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.

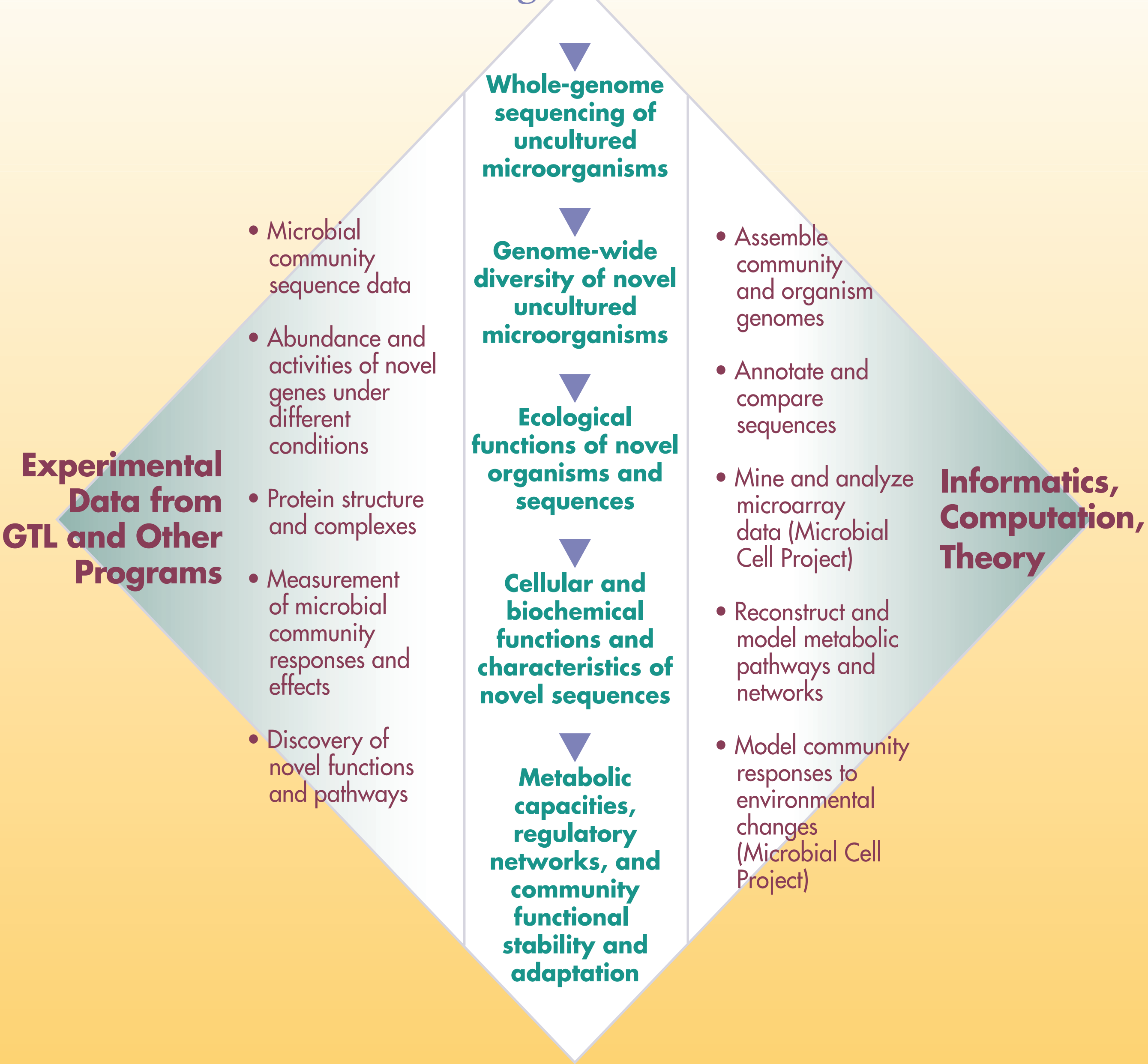
Microorganisms are the largest reservoir of genetic and biochemical diversity on earth. They have been evolving for around 3.7 billion years to colonize virtually every environment, often thriving under extremes of nutrient concentration, pH, salinity, pressure, and temperature. In the past several decades, 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

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

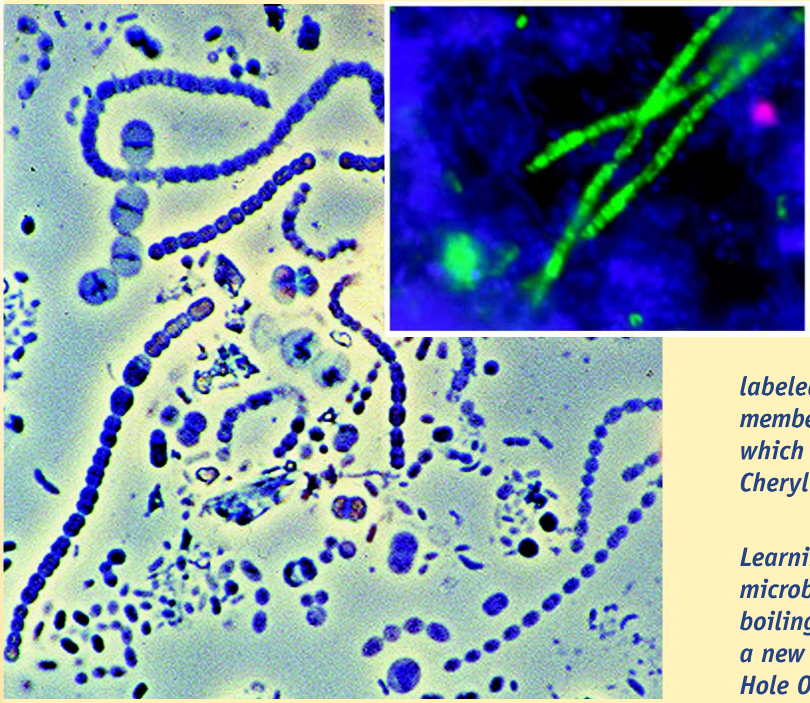
Aim 1. Determine whole-genome sequences of dominant uncultured microorganisms

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.

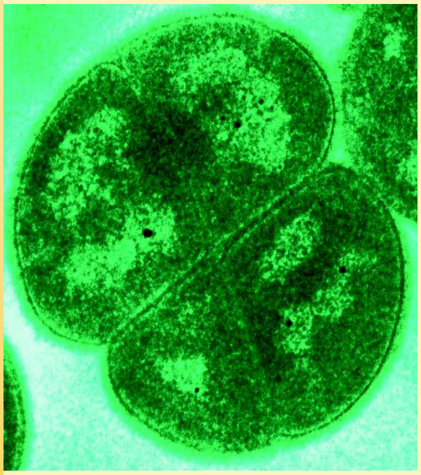
The vast majority of microbes—often thousands of species in a single environmental niche—cannot currently be grown in the laboratory, and estimates are that less than 1% have even been identified. Recent advances in molecular methods now enable an entirely different approach for tapping into the potentially 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.

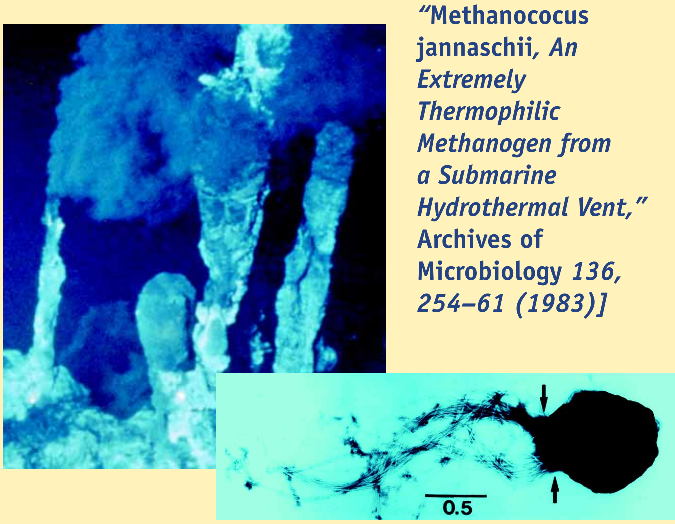
The large image at left illustrates the morphological diversity found in a natural microbial community. [Source: Frank Dazzo, Center for Microbial Ecology, Michigan State University]

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.

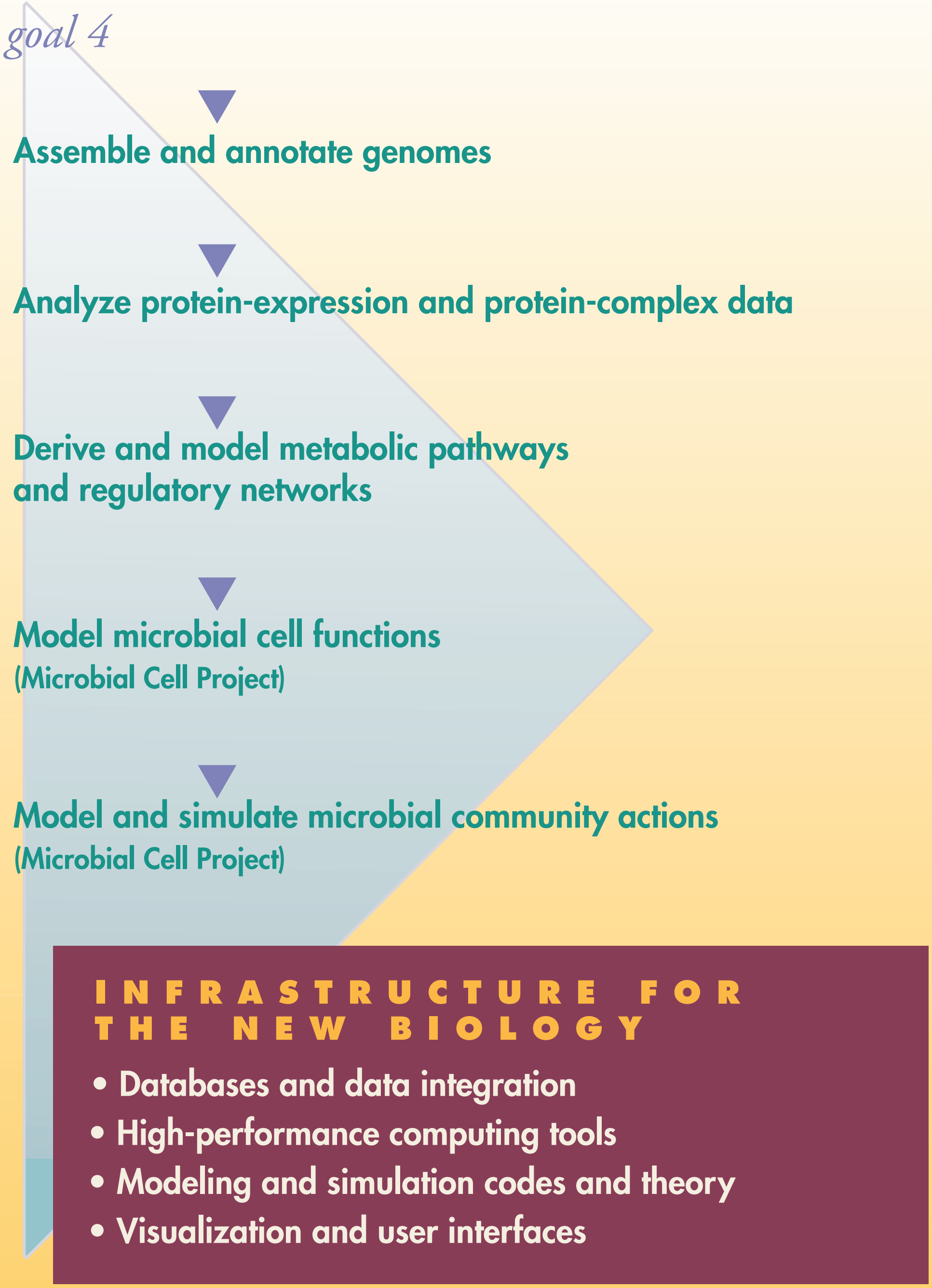
A long-term goal of the computational-modeling section of Genomes to Life is to develop the next generation of methods for simulating cellular behavior and pathways. Other goals are to create molecular-modeling and bioinformatics tools for studying multiprotein complexes, along with new computational 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)

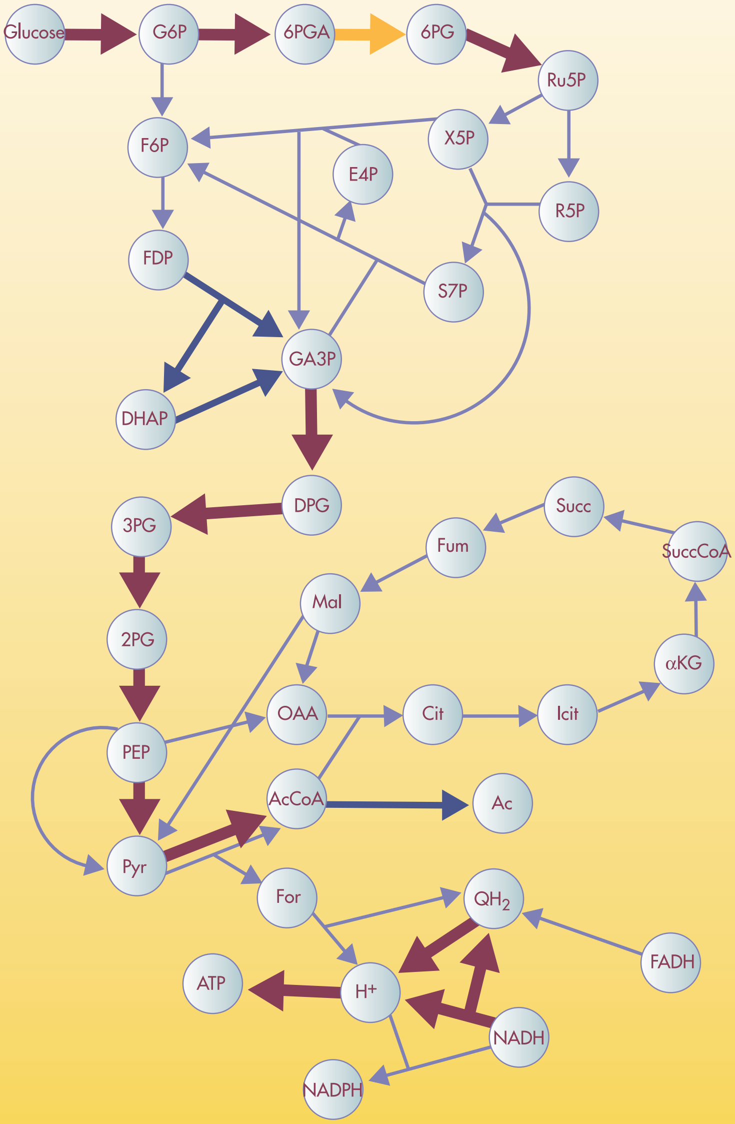
INFRASTRUCTURE FOR THE NEW BIOLOGY

- 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.

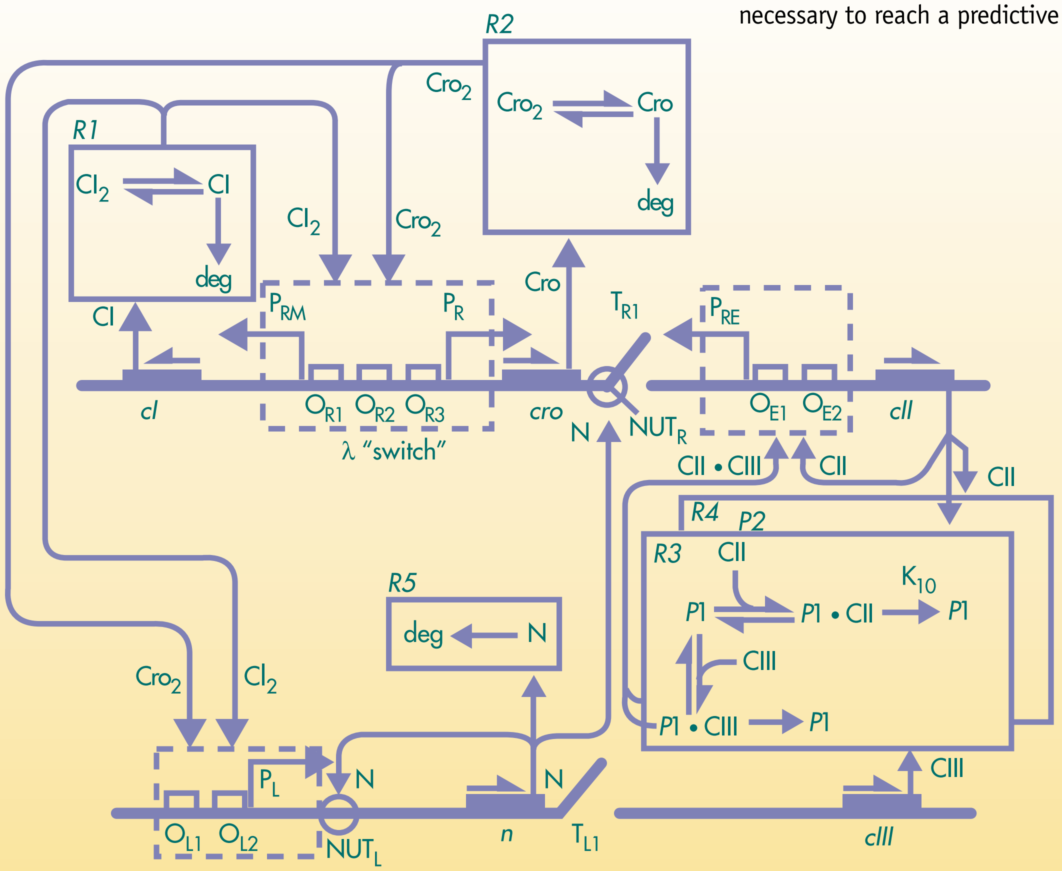
The prospect for pathway-level modeling is demonstrated by recent research using steady-state models of biological networks in whole cells and kinetic models of individual biochemical pathways. The first approach, metabolic network modeling, combines simplified models with successive constraints to identify 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.

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.