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

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

DOEGenomesToLife.org

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).

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

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

Genomes to Life
Contents

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.

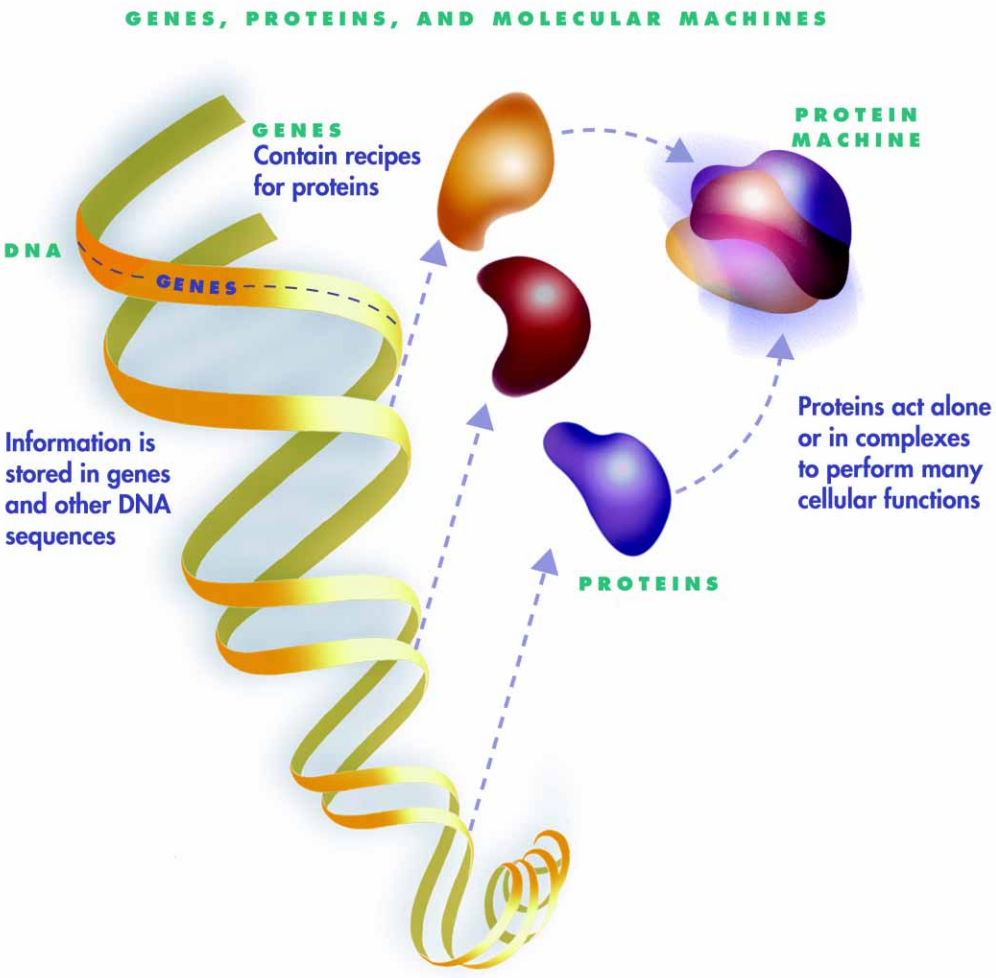
By developing a fundamental understanding of living systems, the Genomes to Life program responds directly to DOE's missions. The results will help the department develop new sources of energy, mitigate the long-term impacts of climate change, clean up the environment, reduce the threat of biological 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:



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

GENOMES to LIFE

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.

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)

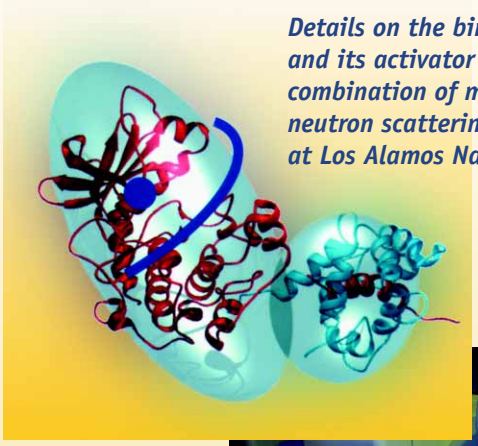
Informatics, Modeling, and Simulation

- 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

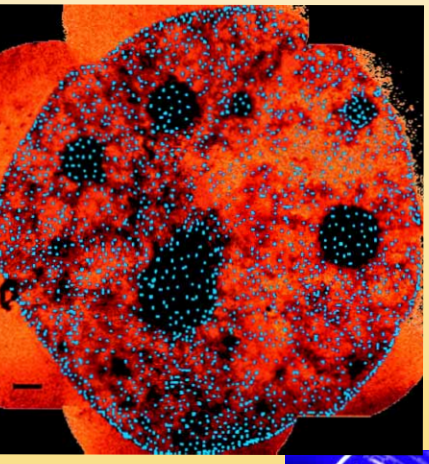
High-throughput methods, advanced computational and imaging resources, and multidisciplinary collaborations are essential elements for using the information contained in DNA sequences as a foundation for expanding knowledge of how living systems function—the focus of the Genomes to Life program. DOE 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

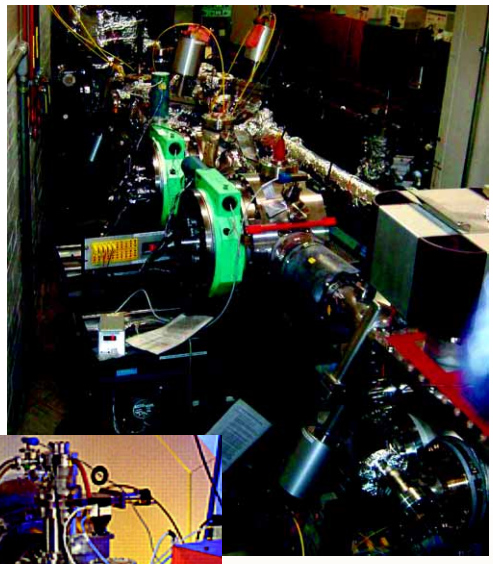


This image of a human mammary cell was produced using soft X-ray microscopy at Lawrence Berkeley National Laboratory. The blue dots label proteins of the nuclear pore complex, through which molecules enter and exit the nucleus.

IBM SP supercomputer at Oak Ridge National Laboratory

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

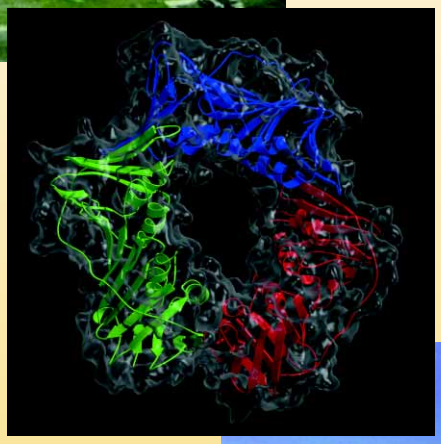
Site plan of the Spallation Neutron Source being built at Oak Ridge National Laboratory in collaboration with Argonne National Laboratory, Brookhaven National Laboratory, Lawrence Berkeley National Laboratory, and Los Alamos National Laboratory



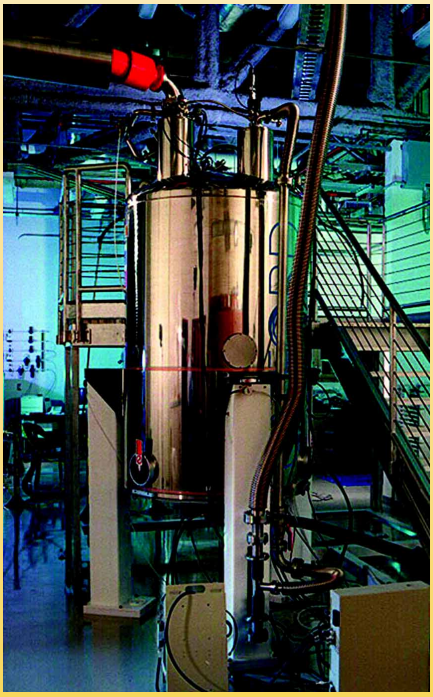
A beamline at the National Synchrotron Light Source at Brookhaven National Laboratory

Advanced Photon Source at Argonne National Laboratory

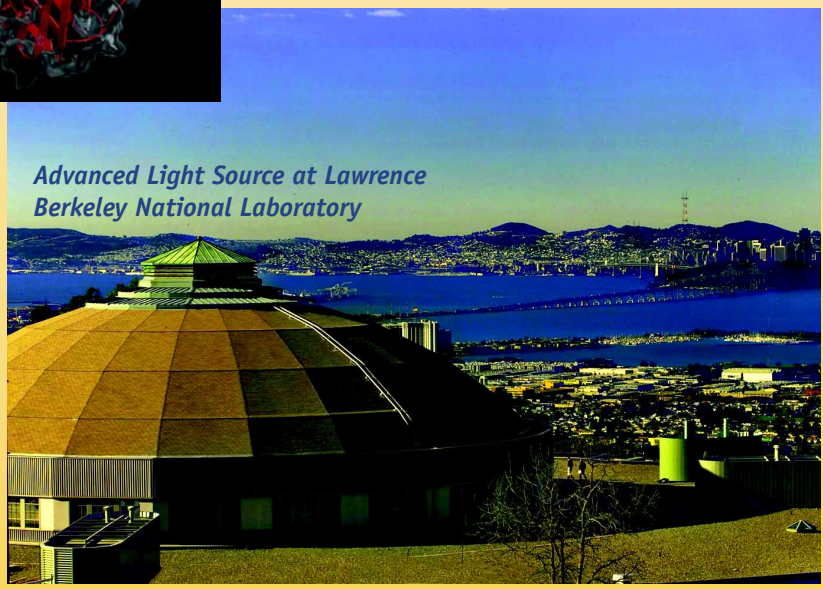
Mass spectrometer in the Environmental Molecular Sciences Laboratory at Pacific Northwest National Laboratory



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.



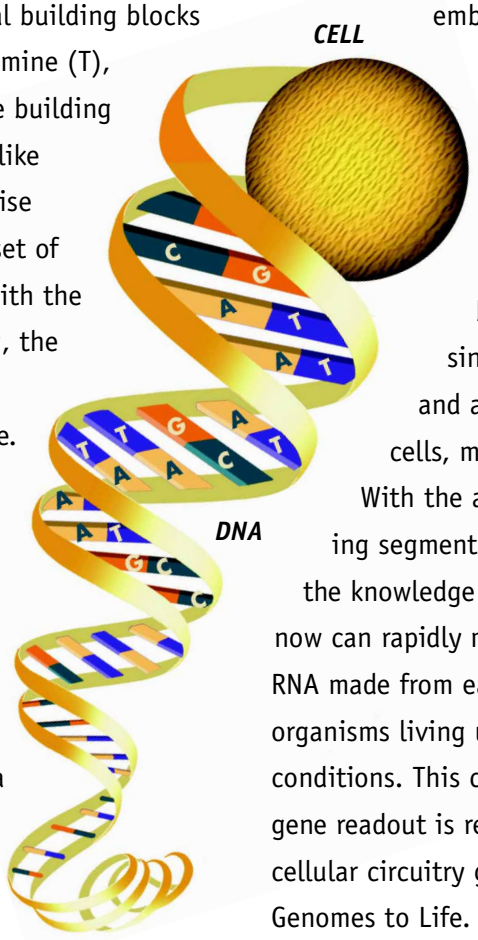
Environmental Molecular Sciences Laboratory's 800-MHz nuclear magnetic resonance spectrometer at Pacific Northwest National Laboratory



Advanced Light Source at Lawrence Berkeley National Laboratory

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.



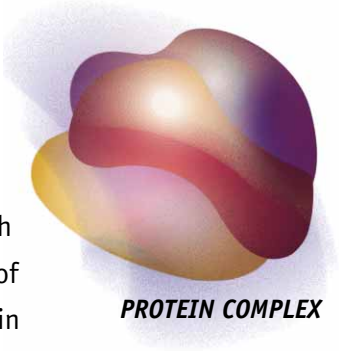
But the information stored in DNA is "lifeless" by itself, just as a recipe in a book is not a delectable dessert, nor a musical score a majestic symphonic performance. In the same way, the DNA sequence must be "expressed" to give life to a cell or organism. Furthermore, the sequence alone does not automatically 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

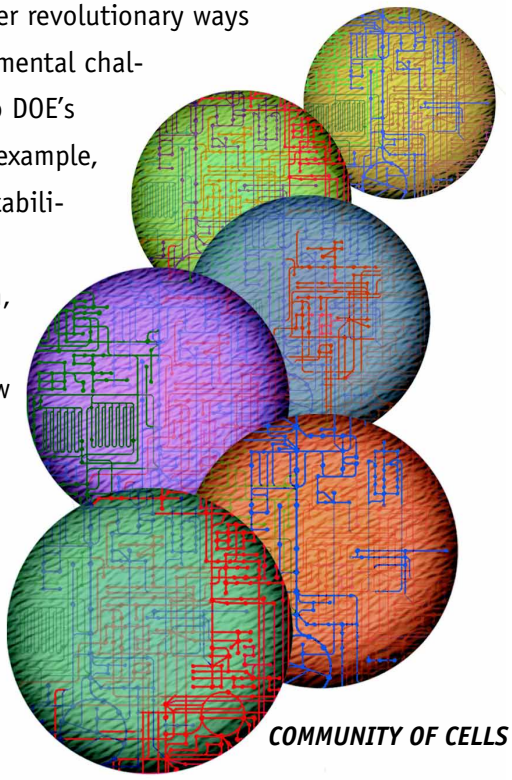
act out the instructions encoded in DNA. Although DNA and RNA are information rich, they are chemically simple and homogeneous. Proteins, by contrast, are chemically complex and diverse, properties that enable them to do so many different jobs. Proteins are "where the action is" in living systems. They 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.)

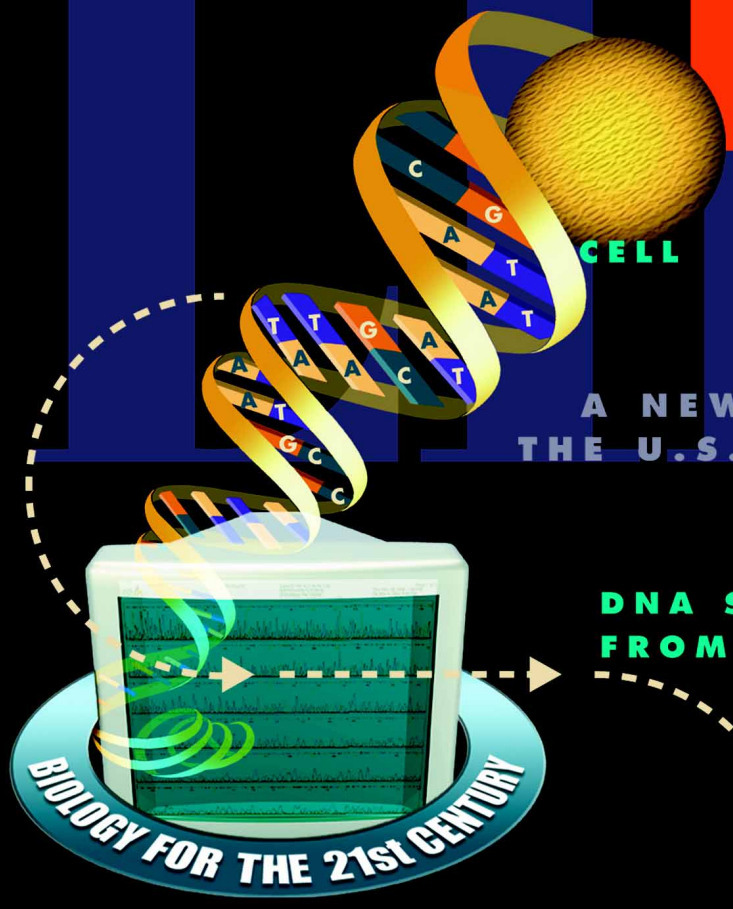


COMMUNITY OF CELLS

GENOMES *to* LIFE

ACCELERATING BIOLOGICAL DISCOVERY

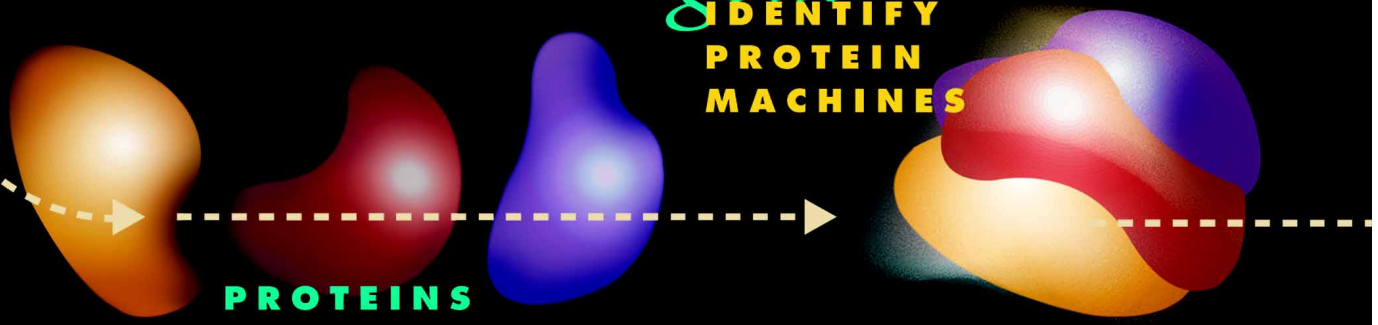
A NEW PROGRAM PROPOSED BY THE U.S. DEPARTMENT OF ENERGY



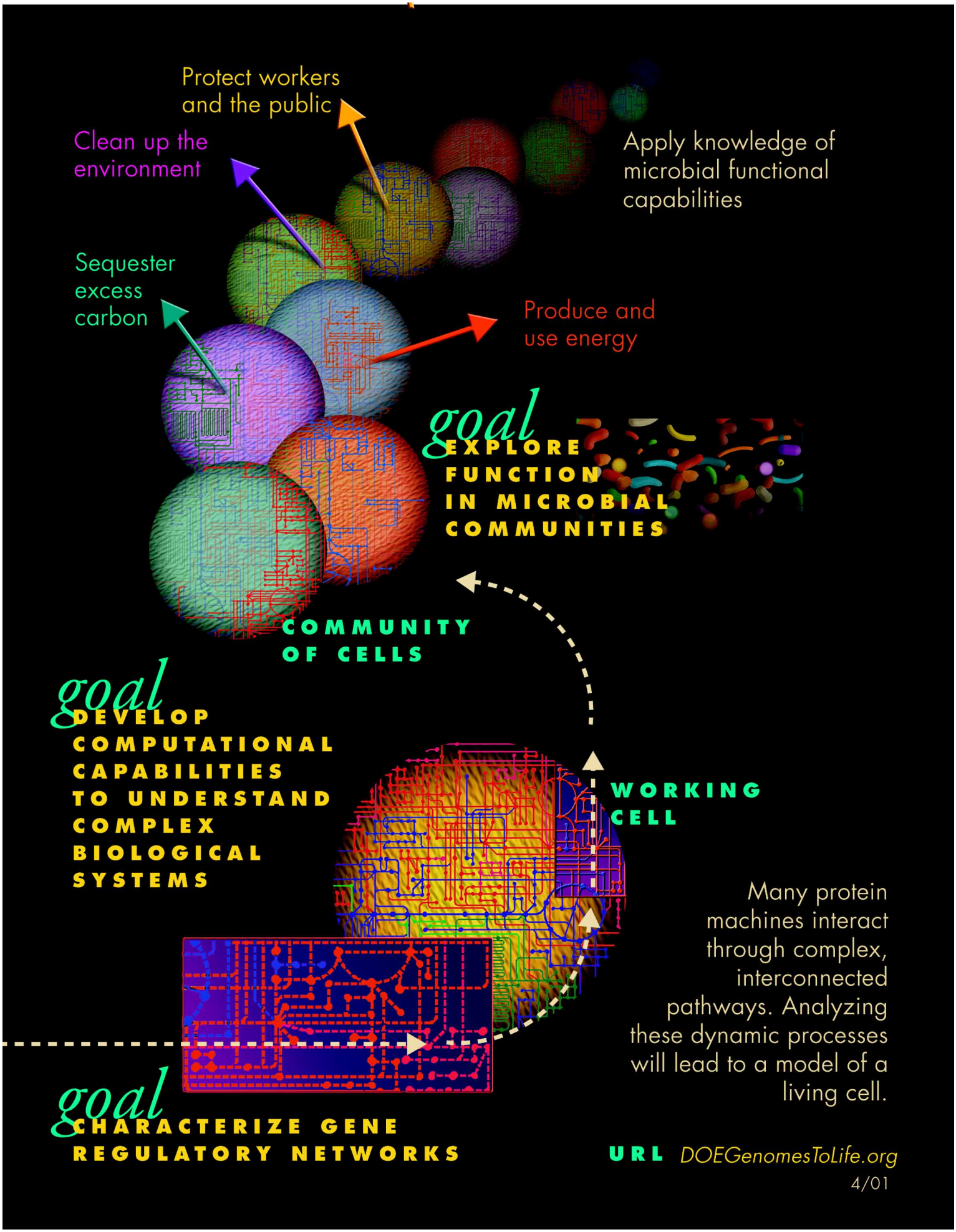
DNA SEQUENCE DATA FROM GENOME PROJECTS

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

goal
IDENTIFY PROTEIN MACHINES



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


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