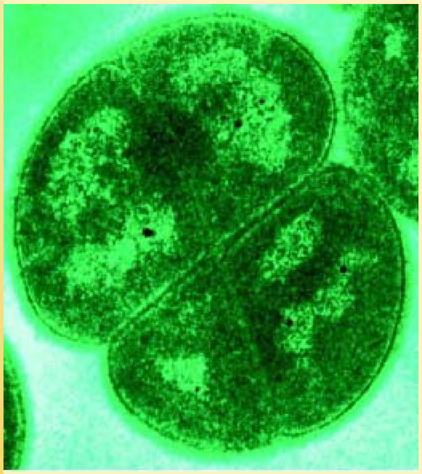
The vast majority of microbes—often thousands of species in a single environmental niche—cannot currently be grown in the laboratory, and estimates are that less than 1% have even been identified. Recent advances in molecular methods now enable an entirely different approach for tapping into the potentially limitless resource of uncultured bacteria: wholesale direct sampling of the DNA present in an entire environmental niche. The genomic information represented by such a "community genome" offers a tremendous resource for examining the extent and patterns of microbial genetic diversity and metabolic capabilities in the natural ecosystems of importance to DOE.



Annotating DNA sequences from microbial communities will offer a first glimpse of the collective metabolic capabilities present in a natural ecosystem, including those of its uncultured members.

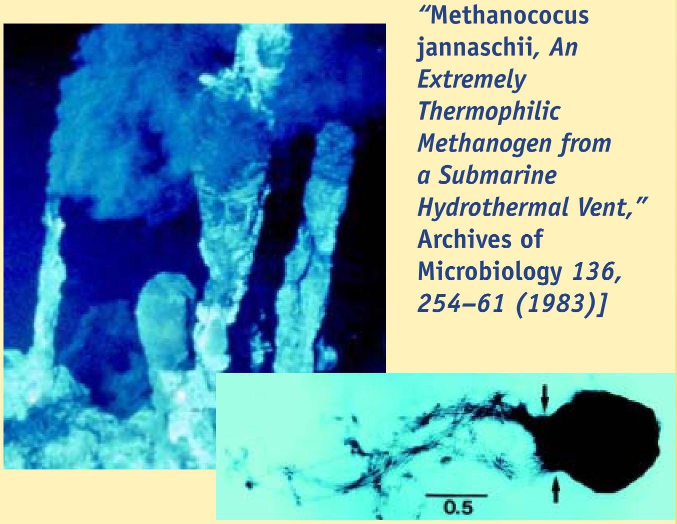
The large image at left illustrates the morphological diversity found in a natural microbial community. [Source: Frank Dazzo, Center for Microbial Ecology, Michigan State University]

The uncultured cells in the inset picture were labeled with a fluorescent molecule used to identify members of the Acidobacterium *division of bacteria, which has only three known cultured members. [Source: Cheryl Kuske et al., Los Alamos National Laboratory]*



Deinococcus radiodurans *thrives in radiation levels thousands of times higher than those that would kill most organisms, including humans, and it may prove useful in bioremediation of toxic waste. [Source: Uniformed Services University of the Health Sciences]*

Learning to control methane-production pathways in microbes such as Methanococcus jannaschii *(found in boiling hydrothermal sea vents) could one day provide a new resource for clean energy. [©Stan Watson, Woods Hole Oceanographic Institute; inset: ©Springer-Verlag, "*Methanococcus jannaschii*, An Extremely Thermophilic Methanogen from a Submarine Hydrothermal Vent,"* Archives of Microbiology *136, 254–61 (1983)]*



- Determine the genetic diversity of uncultured microorganisms.
- Understand the relationships between uncultured microorganisms and cultured microorganisms at the whole-genome level.
- Determine whether known genes, pathways, regulatory networks, and protein machines needed for survival, growth, replication, and environmental adaptation are conserved between cultured and uncultured microorganisms.

To achieve these objectives within a decade, whole or nearly whole genome sequences will be obtained from 100 to 200 closely and distantly related uncultured microbial species from widely distributed microbial groups in the environment. Current strategies for obtaining whole-genome sequences of uncultured microorganisms are to clone and sequence the desired high-molecular-weight DNA directly from community DNA. This strategy will require separation and purification of targeted cells or their DNA in sufficient quantity and purity to enable sequencing of the genome. The insights gained from knowledge of the individual genomes from environmentally important populations of uncultured organisms are expected to be extremely valuable in understanding their biogeochemical roles and in assembling and understanding the microbial community genome from more complex natural environments.

Aim 2. Identify the extent and patterns of genetic diversity in microbial communities

So far, 16S rRNA gene-based phylogenic studies with a variety of environmental samples have yet to adequately define the extent of microbial phylogenetic diversity. Information obtained from the microorganisms' collective genome may be used to assess the extent and reservoir of genetic diversity, the patterns of diversity within the range of phylogenic groups, and the relationships of diversity to site characteristics. The following objectives will be addressed:

- Determine the extent and patterns of phylogenetic diversity in microbial communities from different environments.
- Understand how microbial communities are genetically adapted to different environments.
- Determine whether microbial communities conserve metabolic function in spite of extensive individual phylogenetic diversity.

To achieve these objectives within a decade, genome sequences will be obtained from 10 to 20 microbial communities of various degrees of complexity. One strategy for understanding the extent and pattern of genetic diversity in microbial communities is to sequence bacterial artificial chromosome (BAC) clones from individual microbial communities by the shotgun approach. Comparing BAC clone sequences should lead to insights into community genetic diversity and metabolic capacity.

Aim 3. Understand the ecological functions of the uncultured microorganisms

Once whole-genome sequences are obtained from novel uncultured microorganisms and from microbial communities, the next critical step is to identify the metabolic functions these genomes encode and to understand how those functions contribute to the community's ecological role in the environment. The following issues will be addressed:

- Examine unique roles of novel uncultured microorganisms in ecosystems important to DOE.
- Understand how uncultured microorganisms interact with other microbial populations and how they respond to environmental changes.
- Determine whether the microbes are involved in biogeochemical processes of interest to DOE and how these activities can be managed to improve the environment.

A basic strategy for understanding the ecological roles of uncultured organisms is to extensively evaluate their abundance, distribution, gene expression, and biochemical functions in response to environmental changes in both laboratory and field studies. For the field studies, specific emphasis will be on the habitats important to DOE's missions involving bioremediation, carbon sequestration, global changes, and energy production.

Aim 4. Determine cellular and biochemical functions of genes discovered in uncultured community members

The sequencing accomplished in Aims 1 and 2 is expected to identify a large number of putative genes of completely unknown function as well as known genes likely to have unique and useful characteristics. Aim 4 is directed at discovering the cellular and biochemical functions and useful characteristics of these genes. The following issues will be addressed:

- Determine the cellular and biochemical functions of the unknown genes discovered in uncultured microorganisms.
- Determine the protein complexes unique to uncultured microorganisms.
- Determine whether these unique characteristics can be used for protein engineering.

Determining the functions of genes from uncultured microorganisms at the cellular and biochemical levels is extremely difficult due to the uncultured status of the target microorganisms and the lack of genetic-manipulation systems. A basic strategy for understanding their functions is to express these genes in a heterologous host and subsequently examine their catalytic function, if possible, and, if not, to characterize protein structure with X rays, neutron scattering, nuclear magnetic resonance, and mass spectrometry. The target genes should include those that appear to be members of protein complexes studied in Goals 1 and 2, genes that appear to be novel varieties of those that catalyze important ecosystem functions, and genes that are novel but in dominant, uncultured members of targeted ecosystems.

Aim 5. Understand the genetic basis of microbial community functional stability and adaptation in environments important to DOE

The relationship between diversity and stability of biological communities is a longstanding controversy in macrocommunity ecology. Understanding the genetic basis and factors controlling microbial community stability and adaptation is of great importance in managing microbial communities to bioremediate contaminated sites, sequester carbon from the atmosphere, and contribute to sustainable energy production. The following issues will be addressed:

- Determine the genetic basis for functional stability and adaptation of microbial communities.
- Understand how a microbial community's functional stability is related to its genetic and metabolic diversity.
- Determine whether the functional stability and future status of a microbial community can be predicted based on the conservation of metabolic functions and the differentiation of individual microbial populations.
- Understand whether a desired stable function can be achieved by manipulating a microbial community's metabolic traits.

A basic strategy for understanding the genetic basis and factors controlling microbial community stability and adaptation is to compare their diversity and metabolic capacities; these comparisons will be carried out under different stress conditions in similar habitats by identifying and selectively sequencing both common and different DNA fragments. Laboratory systems to study the responses of microbial communities to environmental stressors also are needed to establish cause-and-effect relationships.

Computation Needs

There are many computational challenges to characterizing the composition and functional capability of microbial communities. New algorithms for DNA sequence assembly and annotation will be required to analyze the multiorganism sequence data, and new modeling methods will be required to predict the behavior of microbial communities. The computational research tasks will be to develop methods to:

- Deconvolute mixtures of genomes sampled in the environment and identify individual organisms.
- Facilitate multiple-organism shotgun-sequence assembly.
- Improve comparative approaches to microbial sequence annotation and gene finding and use them to assign functions to genes where possible.
- Accomplish pathway reconstruction from sequenced or partially sequenced genomes and evaluate the combined metabolic capabilities of heterogeneous microbial populations.
- Integrate regulatory-network, pathway, and expression data into integrated models of microbial community function.

Goal 4

Develop the Computational Methods and Capabilities to Advance Understanding of Complex Biological Systems and Predict Their Behavior

Background and Strategy

The Genomes to Life program involves a new approach to biology. It combines large experimental data sets with advanced data management, analysis, and computational simulations to create predictive models of microbial function and of the protein machines and pathways that embody those behaviors. The program's computational component will require developments ranging from more efficient modeling tools to fundamental breakthroughs in mathematics and computer science as well as algorithms that efficiently use the fastest available supercomputers. Vast sets of genome sequences, protein structures, interactions, and expression profiles will be generated by this and other biology initiatives. The information must be annotated and archived to provide raw data for computer models of biochemical pathways, entire cells, and, ultimately, microbial ecosystems.

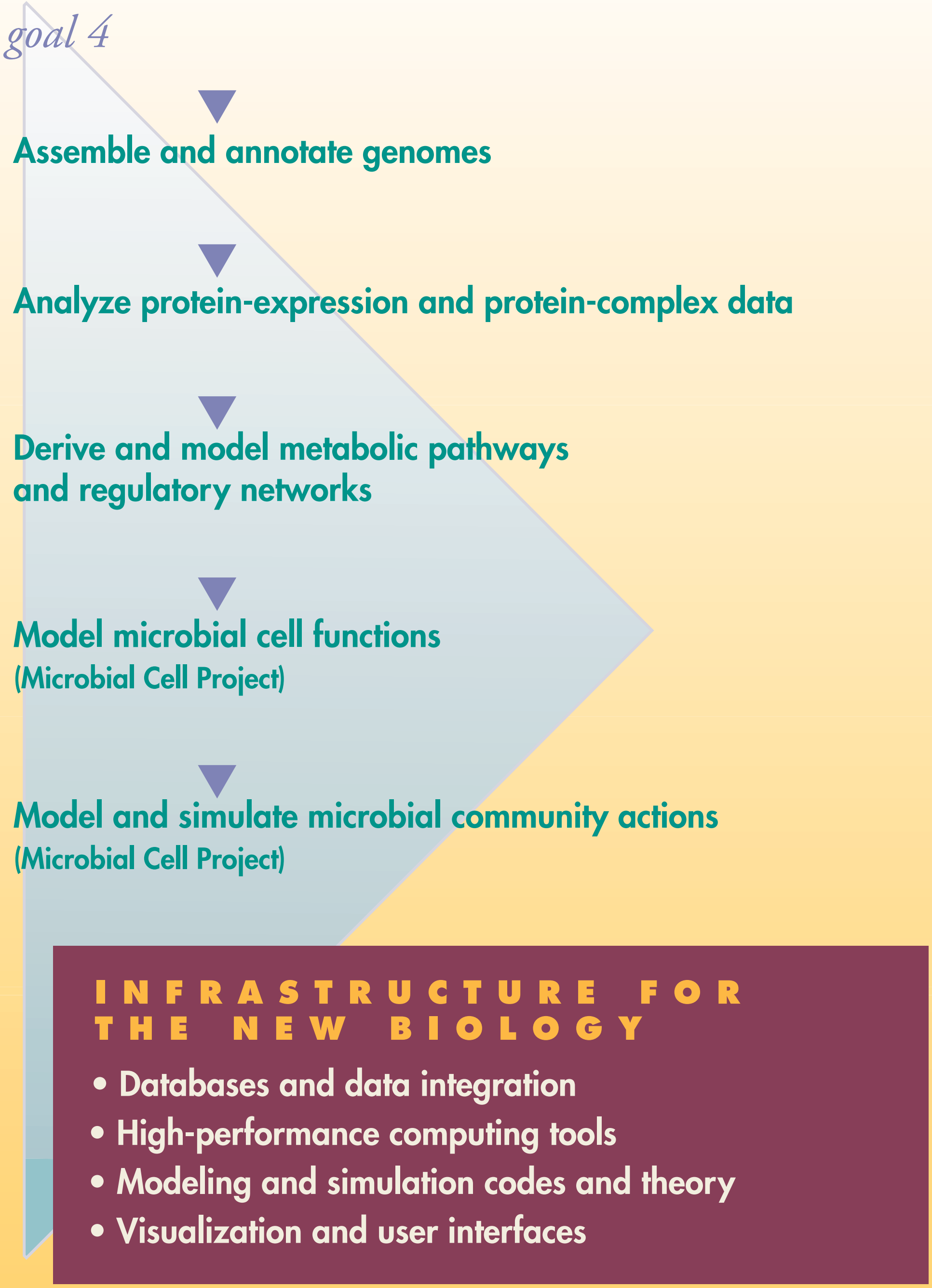
A long-term goal of the computational-modeling section of Genomes to Life is to develop the next generation of methods for simulating cellular behavior and pathways. Other goals are to create molecular-modeling and bioinformatics tools for studying multiprotein complexes, along with new computational methods to explore the functional diversity of microbes (see table, p. 49). In addition to developing new technologies, Genomes to Life will leverage information and methods from a variety of sources, including cell systems data from the DOE Microbial Cell Project, protein structures produced in the NIH Protein Structure Initiative, the Protein Data Bank, databases of metabolic processes such as KEGG and WIT, and a host of available analytical tools in such areas as molecular dynamics, mass spectrometry, and pathway modeling and simulation.

Successful production of advanced tools for computational biology will require the sustained efforts of multidisciplinary teams, teraflop-scale and faster supercomputers, and considerable user expertise. This task for the entire biological community will involve many institutions and federal agencies, led in many aspects by the National Institutes of Health and the National Science Foundation. A central component of Genomes to Life will be the establishment of effective partnerships with these and other agencies to ensure that computational tools and standards are widely adopted and to eliminate redundant efforts.

GENOMES to LIFE

DEVELOP THE COMPUTATIONAL METHODS AND CAPABILITIES TO ADVANCE UNDERSTANDING OF COMPLEX BIOLOGICAL SYSTEMS AND PREDICT THEIR BEHAVIOR

goal 4



Assemble and annotate genomes

Analyze protein-expression and protein-complex data

Derive and model metabolic pathways and regulatory networks

Model microbial cell functions (Microbial Cell Project)

Model and simulate microbial community actions (Microbial Cell Project)

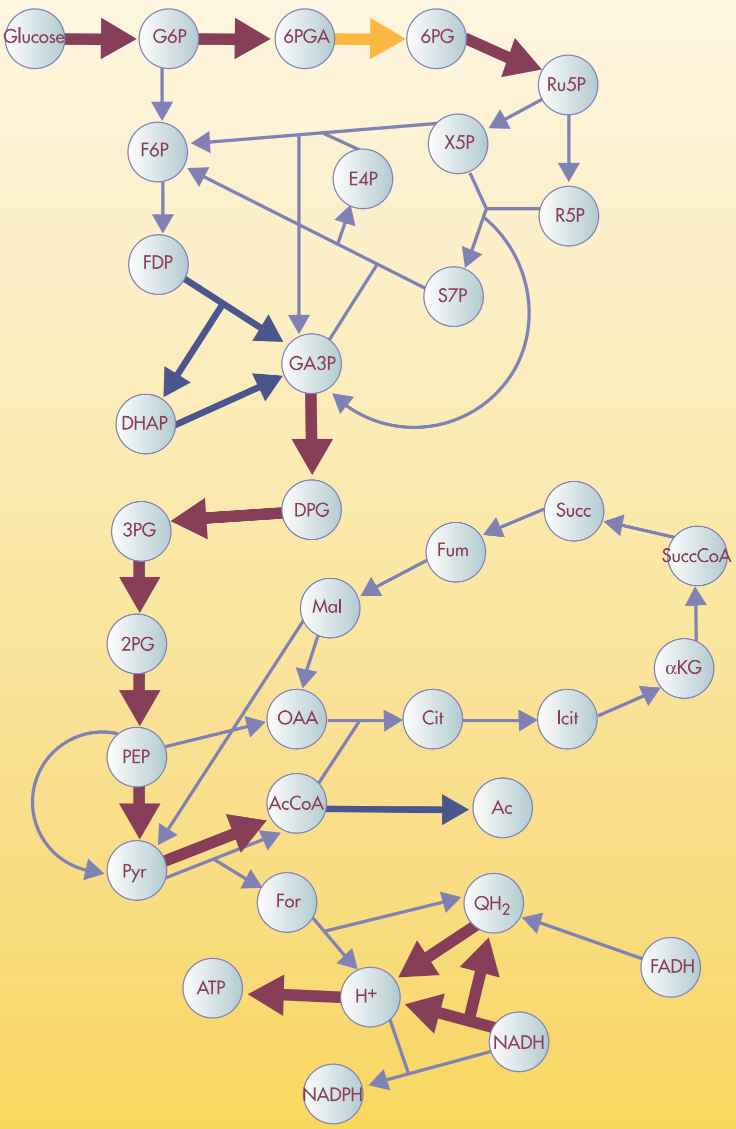
INFRASTRUCTURE FOR THE NEW BIOLOGY

- Databases and data integration
- High-performance computing tools
- Modeling and simulation codes and theory
- Visualization and user interfaces

Models for the New Biology

Even the simplest microbes command a vast repertoire of complex self-regulating chemical and physical processes. An ultimate goal of the Genomes to Life program is to develop predictive models of microbial cell and community functions; because of the complexity of microbes, however, the first generation of models will not reach the level of individual biochemical reactions. Instead, they will operate at a level in which cellular pathways are described either qualitatively (as being present or absent) or quantitatively in terms of average concentrations and activity rates derived from experimental data. Despite their lack of chemical detail, these models will provide a powerful tool for integrating and analyzing the very large new biological data sets and, in some cases, predicting cellular behavior under changing conditions.

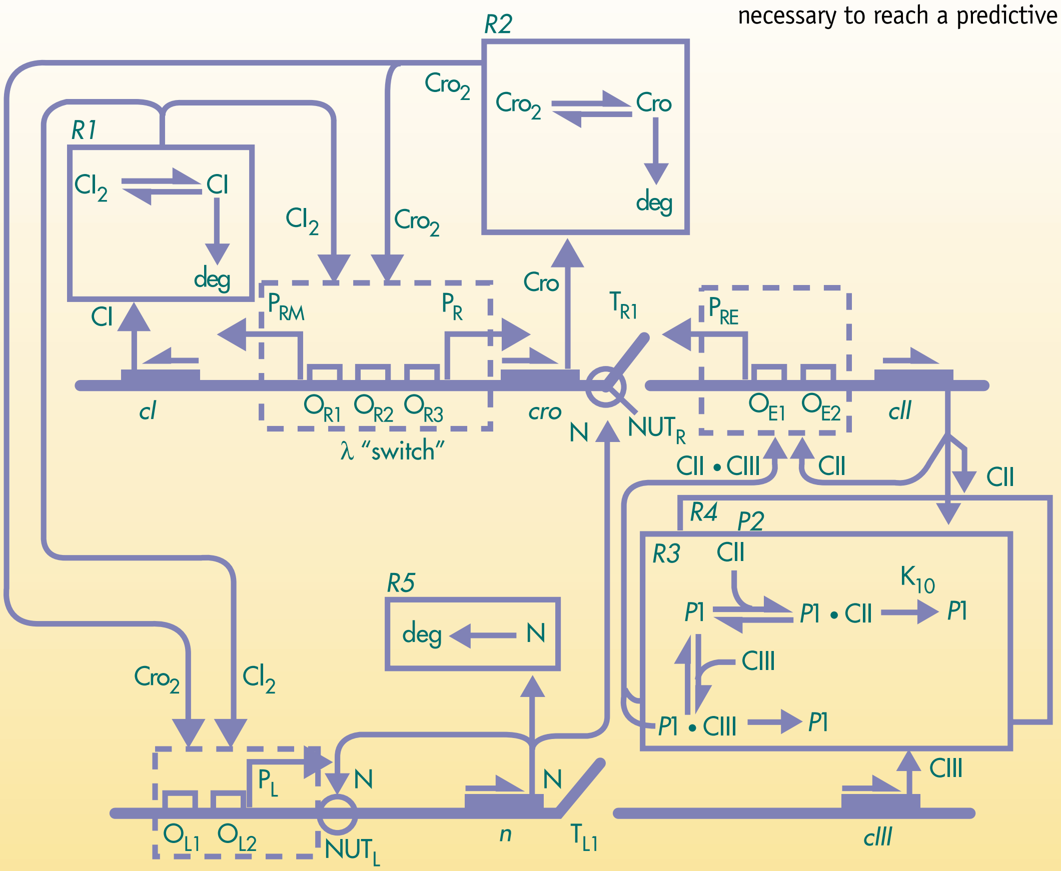
The prospect for pathway-level modeling is demonstrated by recent research using steady-state models of biological networks in whole cells and kinetic models of individual biochemical pathways. The first approach, metabolic network modeling, combines simplified models with successive constraints to identify an “envelope” of expected cell behaviors under different conditions. Such modeling depends only on the nature rather than the rates of reactants and products of metabolic transformations, and most data for building the model can be derived directly from annotated genomes (see figure below). For example, this type of model could identify which nutrients and metabolic pathways are essential under specific conditions. Metabolic network models eventually will allow scientists to infer phenotypic properties directly from functionally annotated genomes. Models can identify possible metabolic processes, but kinetic information about each pathway is necessary to simulate the cells’ dynamic behavior.



Metabolic maps provide a framework for studying the consequences of genotype changes and the relationships between genotypes and phenotypes. This metabolic network model for Escherichia coli *incorporated data on 436 metabolic intermediates undergoing 720 possible enzyme-catalyzed reactions. In this diagram, the circles contain abbreviated names of the metabolic intermediates, and the arrows represent enzymes. The very heavy lines indicate links with high metabolic fluxes. Analyses were correct 90% of the time in predicting the ability of 36 mutants with single-gene deletions to grow on different media. [J. S. Edwards and B. O. Palsson,* Proc. Nat. Acad. Sci. *97, 5528–33 (2000)]*

In the second approach, models of pathway kinetics require very fast computers and extensive empirical data, including reaction rates and substrate concentrations, to study every step of the biological system to be modeled. Kinetic models have been applied successfully to some very well characterized pathways (see figure below). Since detailed biochemical data generally are not available for pathways, however, comprehensive whole-system models will be possible only after further research has been conducted and computing power has advanced significantly.

The full promise of predictive simulations of microbial function will require a sustained partnership among experimental and computational biologists, mathematicians, and computer scientists. A number of advances are critical in data collection, data management, and modeling methods. Additionally, close collaborations between modelers and biologists are needed to collect complete and consistent experimental data sets for constructing models. The Genomes to Life program will establish such multidisciplinary partnerships and create data sets and computational methods necessary to reach a predictive understanding of microbial life.



The pathway kinetics model above depicts the mechanisms of the "decision circuit" that commits a bacterial virus [lambda (λ)] to one of two alternate pathways in its life cycle. The lytic path sets the stage for immediate replication of the virus and destruction of its* Escherichia coli *host cell, while the lysogenic path selects for the incorporation of viral DNA into the host genome, allowing the virus to remain in a dormant state.

In the diagram, bold horizontal lines indicate stretches of double-stranded DNA, arrows over genes show the transcription direction, and dashed boxes enclose operator sites that comprise a promoter control complex. The core of the decision circuit is the four-promoter, five-gene regulatory network; initiation of pathway actions involve other coupled genes not shown. Many pathogenic organisms use a similar mechanism of concentration-dependent probabilistic pathway selection to switch surface features and evade host responses.

In the model above, pathway selection at different virus concentrations, predicted using a kinetic model of the genetic regulatory circuit, is consistent with experimental observations. Developing this model required nearly 40 empirical rate constants and the use of a supercomputer. [A. Arkin, J. Ross, and H. H. McAdams,* Genetics *149, 1633-48 (1998)]

DOE brings a unique combination of capabilities and missions to the broader research landscape and is well suited to producing specific components of next-generation tools. DOE's accomplishments include the establishment of major biological databases and the development of notable expertise in DNA sequence informatics.

Specific Aims

Aim 1. Develop methods for high-throughput automated genome assembly and annotation

The first step in characterizing microbial functional diversity is a comprehensive genome-based analysis of the rapidly emerging genomic data. With changes in sequencing technology and program priorities, the acquisition rate of microbial whole-genome sequences is increasing dramatically. Assembling and interpreting such data will require new and emerging levels of coordination and collaboration in the genome research community to formulate the necessary computing algorithms, data-management approaches, and visualization systems. Moreover, the study of microbial ecosystems will require computational methods for inferring the composition, capability, and phylogenetic relationships of a heterogeneous microbial community from sampled sequence data.

Aim 2. Develop computational tools to support high-throughput experimental measurements of protein-protein interactions and protein-expression profiles

Mass spectrometry, DNA microarrays, and other technologies offer the promise of rapid and comprehensive identification of expressed proteins and protein complexes. This aim, which relies heavily on computers and algorithms to deconvolute and archive the raw data for useful querying, will require high-speed algorithms for matching mass spectrometry tags to protein databases. Databases also will be needed for storing complex data on gene expression and protein interactions.

Aim 3. Develop predictive models of microbial behavior using metabolic-network analysis and kinetic models of biochemical pathways

Such modeling has been applied to a number of well-characterized cells and pathways. The ultimate goal will be to develop methods for automatically collecting and integrating model parameters from large experimental data sets into computational models to simulate cellular capability and behavior in newly characterized microbes.

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

Computational Biology Research and Development Goals

Category	Research Goal
Sequencing Informatics	• Automated microbial genome assembly
	• Laboratory Information Management Systems (LIMS)
Sequence Annotation	• Consistent gene finding, especially for translation start
	• Identification of operon and regulon regions
	• Promoter and ribosome binding-site recognition
	• Repressor and activator-site prediction
Structural Annotation	• High-throughput automated protein-fold recognition
	• Comparative protein modeling from structure homologs
	• Modeling geometry of complexes from component proteins
Functional Annotation	• Computational support for protein identification, post-translational modification, and expression
	• Protein-function inference from sequence homology, fold type, protein interactions, and expression
	• Methods for large-scale comparison of genome sequences
	• Mass spectrometry LIMS and analysis algorithms
	• Image analysis of protein interactions and dynamics
New Databases	• Environmental microbial populations
	• Protein complexes and interactions
	• Protein expression and post-translational modification
Data Integration	• Tools interoperation and database integration
	• Tools for multigene, multigenome comparisons
	• Automated linkage of gene/protein/function catalog to phylogenetic, structural, and metabolic relationships
Microbial Ecology Support	• Statistical methods for analyzing environmental sampling
	• Sequence- and expression-data analysis from heterogeneous samples
	• Pathway inference from known pathways to new organisms and communities
Modeling and Simulations	• Molecular simulations of protein function and macromolecular interactions
	• Development of computational tools for modeling biochemical pathways and cell processes
	• Implementation of computational tools
	• Structural modeling of protein variants
	• Computational tools for modeling complex microbial communities
Visualization	• Methods for hierarchical display of biological data: (System level > Pathway > Multiprotein machines > Proteins > mRNA > Gene)
	• Displays of interspecies comparisons
	• Visualization by functional pathways (e.g., DNA repair, protein synthesis, cell-cycle control)

Aim 4. Develop and apply advanced molecular and structural modeling methods for biological systems

Some challenges to accomplishing this aim include the large size of biomolecules, the long time spans of many biological processes, and the subtle energetics and complex milieu of biochemical reactions. Chemical simulations will aid in understanding biochemical processes through elucidation of the energetic factors underlying protein-protein or protein-DNA interactions and dissection of the catalytic function of certain enzymes. Additionally, improved methods for predicting protein structure offer considerable promise for structural annotation and analysis of protein interactions.

Aim 5: Develop the groundwork for large-scale biological computing infrastructure and applications

With the continued exponential growth of biological data, the data analysis and simulation essential to achieving the long-term goals of the Genomes to Life program will require significantly greater computing power and information infrastructure than are currently available to the biological community. Consequently, another aim of the program is to begin the planning and prototyping processes to determine the New Biology’s computing and information demands and begin the planning and interagency partnerships needed to put the infrastructure into place. Experience in other major computing initiatives has shown that early planning is essential for the development and implementation of such large-scale computing resources for a scientific community.

Additionally, significant investment in the development of high-performance biological computing codes and software libraries will be needed to support a wide range of modeling and simulation tasks in Genomes to Life. These computing codes will include everything from basic bioinformatics algorithms to fundamentally new methods for simulating complex processes. The task will require a concerted strategy that complements the milestones of the Genomes to Life program’s scientific plan and computing-infrastructure development. The development of such a plan and prototypes for the computing infrastructure and related codes and libraries motivate the Genomes to Life partnership between the offices of Biological and Environmental Research and Advanced Scientific Computing Research.

Appendices

Appendix A
Technology Fundamentals

Static vs Dynamic Structures

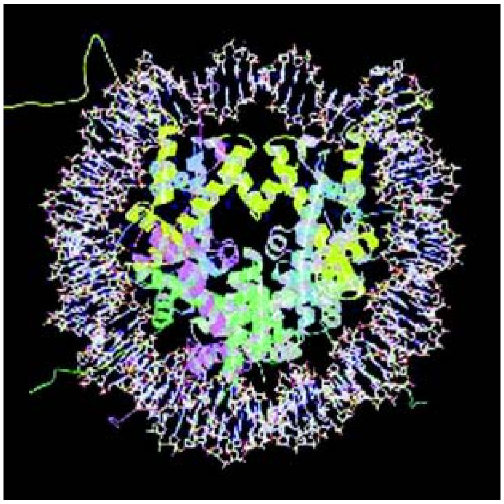
Over the years, the laborious and painstaking process of protein crystallography has resulted in an increasingly large number of protein and nucleic acid structures. Crystallography is essential for studying the 3-D structure of these biomolecules and for revealing some of the mechanisms of their biological activity. However, the structures solved by crystallography are static: they are a snapshot image of the molecule's motion and chemical activity.

In contrast with time-averaged still images from protein crystallography, other analytical tools can provide dynamic structural information, something that could be considered a motion picture of a molecule's behavior. These tools include a variety of spectroscopic techniques such as fluorescence studies, electron paramagnetic resonance, and nuclear magnetic resonance; and scattering techniques such as quasi-elastic and dynamic light scattering, and X-ray and neutron small-angle scattering. These techniques usually cannot provide the same level of molecular detail as that of crystallographic studies, but the measurable changes in structure associated with molecular activity can yield essential insights into how a molecular machine does its job. As a result of increasingly sensitive detection limits, observing molecules directly within a cell is now becoming possible using fluorescent labels, for example. Being able to observe proteins in vivo, without perturbing their natural environment, offers the ultimate means of understanding the many processes that occur within living cells.

Molecular Models from Crystallography

The most common experimental method for obtaining a detailed picture of a protein or protein complex is to interpret the diffraction of X rays from many identical molecules in an ordered array commonly referred to as a crystal. The method thus is called single-crystal protein X-ray crystallography. This experimental technique provides information on the positions of individual atoms within a biological complex. Having a detailed structure or "form" of the macromolecule will aid in beginning to understand its function.

Determining a protein's structure by X-ray crystallography consists of growing high-quality crystals of the purified biomolecule, measuring the directions and intensities of X-ray beams diffracted from the crystals, and using computers to transform the X-ray measurements. This method produces an image of the crystal's contents in much the same manner as a microscope's objective lens. Computers and crystallographer take the place of the microscope's objective lens because no lens can focus the highly divergent beams diffracted from the crystals. Finally, the image must be interpreted, which involves computer graphics to display the electron density of atoms in the molecule and the construction of a consistent molecular model.



Molecular structure of the nucleosome core complex

Recent developments in synchrotron radiation sources have revolutionized protein crystallography, opening the door to high-throughput protein-structure determination. These intense tunable X-ray sources have allowed the development of the Multi-wavelength Anomalous Dispersion (MAD) technique for solving the phasing problem. MAD enables the collection of data in mere minutes compared with the many hours required when conventional X-ray sources were used. To make effective use of synchrotron sources for protein studies, however, new approaches are needed for efficient, high-throughput production of protein crystals. In addition, improvements in detectors, data interpretation, and graphics display all will enhance the quality of the molecular model that is the end product of the macromolecular crystallography process.

Neutron crystallography is a valuable technique to use when details of an enzyme mechanism or binding site for drug development are needed. Although not a high-throughput procedure, neutron crystallography can visualize the hydrogen atoms (about half the atoms in a protein structure). It can also visualize many of the more-mobile water molecules that cannot be seen, even with ultrahigh-resolution synchrotron X-ray radiation.

The accompanying figure shows the molecular structure of the nucleosome core complex, the chromosome's basic building block. This fundamental repeating unit is made of a complex of eight separate protein molecules and two strands of DNA that carry a piece of the genetic code, the blueprint for life. This is the longest segment of DNA ever seen at near-atomic resolution.

Nuclear Magnetic Resonance Spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy uses high magnetic fields and radio-frequency pulses to manipulate the spin states of nuclei—including 1H, 13C, and 15N—that have nonzero-spin angular momentum. For a molecule containing such nuclei, the result is an NMR spectrum with peaks whose positions and intensities reflect the chemical environment and nucleic positions within the molecule. As applied to protein-structure analysis, the accuracy now achievable with NMR spectroscopy is comparable to that obtained with X-ray crystallography.

In several respects, NMR spectroscopy offers a technique complementary to X-ray crystallography and neutron diffraction. An important consideration is that NMR structures typically are obtained from proteins in solution, with no requirement that the protein be crystallizable. Not only does this lead to a protein-structure representation unconstrained by any crystal lattice, but it also allows structures to be determined for proteins that cannot be crystallized. The latter point is especially significant because a substantial fraction of all proteins are thought to contain long, disordered regions (>40 residues) that may prevent crystallization. In these cases NMR may

Environmental Molecular Sciences Laboratory's 800-mhz NMR spectrometer at Pacific Northwest National Laboratory be the best, perhaps the only, method available to characterize the structures.

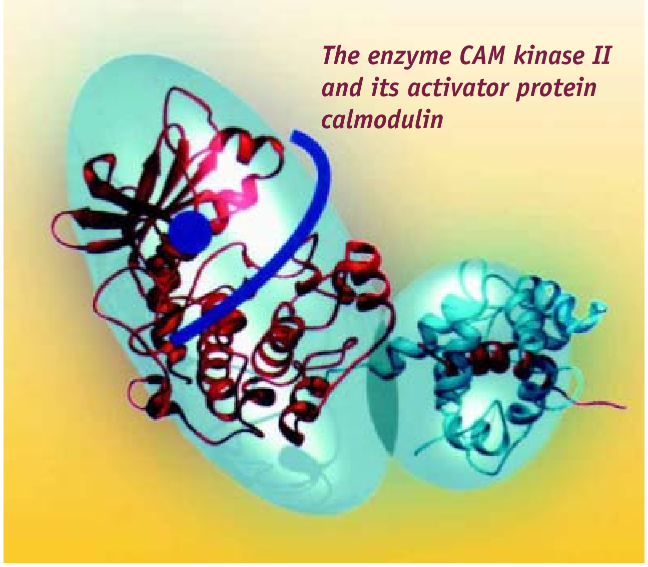
On the other hand, NMR historically has been limited in two important ways. First, protein-structure determination by NMR methods has been limited to relatively small proteins, that is, those smaller than about 40 kD. Second, data collection for a single protein structure typically has required weeks, as compared with the minutes or hours needed for X-ray crystallography. Recent advances on four fronts have significantly lessened these limitations, however. They are (1) development of instruments using higher magnetic fields and higher radio frequencies, thus improving sensitivity and resolution; (2) novel use of isotopic labeling, and (3) development of sophisticated experimental methods that can differentially manipulate nuclear spins, thus enabling studies of proteins up to 150 kD [K. Pervushin et al., *Proc. Nat. Acad. Sci.* **94**, 12366 (1997); R. Reik et al., *Proc. Nat. Acad. Sci.* **96**, 4918 (1999)]. The fourth factor is the significant advances in NMR probe technology that have improved sensitivity by three- to fourfold and substantially reduced data-collection times. Collectively, these advances are having a significant impact on experimental strategies being employed in protein studies; larger protein molecules are being analyzed, proteins with low solubility are being studied at lower concentrations (~0.2 mM), and data for well-behaved proteins of <40 kD are being collected more rapidly (in about 2 days). Other significant advances are being made in the automation of data analysis, reducing the time needed for protein-structure determination from weeks to days.

Neutron Scattering

Neutron scattering helps to resolve how the 3-D parts of protein machines fit together and how proteins communicate in dynamic regulatory and signaling networks. While nuclear magnetic resonance (NMR) spectroscopy and X-ray crystallography provide high-resolution structural information on individual subunits of protein and protein-DNA complexes, neutron scattering provides lower-resolution information on the shapes and arrangements of these subunits in solution, thereby complementing these other tools.

Although X-ray crystallography on its own has delivered a few spectacular examples of high-resolution structures of molecular machines (such as the structure of the ribosome), such structural achievements can take decades of work and require crystalline forms of the proteins studied. Further, if the protein is mechanically flexible, crystallizing the protein without distorting it can be difficult. Neutron scattering, however, analyzes protein complexes in solution, bypassing this potential complication. Together with NMR and X-ray crystallography, neutron scattering enables researchers to develop comprehensive high-resolution models of working molecular machines.

In addition to helping structural biologists fit together the interlocking pieces of complex molecu-


The enzyme CAM kinase II and its activator protein calmodulin

lar machines, state-of-the-art neutron scattering capabilities also can reveal the conformational dynamics of proteins. For example, neutron scattering has shown how the DNA in a nucleosome complex unwinds from the histone core as the solution chemistry changes. Similarly, neutron scattering can monitor conformational changes in an enzyme as a substrate binds. By combining high-resolution data on individual components (from NMR or X-ray crystallography) with neutron scattering's lower-resolution information on overall shapes and positions, scientists can view protein complexes in different functional states, along a signaling pathway, or at different stages of an activation mechanism.

The physical basis for neutron-scattering experiments lies in how neutrons interact with the atoms in a sample. Neutrons are electrically neutral particles scattered by atomic nuclei. The neutron-scattering power of different elements—and even different isotopes of an element—is a complex function of the properties of the compound nucleus that forms briefly between an incoming neutron and the atomic nucleus that scatters it. As a result, the neutron-scattering power does not vary smoothly with the atomic mass of the scattering nucleus. (In contrast, X rays are scattered by electrons, and thus the scattering power of X rays increases monotonically with the atomic number.)

The neutron-scattering signal from a biological macromolecule such as a protein or DNA molecule in solution is proportional to the "contrast" between the molecule studied and the solvent. This contrast is the difference between their neutron-scattering densities, which are calculated by summing the neutron-scattering lengths of all the atoms in the macromolecule or in the volume of solvent containing the macromolecule and then dividing by that volume.

One of the biggest differences in neutron-scattering power is between hydrogen and deuterium. In addition, hydrogen has a neutron-scattering length opposite in sign to those of most elements found in biological molecules. By manipulating the ratios of hydrogen to deuterium in the protein or DNA subunits of a sample, one can "tune" the scattering densities of different subunits relative to the solvent so that the subunits can be made to "disappear" or "appear" selectively. This contrast-variation technique enables researchers to study the shapes of individual protein or DNA subunits within large complexes and determine the relative positions of the subunits.

Neutron scattering shows much promise for structural analyses of protein complexes, but it faces several technical limitations. For example, techniques are needed for inexpensively producing relatively large quantities of soluble, deuterium-labeled samples, which requires robust protein-expression systems that give high yields in deuterated media. In addition, current applications are limited by the relatively low fluxes of neutron beams produced by reactors or accelerators. The development of more powerful neutron sources, such as the Spallation Neutron Source and the higher-intensity cold neutron source planned for the High Flux Isotope Reactor also at Oak Ridge National Laboratory, will give scientists access to state-of-the-art and eventually next-generation instruments.

Imaging Technologies

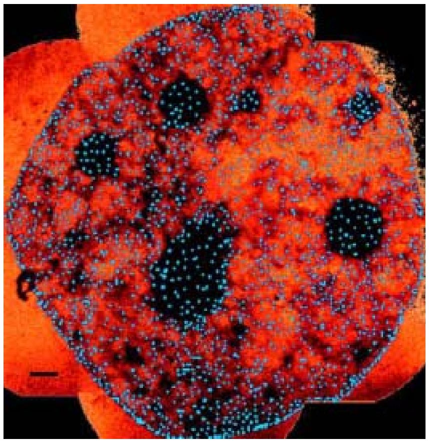
Understanding a complex living system will require a thorough comprehension of the interactions of cells and tissues in the organism. And understanding those cells will necessitate an integrated understanding of all functional units—signal transduction molecules, structural scaffolding, and genetic material. The molecular machinery of life must be studied at all size scales from atoms to complete organisms. Extensive information about the proteins that make up the cells' functional units can be obtained through the use of molecular biology, crystallography, and computational biology. But understanding their function within their natural environment—the cell—will require examining these proteins within the cell, through all phases of cell behavior.

Imaging is a very powerful unifying tool for such studies. Light-microscopic analyses of fluorescently tagged markers yield critical information about the location and behavior of proteins, as well as many details of protein-protein interactions. The information obtained, however, is limited by the light-microscope's resolution (150 to 200 nm). Better resolution can be obtained using electron microscopes, but that approach requires more elaborate cell-processing procedures and typically is limited to sectioned or very thin specimens. An array of imaging techniques will be needed if science is to understand the function of proteins in cells, the behavior of cells, and, ultimately, whole organisms.

Confocal Microscopy. At relatively low resolutions, confocal microscopy can produce three-dimensional (3-D) images of fluorescently tagged gene products to determine their distribution in the cell during different stages of the cell cycle or under various environmental conditions. Such information allows deep insights into cell and organelle biology. Furthermore, confocal microscopy permits analysis of the cell's 3-D architecture, which cannot be achieved by conventional light microscopy. The broad goal is to visualize cellular constituents and general cytoarchitecture in a state as close to native organization as possible.

The availability of laser-based confocal microscopes and the imaginative exploitation of green fluorescent protein from jellyfish have provided new tools of great diversity and usefulness. Watching a protein bind its substrate or its partners in real time with submicrometer resolution within a single cell is now possible. The importance of such processes as self-organization and the assembly of subcellular organelles is well recognized. Self-organization at the intermediate level of multimeric protein complexes is open to inspection.

X-Ray Microscopy. Soft X-ray microscopy, using X rays produced at synchrotron light sources, is an emerging biological imaging technique for the examination of intact, hydrated cells. The shorter wavelengths of X rays permit resolution 5 to 8 times better than that achieved by light microscopy, and the information obtained from the image contrast is highly quantitative in nature. Protein location can be determined at better than 50-nm resolution in whole cells, as shown in the image of the nucleus below, in which nuclear pore proteins located in the



Soft X-ray image of a human mammary cell membrane surrounding the nucleus are labeled blue. Entire cells can be examined and information about the localization of a specific protein determined without extensive processing. X-ray cryotomography facilitates 3-D reconstructions of cells—a technique comparable to CT scans of the brain and other parts of the body—and precise localization of proteins in the cells. Future developments will enable studies of specific proteins in living cells using light microscopy, followed by determination of their ultrastructural localization using soft X-ray tomographic imaging. Single-cell studies with these techniques will combine the power of protein-specific, live-cell, fluorescent light microscopy with X-ray microscopy's unique high-spatial–resolution, whole-cell imaging methods for high-throughput analyses of protein function in cells.

Electron Tomography. At still higher resolutions, electron tomography is the most widely applicable method for obtaining 3-D information. Whereas the wavelength of visible light limits the resolution of light microscopy to hundreds of nanometers, the wavelength of intermediate-voltage electrons is only a fraction of an angstrom. Electron tomographic microscopy is therefore the only method suitable for examining such structures as many supramolecular assemblies, organelles, and cells that vary in structure from one to another. The method recently has been applied to the study of cryofixed and stained sections of mitochondria and secretion organelles. Furthermore, with the development of automated low-dose data-acquisition schemes, molecules and cells embedded in vitreous ice can be studied. This opens new horizons for investigating the functional organization of cellular components with minimal perturbation of the cellular context. The unmatched spatial resolution provided by electron microscopy complements the temporal resolution provided by light-microscopic techniques, which allow movements of molecules to be tracked in vivo.

In contrast to resin embedding and sectioning, preparation in vitreous ice yields a preserved state closer to the native state of biological samples. The sample is spread into a film only a few hundred nanometers thick across holes in a perforated carbon support and plunged into liquid ethane, thereby forming vitreous ice throughout the thin specimen layer [J. Dubochet et al., *Quart. Rev. Biophys.* **21**, 129 (1988)]. This procedure almost instantaneously immobilizes all cellular components in their native, frozen-hydrated environment and allows them to remain stable in the high vacuum of the electron microscope. In principle, image interpretation is also more straightforward because in stained samples—at least on a molecular scale—the relationship of stain distribution to underlying biological structure is not clear. Specimen contrast in vitreous ice is much lower, however, and the total electron dose is limited roughly to 2000 to 6000 e^-/nm^2 because of the samples' radiation sensitivity. Radiation damage results in structural changes at the molecular level and, at higher doses, in bubble formation. Using a computer to optimize focusing and tracking of the sample during the tilt series, however, has enabled the study of virus particles, lipid vesicles with or without cargo, and macromolecules without significant damage.

Theoretical analysis suggests that electron tomographic reconstructions to about 2-nm resolution should be possible [J. Bohm et al., *Proc. Nat. Acad. Sci.* **97**, 14245 (2000); R. Grimm et al., *Biophys. J.* **74**, 1031 (1998)]; this would allow the identity, location, and even conformation of many proteins to be seen. In the best published example, however, resolution was estimated at about 6 nm, sufficient to locate and identify only the largest multiprotein complexes [W. Baumeister et al., *Trends Cell Biol.* **9**, 81 (1999)]. Even better reconstructions, perhaps as good as twice the theoretical limit, are widely anticipated for the newest generation of electron microscopes operating at 300 kV and equipped with field-emission electron sources, liquid helium–cooled specimen stages, and energy filters [A. J. Koster et al., *J. Struct. Biol.* **120**, 276 (1997)].

A serious problem with specimens thick enough to provide useful 3-D information is that many electrons lose energy in passing through the specimen. Energy filtering improves image contrast by selecting electrons within a narrow energy window, thereby minimizing loss of resolution and contrast due to chromatic aberration. When ice-embedded specimens are investigated, the best contrast and resolution generally are achieved by selecting electrons that have lost no energy. The long-term goal of these studies is to detect macromolecular structures in their native environment, thus providing insights into their cellular function. Whole-cell tomography using the next generation of intermediate-voltage, field-emission–gun, energy-filtered electron microscopes will provide detailed 3-D information about the distribution of gene products tagged with labels absorbing at specific electron energies.

Magnetic Resonance Microscopy. Confocal or optical microscopy (OM) and magnetic resonance microscopy (MRM) have developed as important tools for cellular research. MRM is noninvasive and nondestructive, and OM requires only the expression or uptake of fluorescently labeled molecules for detection. Both methods have their advantages and disadvantages. MRM provides access to several observable quantities that cannot be determined with OM alone (e.g., metabolite concentrations, chemical shifts, spin couplings, T1 and T2 relaxation times, and diffusion constants). These quantities have been related to a variety of such cellular events as tumor formation, programmed death (apoptosis), necrosis, and increased proliferation. Instruments combining OM and MRM allow live cells to be studied simultaneously using both techniques, providing a necessary link between cellular response and molecular information on proteins and other biochemicals involved in a certain cellular event. Two combined OM-MRM microscopes are under development at Pacific Northwest National Laboratory.

Mass Spectrometry

Over the past decade, mass spectrometry (MS) has become an important tool for the analysis of proteins. In its simplest form, MS sorts and measures the mass of individual ions (charged molecules). Ions can be formed from proteins using either electrospray (ES) or matrix-assisted laser desorption ionization (MALDI), each of which typically adds one proton (in the case of MALDI) or many protons (ES) to the protein. These positively charged protein ions can be analyzed directly to establish the protein's mass. Alternatively, the ions can be fragmented while inside the mass spectrometer by techniques such as collision-induced dissociation to provide more detailed information on the protein. Of particular relevance to the Genomes To Life program is the unique ability of MS to identify a protein unambiguously, establishing the amino acid sequence (the order in which these building blocks of proteins are arranged) and determining the presence of post-translational modifications that can impact the protein's function.

Scaling up MS from one-protein-at-a-time analysis to high-throughput proteome-wide analysis is of high importance to the Genomes To Life program. Currently, the majority of MS-based techniques for proteome analysis are linked with two-dimensional (2-D) electrophoresis. These 2-D separations typically combine isoelectric focusing and sodium dodecyl sulfate polyacrylamide gel electrophoresis to separate the many proteins found in a cell. This approach offers two advantages. First, 2-D gel technology allows the visualization of a large number of proteins simultaneously. And second, comparing a 2-D gel from one organism (or cell) with that from another organism allows differences in expressed proteins to be observed clearly. A wealth of information regarding the isolated protein then can be obtained by excising individual spots, digesting proteins, and analyzing the resulting fragments by MS, typically using MALDI and a time-of-flight mass analyzer. The resulting MS data can be correlated with protein, genome, or expressed sequence tag databases. This technique will become increasingly popular as more and more genomic sequence data become available.

A major drawback of the 2-D gel approach, however, is that the entire process is quite labor- and time-intensive. The gel work alone can require hours to days of effort. Each protein has to be removed from the gel and prepared individually for MS analysis. Efforts to automate many of the spot-excision, digestion, and sample-preparation steps involved in 2-D gel and MS assays are being pursued. Other drawbacks include the limited dynamic range, quantification capabilities, and usefulness in analyzying hydrophobic proteins.

Mass spectrometer in Environmental Molecular Sciences Laboratory at Pacific Northwest National Laboratory

Recently, alternative techniques for analyzing entire proteomes have shown great promise by addressing many shortcomings of 2-D gel electrophoretic strategies. One such technique is isotope-coded affinity-tag peptide labeling. In this technique, quantitative differences can be measured readily in the levels of proteins expressed in a reference organism and in one grown with an isotope label (i.e., 180 or 15N). The isotopically labeled peptides are separated by liquid chromatography (LC) and then analyzed directly online by ES MS, providing full identification of the proteins. This technique, which has the advantage of eliminating the labor-intensive 2-D gel electrophoresis step, is potentially scalable to high-throughput techniques and is amenable to the analysis of minor expression products in complex mixtures ("large dynamic range").

Another MS-based approach that shows great promise is multidimensional LC coupled with MS. By use of 2-D capillary LC columns composed of a strong cation exchanger combined with a reverse-phase resin, samples of proteins in complex mixtures can be introduced directly into the MS using ES and the MS data can be produced in a fully automated fashion. This flowing 2-D LC approach offers substantial advantages in analysis time over conventional 2-D gel electrophoresis, and it has demonstrated a very wide dynamic range. Another key feature of this emerging technique is that it overcomes the protein-solubility problem often encountered in 2-D gel assays by including a proteolytic digestion step on the entire protein extract before sample loading.

Other MS-based approaches include the use of accurate mass tags derived from high-resolution capillary LC separation, combined with ES ionization and a Fourier transform ion cyclotron resonance (FTICR) MS. FTICR is a specialized MS that has capabilities for both high and accurate mass resolution (about one part per billion). FTICR allows a significant fraction of peptides obtained from a crude protein extract's complete proteolytic digest to be distinguished uniquely by mass alone. The technique thus can identify individual protein components from a complex mixture.

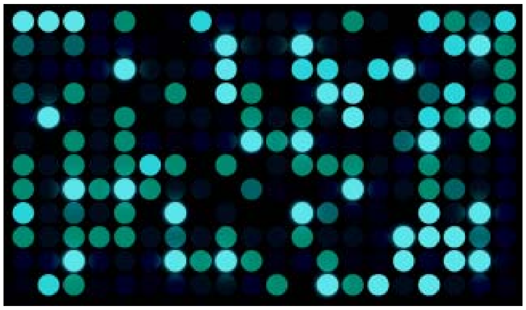
Still other promising new high-sensitivity techniques on the horizon require minimal sample preparation and therefore have potential for high-throughput proteome analysis. Microscale ("lab-on-a-chip") sample-preparation and separation technologies are being examined for online analysis of protein mixtures with MS to minimize sample-handling steps and realize high-throughput protein analysis. This high-sensitivity technique has the additional advantage of requiring minimum quantities of the sample and expensive reagents.

In another approach, the sample-separation steps are virtually eliminated when ion-ion recombination techniques are employed in a quadrupole ion trap (QIT) MS. In this technique, complete cell extracts are introduced into MS, and reactions within the QIT's trapped ion cell are used to simplify the mixture without physically isolating individual components. The resulting spectra are interpreted with the aid of computational techniques.

Although not as mature as those for proteomics, MS-based techniques also are being developed for analyzing protein complexes to rapidly assess the effects of minor protein modifications.

Microarray Technologies: DNA, Proteins, and Beyond

Historically, biochemical assays have focused on analyzing one reaction at a time. In the era of the New Biology, however, higher-throughput techniques are needed to make genomic-scale analyses practical. Microarrays are a promising technique that allows for massively parallel analyses by densely arranging miniscule samples on a glass chip or other solid surface. Most current applications involve analyzing samples with labeled probes and reading the results with a computerized image-analysis system, although mass spectrometry and other "label-less" detection techniques are becoming increasingly available.



Microarrays for simultaneous analyses of tens of thousands of samples at Oak Ridge National Laboratory

Much of the promise of microarrays lies in their small dimensions, which reduce sample and reagent requirements (samples are typically in the submicroliter range) and reaction times, while increasing the amount of data available from a single assay. In addition, through the use of different labels such as multicolor fluorescent tags, multiple tests can be conducted on the same array. The efficiency of microarrays is appealing, but their parallelism offers perhaps the most important benefit; microarrays enable many samples to be analyzed simultaneously, so standardizing data from multiple separate experiments is unnecessary and truly meaningful comparisons can be made.

Microarrays yield information on complex metabolic pathways, detailed genotypes, and the functional context of genes. One well-established use of DNA microarrays is to create transcription profiles, a measure of gene expression. Each microarray consists of a pattern of different known DNA sequences that is “probed” with fluorescently labeled mRNAs extracted from their cDNA complements or from cells. The mRNAs from expressed genes hybridize with the immobilized DNA on the chip. The fluorescence intensities reflect the amount of bound mRNA, which is in turn a relative measure of gene expression. Profiles can be generated readily to determine baseline expression levels, compare expression in cells under different conditions, and compare expression from different genotypes. DNA microarrays also are being used for such large-scale DNA sequence studies as genotyping single nucleotide polymorphisms and for investigating DNA-protein interactions.

Building on the success of DNA microarrays, protein arrays are being developed for high-speed assays of protein function, including protein-protein interactions and ligand-receptor interactions. Similarly, they can be used to screen for antibodies to use as reagents, or antibodies can be arrayed to simultaneously determine the presence and concentration of multiple analytes.

Another emerging technology involves using flow cytometry to analyze suspension arrays of fluorescent microspheres. In suspension arrays, individual array elements are defined by microspheres bearing different amounts of two or more fluorescent dyes, rather than a physical position on a flat surface. Not only can suspension arrays be handled like any other liquid, manually or in an automated system, they can be analyzed rapidly with a flow cytometer.

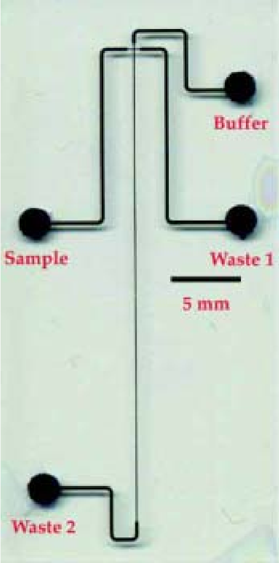
The success of microarray technologies has driven the development of new instruments and techniques for creating small but incredibly dense arrays; most arrays currently are prepared by robotic application of previously prepared samples or by light-directed in situ syntheses. Microarray technologies also have led to improved methods for rapidly reading and integrating the results of large-scale assays, although further advances are needed to keep pace with the massive amounts of data becoming available. Continued development of these tools will result in even greater miniaturization, sensitivity, and automation in the future.

“Lab-on-a-Chip” Microfluidics

A promising analytical tool for analyzing proteins and protein complexes in the biology laboratory of the future is a microfluidic device commonly called a “Lab-on-a-Chip.” These “laboratories” are fabricated using photolithographic processes developed in the microelectronics industry to create circuits of tiny chambers and channels in a quartz, silica, or glass chip. They direct the flow of liquid chemical reagents just as semiconductors direct the flow of electrons. These reagents can be diluted, mixed, reacted with other reagents, or separated by capillary electrophoresis or electrochromatography—all on a single chip.

These microfluidic circuits can be designed to accommodate virtually any analytic biochemical process. For example, a lab-on-a-chip for immunological assays probably would integrate sample input, dilution, reaction, and separation, whereas one designed to map restriction enzyme fragments might have an enzymatic digestion chamber followed by a relatively long separation column. Many features of these labs-on-a-chip make them well suited for high-throughput analyses. Their small dimensions reduce both processing times and the amount of reagents necessary for an assay, substantially reducing costs. Just as microelectronic devices can be manufactured with many elements on a single chip, microfluidic devices can be fabricated with many channels, allowing for massively parallel chemical analyses at a reasonable cost. They are uniquely suited to small-scale analyses; sample volumes for a single experiment often are in the nano- to picoliter range, opening the door to the possibility of analyzing components from single cells.

Relatively simple labs-on-a-chip already are being used for some nucleic acid and protein analyses, but microfluidics technology may someday allow millions of automated biochemical experiments to be performed per day using miniscule quantities of reagents. Eventually, individual analyses may be replaced by protocols in which tens to thousands of analytical measurements are made in parallel, either on the same or multiple samples.



Microfabricated electrophoresis device at Oak Ridge National Laboratory. This "Lab-on-a Chip" electrophoresis device allows mixtures of DNA or proteins to be separated at 1% of the time required by conventional capillary electrophoresis while using much less sample.

Phage Display

Antibodies are well recognized as indispensable tools for recognizing and tracking target molecules. However, traditional methods for preparing antibodies are cumbersome and labor intensive. As a result, researchers are working to develop faster and easier ways to capitalize on the target-recognition qualities of antibodies. Phage display is a new method that enables researchers to quickly evaluate a huge range of potentially useful antibodies and then produce large quantities of the selected ones.

Phage display uses bacteria and bacterial viruses known as phage to produce and select synthetic antibodies that have all the target-recognition qualities of natural antibodies. In fact, these synthetic antibodies are produced using the same genes that code for the target-recognition or variable region in natural antibodies from mammalian systems. The phage are genetically engineered so that a particular antibody is fused to a protein on the phage's coat and the gene encoding the displayed antibody is contained inside the phage particle. This technology thus couples the displayed antibody's phenotype to its genotype, allowing the DNA that codes for the selected antibody to be retrieved easily for future use. Collections of these antibody-covered phage are called a library. Phage libraries each typically contain a billion different antibodies, a number comparable to that in human immune systems.

To select the phage with the desired antibody from a library, the phage are allowed to bind to the target molecule, which is attached to a solid surface. The phage with antibodies that recognize the target molecule bind tightly, and the remaining (unbinding) phage are simply washed away. (Phage display even permits researchers to select antibodies with different binding characteristics for a given target.) The DNA contained within the desired phage then can be used to produce more of the selected antibody for use in research or medical diagnostics.

Appendix B
DOE Partners in Genomes to Life Program

Office of Advanced Scientific Computing Research

The primary mission of the Office of Advanced Scientific Computing Research (ASCR) program is to discover, develop, and deploy computational and networking tools that enable researchers to analyze, model, simulate, and predict complex phenomena important to the U.S. Department of Energy (DOE). This mission is carried out by the Mathematical, Information, and Computational Sciences (MICS) Division. To accomplish its mission, ASCR fosters and supports fundamental research in advanced scientific computing—applied mathematics, computer science, and networking—and operates supercomputer, networking, and related facilities.

ASCR supports the Office of Science Strategic Plan's goal to provide extraordinary tools for extraordinary science and to build a research foundation in support of the plan's other goals. In the course of accomplishing this aim, ASCR research programs have played a critical role in the evolution of high-performance computing and networks.

High-performance computing resources are expected to support the objectives of Genomes to Life through the National Energy Research Scientific Computing Center (NERSC). Platforms at NERSC include an IBM SP RS/6000 system called Gseaborg, a 512-processor IBM RS/6000 SP system with a peak performance of 410 gigaflop/s, 256 gigabytes of memory, and 10 terabytes of disk storage (hpcf.nersc.gov/computers/SP).

ASCR also provides support for Advanced Computing Research Test Beds (ACRTs). The primary objective of an ACRT is to assess the potential of new computing technologies for advancing scientific applications critical to Office of Science missions. The assessment process can, at times, include making the platforms available for specialty computing applications. Some examples of ASCR accomplishments follow.

Information on ACRTs

- www.csm.ornl.gov/ccs/Falcon.html
- www-unix.mcs.anl.gov/chibaindex.html
- www.acl.lanl.gov/news/releases/99-001.html

- Established First National Supercomputer Center. In 1974, DOE established the National Magnetic Fusion Energy Computing Center (predecessor to NERSC) and pioneered the concept of remote, interactive access to supercomputers. Before that time, scientists had to travel to the supercomputer, submit jobs, and wait for hours or days to see the output. The MICS subprogram developed the first interactive operating system for supercomputers, the Cray Time-Sharing System (CTSS), as well as a nationwide network to allow effective computer access to remote users. This revolutionary operating system also enabled users to monitor their jobs as they executed. When the National Science Foundation (NSF) initiated its Supercomputer Centers program in the 1970s, the CTSS operating system was adopted by the San Diego Supercomputing Center and the National Center for Supercomputing Applications to enable access to NSF's first Cray machines.

- Developed High-Speed Interconnects for Supercomputers. To provide a standard interface between supercomputers and such other devices as disk arrays, archival tape systems, and visualization computers, DOE laboratories developed a high-performance network interface. They also led a consortium of vendors to make it the industry standard for the highest bandwidth interconnects between computers and peripheral devices. This advance required the solution of many problems in high-speed signaling, data parallelism, and high-speed protocol design.

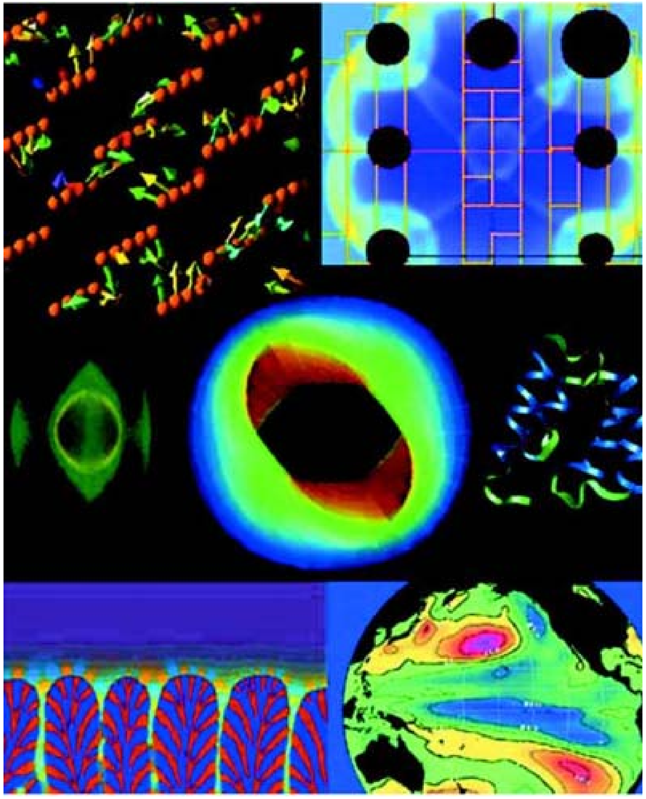
ASCR's mission and accomplishments

- www.sc.doe.gov/production/octr

Scientific Discovery Through Advanced Computing

- www.sc.doe.gov/images/news_photos SDAC_Overview_000330.pdf

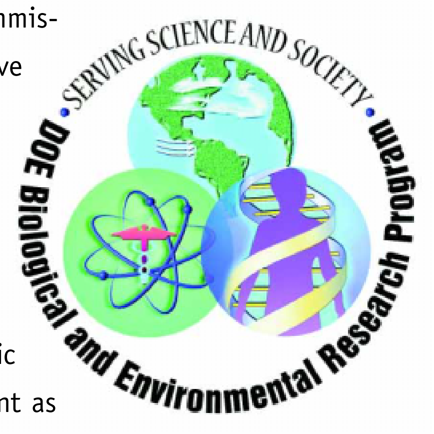
- Installed New Test Bed for Open-Source Software. A 512-CPU Linux cluster has been installed at Argonne National Laboratory's Mathematics and Computer Science Division. The cluster provides a flexible development environment for scalable open-source software in four key categories: cluster management, high-performance systems software (file systems, schedulers, and libraries), scientific visualization, and distributed computing. Its modular design makes the cluster easily reconfigurable for systems-management experiments, and its availability for testing open-source code and algorithms ensures broad use by researchers both within the laboratory and externally.



A display of NERSC-enabled research accomplishments (www.nersc.gov)

Note: Archival document. Some aspects of the Genomes to Life program have changed since publication of this April 2001 roadmap. For a current overview of the program see "Genomes to Life: Realizing the Potential of the Genome Revolution."

Office of Biological and Environmental Research

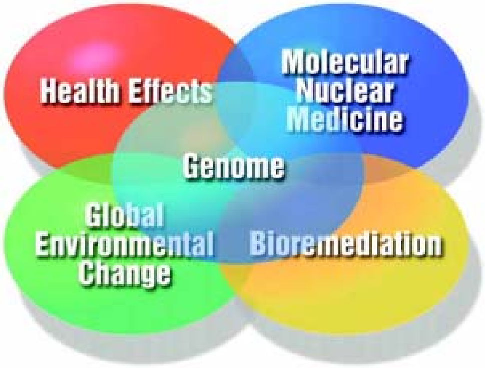

SERVING SCIENCE AND SOCIETY
DOE Biological and Environmental Research Program

For over half a century since the establishment of the Atomic Energy Commission, the U.S. Department of Energy and its predecessor agencies have pursued biological and environmental research with an unwavering commitment to understand the health and environmental consequences of energy technologies and byproducts. To address these goals, the Office of Biological and Environmental Research (BER) relies on investigators supported in all three components of the nation's research community: multidisciplinary national laboratories, the academic community, and the private sector. Scientific diversity, always a hallmark of BER programs, has become even more important as science advances at the interfaces of such disciplines as biology and computational science.

For more details on the history of BER, see the booklet, A Vital Legacy (www.ornl.gov/hgmis/publicat/miscpubs/ober-lay.pdf).

National User Facilities

As a further service to our nation's biologists and environmental scientists, the Office of Science makes advanced instrumentation and other specialized resources available through its National User Facilities supported by the Office of Basic Energy Sciences and BER. Access to these facilities, listed below, enables the broader scientific community to increase the understanding of relationships between biological structure and function, study disease pathways, develop new pharmaceuticals, and conduct basic research in molecular biology and environmental processes.



BER programs cross traditional research boundaries to seek revolutionary solutions to energy-related biological and environmental challenges.

- **Advanced Light Source**
  www-als.lbl.gov/index.html
- **Structural Biology Center at the Advanced Photon Source**
  www.sbc.anl.gov
- **Environmental Molecular Sciences Laboratory**
  www.emsl.pnl.gov:2080
- **High Flux Isotope Reactor Facility**
  www.ornl.gov/hfir/hfirhome.html
- **Joint Genome Institute**
  www.jgi.doe.gov/tempweb
- **Los Alamos Neutron Science Center**
  lansce.lanl.gov/index_ext.htm
- **Mouse Genetics Research Facility**
  www.bio.ornl.gov/htpages/mgd/mouse_fac.htmlx
- **National Synchrotron Light Source**
  nslsweb.nsls.bnl.gov/nsls/Default.htm
- **Stanford Synchrotron Radiation Laboratory**
  www-ssrl.slac.stanford.edu/welcome.html

Appendix C
Web Sites of Research Programs and Resources Complementary to Genomes to Life

Web Sites of Research Programs and Resources Complementary to Genomes to Life

U.S. DEPARTMENT OF ENERGY

OFFICE OF SCIENCE

- www.science.doe.gov

Funding

- www.science.doe.gov/production/grants/grants.html

Office of Advanced Scientific Computing Research

- www.sc.doe.gov/production/octr

Office of Biological and Environmental Research (OBER)

- www.science.doe.gov/ober/ober_top.html

Carbon Sequestration Research

- cdiac2.esd.ornl.gov/

Environmental Molecular Science Laboratory

- www.emsl.pnl.gov:2080

Genome Programs

Human Genome Program

- www.ornl.gov/hgmis

Microbial Genome Program

- www.science.doe.gov/ober/microbial.html

Joint Genome Institute

- www.jgi.doe.gov

Microbial Cell Project*

- microbialcellproject.org

Ethical, Legal, and Social Issues

- www.ornl.gov/hgmis/elsi/elsi.html

Global Climate Change

- www.science.doe.gov/ober/esdrestopic.html

Low Dose Radiation Research Program

- www.lowdose.org

Natural and Accelerated Bioremediation Research (NABIR) Program

- www.lbl.gov/NABIR

Structural Biology Research Program

- www.science.doe.gov/ober/msd_struct_bio.html

OFFICE OF BASIC ENERGY SCIENCES (OBES)

- www.science.doe.gov/production/bes/bes.html
  Renewable energy, carbon sequestration, nanotechnology

OFFICE OF DEFENSE NUCLEAR NONPROLIFERATION

- www.nn.doe.gov
  Characterization and detection of potential biological warfare agents

OFFICE OF ENERGY EFFICIENCY AND RENEWABLE ENERGY

- www.eren.doe.gov/overview
  Renewable energy, hydrogen and ethanol production, organic acid synthesis, cellulose and lignin degradation

OFFICE OF ENVIRONMENTAL MANAGEMENT

- www.em.doe.gov
  Bioremediation research (organics)

OFFICE OF FOSSIL ENERGY

- www.fe.doe.gov
  Carbon sequestration

*Jointly funded by OBER and OBES; part of Genomes to Life program.

OTHER AGENCIES

ENVIRONMENTAL PROTECTION AGENCY

- www.epa.gov

FOOD AND DRUG ADMINISTRATION

- www.fda.gov

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

- www.nasa.gov

NATIONAL INSTITUTES OF HEALTH

- www.nih.gov

Human Genome Project: www.nhgri.nih.gov

Protein Structure Initiative:
www.nigms.nih.gov/funding/psi.html

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

- www.noaa.gov

NATIONAL SCIENCE FOUNDATION

- www.nsf.gov

Biocomplexity in the Environment Initiative:
www.nsf.gov/home/crssprgm/be/start.htm

U.S. DEPARTMENT OF AGRICULTURE

- www.usda.gov

PUBLICATIONS

Bringing the Genome to Life: Energy-Related Biology in the New Genomic World

- www.science.doe.gov/ober/berac/genome-to-life-rpt.html

Interagency Report on the Federal Investment in Microbial Genomics

- www.ostp.gov/html/microbial/start.htm

Microbial Genome Program Report

- www.ornl.gov/hgmis/publicat/microbial

***Science Special Human Genome Issue* (Feb. 16, 2001)**

- www.sciencemag.org/content/vol291/issue5507

***Nature* and *Nature Genetics Genome Gateway* (Feb. 15, 2001)**

- www.nature.com/genomics/human

SELECTED BIOINFORMATICS SITES

Celera Genomics

- www.celera.com

Computational Biosciences, ORNL

- compbio.ornl.gov

DNA Data Bank of Japan

- www.ddbj.nig.ac.jp

European Bioinformatics Institute

- www.ebi.ac.uk

KEGG: Kyoto Encyclopedia of Genes and Genomes

- www.genome.ad.jp/kegg

National Center for Biotechnology Information

- www.ncbi.nlm.nih.gov

Protein Data Bank

- www.rcsb.org/pdb

TIGR Microbial Web Page

- www.tigr.org/tdb

WIT at Argonne National Laboratory

- wit.mcs.anl.gov/WIT2