DOE/SC-0016

Microbial Genome Program

Date Published: February 2000

Prepared for the
U.S. Department of Energy
Office of Science
Office of Biological and Environmental Research
Germantown, MD 20874-1290

Prepared by the
Human Genome Management Information System
Oak Ridge National Laboratory
Oak Ridge, TN 37830
managed by
Lockheed Martin Energy Research Corporation
for the
U.S. Department of Energy
Under Contract DE-AC05-96OR22464

Foreword

These are truly extraordinary times for microbiology, and the field is enjoying a resurgence. The genomics revolution has now placed nearly 3 dozen complete microbial genomes, 11 supported by the Department of Energy's (DOE) Office of Biological and Environmental Research (OBER), into the public domain for unrestricted use by the scientific community. The number of sequenced genomes is growing rapidly. (See *www.tigr.org/tdb/mdb/mdb.html* for a current list.)

A complete genome is the ultimate "parts list" for an organism. The availability of complete genomes is now upending the traditional research approach. Previously, no alternative existed but the "reductionist" approach—to dissect an organism, its systems, and its component parts down to their simplest elements. With all the parts for a microbe now in hand (although many still have unknown functions), we now must reverse direction and do biology differently than in the past, to take a more "reconstructionist" approach. Today, we are like children with a new set of Lego blocks, dreaming of all that we can build with them if only we could understand how the pieces fit together. Obtaining the parts lists for an increasing number of microbial genomes has also launched explorations at a new level, ushering in the now flourishing field of comparative genomics.

Microbes, which make up most of the Earth's biomass, have evolved for some 3.8 billion years. They have been found in virtually every environment, surviving and thriving in extremes of heat, cold, radiation, pressure, salt, acidity, and darkness, often where no other forms of life are found and where the only nutrients come from inorganic matter. It is thought that less than 1% of all microbial species have ever even been described, since the repertoire of known species is highly dependent on the ability to culture them and most microbes are difficult to grow in the lab. (Most microbes are not responsible for diseases in humans, animals, or plants, which makes them "silent" to investigations.) The diversity and range of their environmental adaptations means that microbes have long ago "solved" many problems for which scientists are still actively seeking solutions.

To aid in carrying out its missions, DOE initiated the Microbial Genome Program in late 1994 as a spinoff of its then 8-year-old Human Genome Program. DOE missions include (besides supporting innovative high-impact and peer-reviewed science) a range of difficult challenges such as environmental waste cleanup, energy production, and biotechnology. Moreover, DOE also supports research into global climate processes, a field in which scientists are beginning to appreciate the role of microbial life. Thus the MGP not only will contribute to established DOE missions but also will generate novel insights into both the biological underpinnings of climate change and the microbial role in the overall processing of carbon and nitrogen on Earth. These capabilities can then be added to long-known uses of microbes in the brewing, baking, dairy, and other industries.

This first Microbial Genome Program report displays the significant accomplishments of this young program. While the program has supported the sequencing of 11 microbial genomes to date, with an additional 17 in various stages of progress, the real measure of this program's impact is that at least five other U.S. government agencies have more recently initiated microbial sequencing efforts. It is significant that all these agencies are establishing the firm policy that the sequence data will be made public for unrestricted use by the scientific community.

But the real impact of the Microbial Genome Program will come from the delivery of new science, new insights, and new approaches to the difficult challenges that DOE faces in carrying out its varied and demanding missions. We anticipate enormous progress as the vast repertoire of microbial genes, honed by billions of years of evolution and hitherto largely concealed from us, becomes available. An exciting future lies ahead.

Daniel Drell, Program Manager
Microbial Genome Program
Office of Biological and Environmental Research
U.S. Department of Energy
301/903-6488, *daniel.drell@science.doe.gov*

Contents

Microbial Genome Program

IMAGINE!

A future in which we can

- use "super bugs" to detect chemical contamination in soil, air, and water and clean up oil spills and chemicals in landfills;
- cook and heat with natural gas collected from a backyard septic tank or bottled at a local waste-treatment facility;
- obtain affordable alcohol-based fuels and solvents from cornstalks, wood chips, and other plant by-products; and
- produce new classes of antibiotics and process food and chemicals more efficiently.

These scenarios represent only a few of the possible ways that microbes—the invisible bacteria, archaea, protozoa, and fungi that inhabit our environment, our bodies, our food and water, and even the air we breathe—can be harnessed to serve humankind. Technological advances developed over the last decade, particularly in genetic research conducted as part of the international Human Genome Project, are enabling researchers to learn about microbes at their most fundamental level and to begin to ask questions about how the basic parts work together to form a functioning organism. The answers may challenge accepted scientific thought and offer beneficial applications in areas important to DOE's Biological and Environmental Research (BER) program, among them bioremediation, global climate change, biotechnology, and energy production.

Microbes drive the chemistry of life and affect the global climate.

Why Microbes?

By some estimates, microbes make up about 60% of the Earth's biomass, yet less than 1% of microbial species have been identified. Microbes play a critical role in natural biogeochemical cycles. Because

Microbes make up most living matter and display tremendous diversity, yet less than 1% have been cultured and studied.

What's a Genome?

A Short Primer on DNA Science

All living things, including microbes, have a chemical called *DNA* (for deoxyribonucleic acid) that contains information used by the organism to build, maintain, and reproduce itself. DNA is made up of four chemical building blocks (bases) that are abbreviated A, T, C, and G. Thousands to millions of these bases, depending on the organism, form long strands that pair together (A with T and C with G) in a twisted zipper-like structure often described as a double helix.

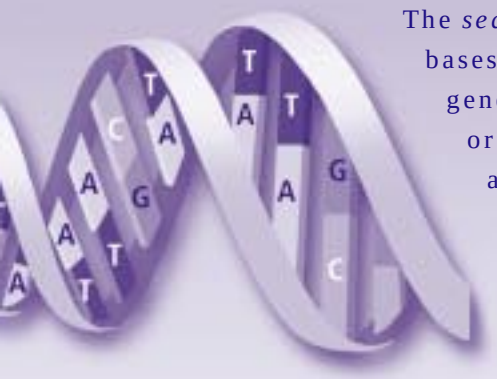
All the DNA in an organism is called its *genome*. Genomes range widely in size: the smallest known bacterial genome contains about 600,000 base pairs and the human genome some 3 billion. Except for the order and number of bases in the genome, DNA from all organisms is made of the same chemical and physical components.

Genes are segments of DNA that contain instructions on how to make the proteins that comprise all the organism's structures and run its functions. The genome of the microbe *Mycoplasma genitalium*, considered the smallest living organism, contains about 500 genes; in contrast, the human genome is estimated to contain some 140,000 genes. Combinations of genes, often interacting with environ-

most do not cause disease in humans, animals, or plants and are difficult to culture, they have received little attention. Microbes have been found surviving and thriving in an amazing diversity of habitats, in extremes of heat, cold, radiation, pressure, salinity, and acidity, often where no other life forms could exist. Identifying and harnessing their unique capabilities, which have evolved over 3.8 billion years, will offer us new solutions to longstanding challenges in environmental and waste cleanup, energy production and use, medicine, industrial processes, agriculture, and other areas. Scientists also are starting to appreciate the role played by microbes in global climate processes, and we can expect insights about both the biological underpinnings of climate change and the contributions of microbes to Earth's biosphere. Their capabilities soon will be added to the list of traditional commercial uses for microbes in the brewing, baking, dairy, and other industries.

Why is it Important?

mental factors, ultimately determine all the physical characteristics of an organism.



The *sequence* of DNA bases for a particular gene (e.g., ATCCTGC or CCATTCG) acts as a code that an organism's cellular machinery understands and can translate into a specific protein. The ultimate goal of most *genome projects* is to determine the animal's, plant's, or microbe's unique DNA sequence and identify the genes contained in that sequence. The genomes of many organisms, including humans, have much nongene (noncoding) DNA.

An increasing number of complete genome sequences are being generated and deposited into large and small public databases. The next great challenge—to understand the genome data—is the focus of a new and growing field called *functional genomics*, which will require the expertise of computer scientists and other investigators across a broad range of disciplines. Questions they might ask include, What are the functions of each gene? and How does it interact with all the other genes and environmental factors to create and maintain an organism?

The focus of DOE's MGP is to use high-throughput, cost-effective DNA sequencing capabilities to provide information on microorganisms with potential environmental, energy, or commercial applications.

Microbes do most of the biogeochemical cycling that keeps the world habitable.

Program Origins

To explore the possibilities for new applications, in 1994 the U.S. Department of Energy (DOE) established the Microbial Genome Program (MGP) as a companion to its Human Genome Program (HGP). From the start, the MGP experienced remarkable success, and microbial genomics has become one of the most exciting and high-profile fields in biology today.

A principal goal of this spin-off project is to determine the complete DNA sequence—the genome—of a number of nonpathogenic microbes that may be useful to DOE in carrying out its missions (nonpathogenic microbes do not cause disease). The microbes chosen for genomic sequencing were selected with broad input from the scientific community. "The microbial diversity of the program is an absolute treasure trove for [research in] biotechnology, ecology, evolution, and bioremediation," notes David Schlessinger (National Institute on Aging).

Only a few years ago, scientists could not have imagined having full access to the genetic structure of more than a few such organisms. Today, nearly three dozen complete microbial genomes, eleven supported by DOE's MGP, have been sequenced, and the rate of reported new genome sequences is increasing rapidly. (For a current listing, see *www.tigr.org/tdb/mdb/mdb/html*.) These DNA sequences, along with those from many viruses and more complex organisms such as fruitfly, roundworm, and yeast, are freely available in public databases. This information is being used by governmental, academic, medical, and industrial scientists. The number of possible applications of this information is staggering. Sequenced genomes provide us with a genetic "parts" list; the next challenge is to explore how these parts come together to form a functioning organism.

This booklet describes projects, accomplishments, and potential benefits of the innovative work supported by DOE in its MGP. Although much more remains to be studied, this program represents a first but vital step toward a greater understanding of the bountiful microbial resources surrounding us, as well as safe ways to exploit their unique beneficial qualities.

A Historic Microbe

The bacterium* Clostridium acetobutylicum *enjoys an unusual place in history. This microbe was discovered in 1915 by Chaim Weizmann, who noted its ability to convert starch into the organic solvents acetone and butanol, which have multiple applications in industrial settings. Shortly afterward, Great Britain used* C. acetobutylicum *to generate acetone for producing cordite for artillery shells in World War I. In gratitude for Weizmann's work, the British government offered to honor him, but he asked instead for support of a Jewish homeland in Palestine. This led to the Balfour Declaration of 1917, committing Britain to sanction what became in 1948 the state of Israel, with Weizmann as its first president.

Scientists hope the availability of this bacterium's genomic sequence, finished in 1999, will lead to a better understanding of its biochemistry and eventually to the replacement of current processes that rely on petroleum and natural gas for organic solvent production. Additionally, some* Clostridia *species are major pathogens. One produces the food toxin that causes botulism, and others are responsible for such rapidly spreading infections as tetanus and gangrene. DNA sequence comparisons of these species could yield insights into what enables some to cause harm to humans.

Organisms Chosen for MGP Studies

In 1995, the MGP's first full year, DOE funded four microbial genome sequencing projects focused on the bacterium

Microbes first appeared on Earth about 3.8 billion years ago. They are critically important in sustaining life on our planet.

Mycoplasma genitalium and three other microbes. Now fully characterized, the tiny *M. genitalium* genome—thought to have the smallest genome of any known free-living bacterium—provides a model for a minimal set of genes necessary for life. Its genome contains only 580,000 base pairs of DNA and yet encodes 470 genes. Future studies on this and other minimal genomes will help increase our understanding of more complex genomes.

Among the oldest life forms known, the Archaea make up one of three phylogenetic or evolutionary domains into which all life is classified. The other two are the Eukarya and the Bacteria (see diagram, p. 19). Archaea found thriving in extreme environments of heat and cold, acidity, pressure, and salinity are known as extremophiles ("extreme-loving" organisms). Understanding the biological mechanisms underlying their hardiness may help researchers develop new industrial, biomedical, and environmental applications.

Microbes may, for example, contain enzymes that are effective in driving chemical reactions in extreme environments. Some may provide enzymes useful in research; one such "extremozyme" derived from a bacterium living in hot springs in Yellowstone National Park has become critical to current protocols for sequencing any genome, including that of humans. Other microbes have metabolic processes with potential for breaking down toxic waste or even producing methane, an energy source.

Comparisons of the genomes of organisms from all three domains are helping scientists better understand the evolution of all living things. Descriptions of MGP-supported research on some other microbes follow.

- *Methanococcus jannaschii* was among the first archaea chosen for sequencing. In 1996 its completed sequencing and analysis confirmed that the "tree of life" has three domains, a hypothesis first advanced nearly 20 years ago by Carl Woese (University of Illinois) but not given much credence at the time. The single-celled *M. jannaschii* was isolated from a sample collected beneath more than 8000 feet of water at the base of a deep-sea thermal vent on the floor of the Pacific Ocean. The microbe lives without the sunlight, oxygen, and

Microbes are at the roots of life's family tree. They will help us understand how more complex genomes developed.

organic carbon important to most other forms of life and uses carbon dioxide, nitrogen, and hydrogen expelled from the thermal vent for its life functions. When the entire DNA sequence *of M. jannaschii* was determined, scientists found that about 65% of its potential gene sequences were not related to any gene previously discovered, representing an exciting area for future investigation.

- The archaeon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmental remediation. Because they thrive in water temperatures above the boiling point, these organisms may provide DOE, the Department of Defense, and private companies with heat-stable enzymes for use in industrial processes. These processes could include conversion of wastes to useful chemicals. *A. fulgidus* has the added capability of surviving at the high pressures associated with deep oil wells, and *T. maritima* metabolizes simple and complex carbohydrates, including glucose, sucrose, starch, xylan, and cellulose. Cellulose and xylan are the most abundant biopolymers on Earth and, through their conversion to fuels such as ethanol, have major potential as sources of renewable

"Superbug" Survives Radiation, Eats Toxic Wastes

A can of spoiled meat and nuclear waste may appear to have little in common, but the bacterium* Deinococcus radiodurans *thrives in both environments. This bacterium was discovered in 1956 when it was identified as the culprit in a can of spoiled ground beef thought to be radiation "sterilized." Scientists subsequently learned that its extreme radiation resistance enables the microbe to survive doses thousands of times higher than would kill most organisms, including humans. The remarkable DNA-repair processes of* D. radiodurans *allow it to stitch together flawlessly its own radiation-shattered genome in about 24 hours.

DOE chose this organism for DNA sequencing because of its potential usefulness in cleaning up waste sites containing radiation and toxic chemicals. Its DNA sequence was completely determined in 1999, and scientists now are exploring ways to add genes from other organisms to expand* D. radiodurans*' capabilities for removing toxic wastes from contaminated sites. The added genes encode proteins that transform heavy metals to a more benign biomass and allow the concentration of heavy metals and the breakdown of organic solvents such as toluene. Studies into this organism's remarkable DNA-repair pathways also may help scientists better understand how defects in human cellular processes might lead to the development of cancers.

energy. Comparisons of the genomic sequences of these two microbes will contribute to a greater understanding of evolutionary relationships as well as high-temperature protein function.

- The archaeon *Pyrobaculum aerophilum*, first isolated from a boiling marine vent, thrives at temperatures close to the maximum tolerated by living systems (113°C). Unlike most hyperthermophiles, *P. aerophilum* is able to withstand exposure to oxygen and can thus be manipulated more easily in the laboratory. Also, the proteins encoded by hyperthermophilic genomes are more stable than those of organisms living in more temperate environments.

- The bacterium *Shewanella putrefaciens*, which can grow with or without oxygen, is an excellent model system for manipulating organisms for remediation. Whole-genome sequencing will elucidate metabolic pathways including those involved in corrosion, consumption of toxic organic pollutants, and removal of toxic metals and radiation waste by conversion to insoluble forms.

Other organisms that could be of great genetic and biochemical interest are present in extreme surface environments but are almost impossible to grow in the laboratory. The MGP funds a project to identify and determine the abundance and activity of novel hard-to-cultivate organisms in two extreme surface environments in the arid southwestern United States. Preliminary samples indicate that most of these bacterial species contain few similarities to the previously described cultivated bacteria. These collections offer a rich resource for identifying and isolating novel species with potentially unique sets of genes as well as proteins with environmental, energy, biotechnological, and other applications.

A Closer Look at the Data

After a microbe's complete genomic sequence is determined, these data are analyzed, or annotated, to identify all the potential genes and to get clues about possible gene functions.

One of the surprises emerging from the study of a number of microbial sequences is the presence of genetic segments

Microbes generate at least half the oxygen we breathe.

Microbial studies will help us define the entire repertoire of organisms in specialized niches and, ultimately, the mechanisms by which they interact in the biosphere.

containing not just single genes but entire blocks of multiple genes that appear to have been acquired intact during evolution from other microbes in very distant parts of the tree of life. The bacterium *T. maritima* is hypothesized to have acquired a quarter of its genome through this process, which is termed "lateral gene transfer." These findings present exciting challenges to our understanding of how microbial species live and evolve.

The availability of complete genomes has opened up the entirely new field of comparative genomics, which is allowing researchers to identify genes that are similar across species. Comparative genomics is providing clues into the functions of genes and how genomes change over time. Comparative studies also are having a profound impact on the ability to discover novel genes and biochemical pathways. To interpret genome sequences, scientists first

Genetic Engineering and Biotechnology

Scientists can now identify genes that influence desirable physical features in one organism and transfer them into others. Such **genetic engineering** ***results in altered (or recombinant) organisms having a combination of desired traits. Using genetically modified living organisms or their products for commercial purposes is an emerging area in biotechnology.***

In the Microbial Genome Program, scientists are altering the genome of the bacterium **Deinococcus radiodurans** ***to increase its potential usefulness in cleaning up toxic-waste sites around the globe. Studies have revealed that the microbe's extraordinary DNA-repair processes enable it to thrive in high-radiation environments. Through the use of biotechnological processes, scientists hope to add genes from other organisms that will confer the ability to degrade toxic chemicals such as toluene, commonly found in mixed, chemical, and radiation waste sites.***

Other examples of current and potential applications of genetic engineering follow.

compare them to other entries in DNA sequence databases. Astonishingly, some one-third of genes discovered in newly sequenced genomes do not have database matches and, therefore, no easily identifiable functions.

Microbes offer unusual capabilities reflecting the diversity of their environmental niches. These may prove to be useful as a source of new genes and organisms of value in addressing bioremediation, global change, biotechnology, and energy production.

These data demonstrate how little is known about microbial species, and, as each completed genome sequence reveals novel sets of genes, many of them may be identified as unique to a particular species or to particular functions present in some strains of microbes but lacking in others. These newly identified genes represent exciting opportunities for future basic research and potential sources of biological resources to be explored for future use. The usefulness of comparative techniques improves dramatically as more genomes

- ***Production of pharmaceuticals by bacteria that produce human insulin for diabetics or human growth hormone for individuals with dwarfism. Scientists are perfecting ways to transfer human genes for important proteins into cows, sheep, and goats to obtain medically significant products from the milk of these animals.***
- ***Development of diagnostics to detect disease-causing organisms and monitor the safety of food and water supplies. Investigators also are developing systems for identifying pathogens that may someday be used as biological weapons by rogue nations or even terrorist groups.***
- ***Use of bacteria as living sensors (biosensors) of particular chemicals in soil, air, and water. In some studies, bacteria have been genetically altered to emit a green fluorescent protein visible in ultraviolet light when they metabolize the explosive TNT leaking from land mines. Researchers envision a day when bacteria can be applied to a tract of land with a crop duster and then analyzed from a helicopter.***

Resources on the Web

Microbial Web Sites

Genomes and Genome Projects

http://www.er.doe.gov/production/ober/EPR/mig_top.html

http://www.ornl.gov/hgmis/publicat/microbial

http://bbrp.llnl.gov/jgi/microbial

http://compbio.ornl.gov

http://www.tigr.org/tdb/mdb/mdb.html

http://geta.life.uiuc.edu/~nikos/genomes.html

http://www.beowulf.org.uk/home.htm

http://www.sanger.ac.uk/Projects/Microbes

http://www3.ncbi.nlm.nih.gov/Entrez/Genome/org.html

http://www3.ncbi.nlm.nih.gov/BLAST/unfinishedgenome.html

http://ncgr.org/microbe (archive only)

Microbial Information Broker

http://mol.genes.nig.ac.jp/gib

Discussion Group

http://www.medmicro.mds.qmw.ac.uk/microbial-genomes

Metabolic Pathways

Clusters of Orthologous Groups (COG) Database

http://www.ncbi.nlm.nih/COG

EcoCyc and MetaCyc

http://ecocyc.pangeasystems.com/ecocyc

KEGG (Kyoto Encyclopedia of Genes and Genomes)

http://www.genome.ad.jp/kegg

Metabolic Pathways on Internet

http://home.wxs.nl/~pvsanten/mmp/mmp.html

WIT

http://wit.mcs.anl.gov/WIT2

URLs Associated with Research Abstracts

Archaeal Genome Sequence Database

http://comb5-156.umbi.umd.edu

Caltech Genome Research Laboratory

http://www.tree.caltech.edu

Chisholm Laboratory

http://web.mit.edu/chisholm/www

Computational Biosciences, Oak Ridge National Laboratory

http://compbio.ornl.gov

Department of Human Genetics, University of Utah

http://www-genetics.med.utah.edu/index.html

Diversa Corporation

http://www.diversa.com

Genome Therapeutics

http://www.genomecorp.com

Genomics Group, Brookhaven National Laboratory

http://www.genome.bnl.gov

Microbe World

http://www.microbeworld.org

Microbial Sequencing, Joint Genome Institute

http://bbrp.llnl.gov/jgi/microbial

Ribosomal Database Project II

http://www.cme.msu.edu/RDP

Schwartz Laboratory

http://www.chem.wisc.edu/~schwartz

The Institute for Genomic Research

http://www.tigr.org

WIT2

http://wit.mcs.anl.gov/WIT2

become available, but better methods are needed to expand the types of analyses that can be performed.

While the DOE MGP focuses primarily on environmental, energy, and biotechnological areas, there may be spinoff applications in medicine as well. Potential biomedical benefits from comparative genomics studies include insights into the specialized, shared systems used by disease-causing organisms (pathogens) to disable or destroy human cells. Comparing these genomic data with those of other microbes may help scientists understand a diverse range of pathogens that have remarkably similar methods for infiltrating organisms with protein-coding genes capable of sneaking past human defense systems. These protein structures may provide ideal targets for developing completely new types of antibiotics.

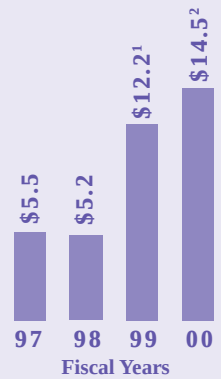
Researchers have only scratched the surface of microbial biodiversity. Given the pervasiveness of microbes in all environments as well as their ancient evolutionary history, one can expect to find a vast repertoire of useful functions in the microbial world that could be applied to solving challenges in the human world.

Microbial genomes are modest in size and relatively easy to study (usually no more than 10 million DNA bases, compared with some 3 billion in the human or mouse genomes).

What's in the Future?

The major focus of the DOE MGP will continue to be on genomic sequencing of microbes relevant to DOE missions. To avoid "starting from scratch" in sequencing new microbes, investigators are developing novel strategies to cost-effectively determine the DNA sequence of microbes that are very closely related to others whose sequence already is known. Additionally, the MGP is developing new tools to study how groups of genes work together to produce specific products or determine particular behaviors. Other

DOE Microbial Genome Budgets
(millions)



Totals include [1]$2.4M and [2]$5.8M from DOE Carbon Management Science Program (CMSP). CMSP is part of the federal Climate Change Technology Initiative.

Diversity patterns of microorganisms can be used for monitoring and predicting environmental change.

Uncovering Microbial Genome Potential

DOE-Supported Projects

**Aquifex aeolicus* (bacteria extremophile, 1.5 Mb): Potential for identifying high-temperature enzymes.

**Archaeoglobus fulgidus* (archaea extremophile, 2.3 Mb): Potential for identifying high-temperature and high-pressure enzymes; useful in oil industry.

**Borrelia burgdorferi* (bacteria, 1.3 Mb): Human pathogen that causes Lyme disease. One linear chromosome (915 kb) supported by DOE. Entire genome published by TIGR.

Caulobacter crescentus (bacteria, 3.8 Mb): Potential for heavy-metal remediation in waste-treatment plant wastewater; simple developmental cycle.

Chlorobium tepidum (bacteria, 2.1 Mb): Photosynthetic; may play important role in Earth's overall carbon cycle.

***Clostridium acetobutylicum* (bacteria, 4.1 Mb): Produces acetone, butanol, and ethanol; useful for industrial enzymology.

Dehalococcoides ethenogenes (bacteria, less than 2 Mb): Degrades dangerous solvent trichloroethene to benign products.

**Deinococcus radiodurans* (bacteria, 3 Mb): Survives extremely high levels of radiation; possesses DNA-repair capabilities for radioactive waste cleanup.

Desulfovibrio vulgaris (bacteria, 1.7 Mb): High potential for bioremediation through metal and sulfate reduction.

Geobacter sulfurreducens (bacteria, 1 Mb): Reduces a variety of metals, including iron and uranium.

**Halobacterium halobium* plasmid (archaea, genome size 2 Mb, NRC100 plasmid size 190 kb): Potential for identifying high-salinity enzymes.

**Methanobacterium thermoautotrophicum* (archaea, 1.7 Mb): Produces methane; plays role in earth's overall carbon cycle.

**Methanococcus jannaschii* (archaea extremophile, 1.7 Mb): Potential for identifying high-temperature, high-pressure enzymes; produces methane.

Methylococcus capsulatus (bacteria, 4.6 Mb): Uses methane as single carbon and energy source; generates enzymes that oxidize some pollutants; used commercially to produce biomass and other proteins of interest.

**Mycoplasma genitalium* (bacteria, 580 kb): Human pathogen; serves as model for minimal set of genes sufficient for free-living existence.

†Nitrosomonas europaea (bacteria, 1.6 Mb): Important in soil nitrogen cycling and ammonia oxidation; promotes incorporation of carbon dioxide into biomass.

†Nostoc punctiforme PCC-73102 (bacteria, 8 Mb): Fixes carbon dioxide and nitrogen; produces hydrogen; survives acidic, anaerobic, and low-temperature conditions.

†Prochlorococcus marinus MED4 (bacteria, 2 Mb): Abundant in temperate and tropical oceans; absorbs blue light efficiently; important in ocean carbon cycling.

Pseudomonas putida (bacteria, 5 Mb): High potential for bioremediation by reducing metal and pollutants.

***Pyrobaculum aerophilum* (archaea extremophile, 1.8 Mb): Potential for identifying high-temperature enzymes.

***Pyrococcus furiosus* (archaea extremophile, 2.1 Mb): Potential for identifying high-temperature enzymes.

†Rhodopseudomonas palustris (bacteria, 4 to 5 Mb): Fixes carbon dioxide; produces hydrogen; biodegrades organic pollutants under both aerobic and anaerobic conditions.

Shewanella putrefaciens (bacteria, 4.5 Mb): Potential for degrading toxic organic wastes and for sequestering toxic metals.

**Sphingomonas aromaticivorans* F199pNRC100 plasmid (bacteria, genome size est. 4 Mb; pNL1 plasmid size 184 kb): Deep-soil organism with potential for degrading toxic organic compounds.

†Synechococcus (bacteria, genome size undetermined): Photosynthetic; uses nitrate and ammonia as nitrogen sources.

**Thermotoga maritima* (bacteria extremophile, 1.8 Mb): Potential for identifying high-temperature, high-pressure enzymes.

Thiobacillus ferrooxidans (bacteria, 2.9 Mb): Used in mining industry to sequester iron and sulfide.

---

*Completed and published (see *www.tigr.org/tdb/tdb.html*)

**Completed, not published (as of January 2000)

†BER Carbon Management Science Program, a part of the federal Climate Change Technology Initiative

objectives are to mine genomic information from sequenced microbes, improve tools for annotation and analysis of sequence data, develop high-throughput methods for determining gene function and gene expression, and develop methods for examining protein-protein and protein-nucleic acid interaction.

The future promises many exciting developments as the fruits of the MGP mature. Already, we have become more appreciative of the extent of the microbial world's effect on Earth, realizing how little we know about this kingdom and wondering at its potential benefits to our world—if only we are wise enough to discover them.

Most microbes do not cause disease.

DOE Microbial Research and Related Activities

Office of Biological and Environmental Research (OBER), Office of Science

Overview of OBER Research: *http://www.sc.doe.gov/production/ober/ober_top.html*
Programs and Staff Contacts: *http://www.sc.doe.gov/production/ober/restaff.html*
Funding: *http://www.sc.doe.gov/production/grants/grants.html*

Carbon Management Science Program

http://www.sc.doe.gov/production/ober/carbseq.html

Genome Programs

Microbial Genome Program

http://www.sc.doe.gov/production/ober/microbial.html
http://www.ornl.gov/hgmis/publicat/microbial

Human Genome Program

http://www.sc.doe.gov/production/ober/hug_top.html
http://www.ornl.gov/hgmis

Global Change Research Program

http://www.sc.doe.gov/production/ober/esdrestopic.html

Natural and Accelerated Bioremediation Research (NABIR) Program

http://www.lbl.gov/NABIR

Bioremediation of Metals and Radionuclides: A NABIR Primer

http://www.lbl.gov/NABIR/primer/primer.html

Structural Biology Research Program

http://www.sc.doe.gov/production/ober/msd_struct_bio.html

Environmental Molecular Science Laboratory

http://www.emsl.pnl.gov:2080

Other DOE Offices Supporting Microbial Research

DOE is a major funder of nonmedical microbiology in the U.S. government. Some programmatic areas of microbial study are listed below (for more details, see *www.DOE.gov/people/peoppo.htm*).

Office of Basic Energy Sciences, Office of Science

Renewable energy, carbon sequestration

Office of Energy Efficiency and Renewable Energy

Renewable energy, hydrogen production, ethanol production, organic acid synthesis, cellulose and lignin degradation

Office of Environmental Management

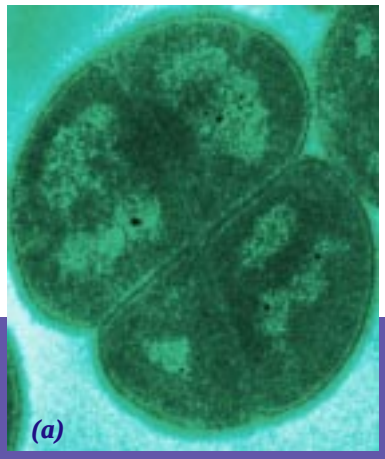
Bioremediation research (organics)

Office of Nonproliferation and National Security

Characterization and detection of potential biological warfare agents

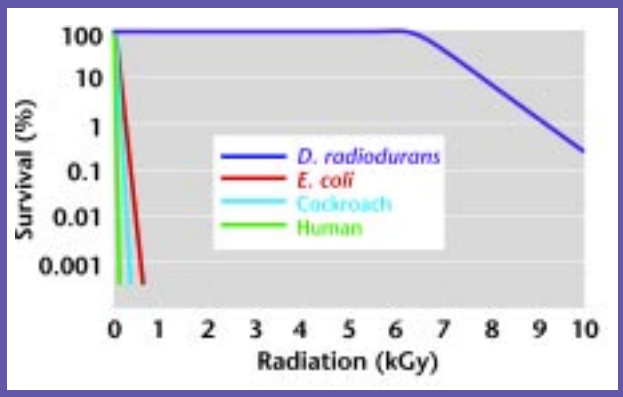
Microbial Genomes in View

Super Survivor *Deinococcus radiodurans*



Although the ability of the lowly cockroach to withstand radiation has long been admired, it is far surpassed by that of the bacterium* Deinococcus radiodurans. *Scientists are eager to learn how* D. radiodurans *thrives in massive amounts of radiation and how to exploit this property to clean up mixed-waste sites around the world, a legacy resulting from nuclear weapons production between 1945 and 1986. Researchers are adding genes from other species to* D. radiodurans *to increase its ability to remediate radioactive sites containing such metallic and organic contaminants as mercury and toluene.

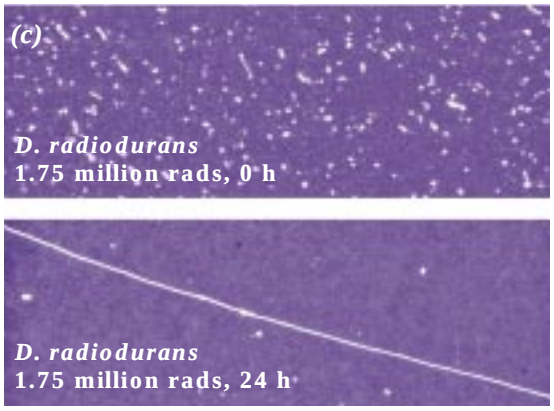
(a) Electron photomicrograph of* D. radiodurans *(sequenced in the DOE MGP), typically found as a cluster of four cells (a tetrad).* D. radiodurans *and related species have been identified worldwide, including in antarctic granite and in water-shielding tanks of powerful ^{60}Cobalt irradiators in Denmark.

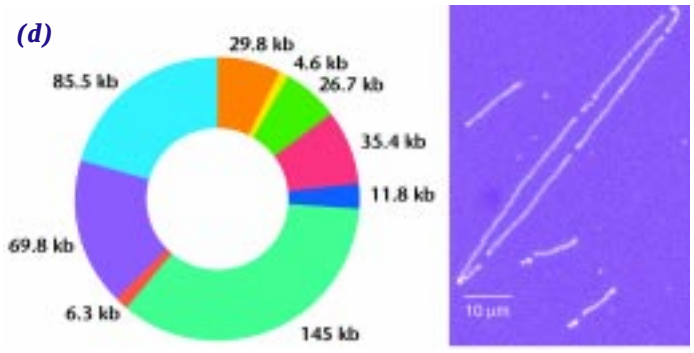


(b) The radiation-resistance profile of* D. radiodurans *compared to such other organisms as the common intestinal bacterium* Escherichia coli, *cockroaches, and humans. When older colonies of* D. radiodurans *are used, their survival extends much farther, to around 17 kGy (1.7 million rads). Scientists believe this extreme radiation resistance may be a side effect of* D. radiodurans' *ability to survive severe dehydration, which also fragments DNA.* [Nature Biotechnology *18, 85–90 (January 2000)]

(a, b) Image and graph data by the* Deinococcus *team (Uniformed Services University of the Health Sciences)

D. radiodurans: The Ultimate Assembly Machine



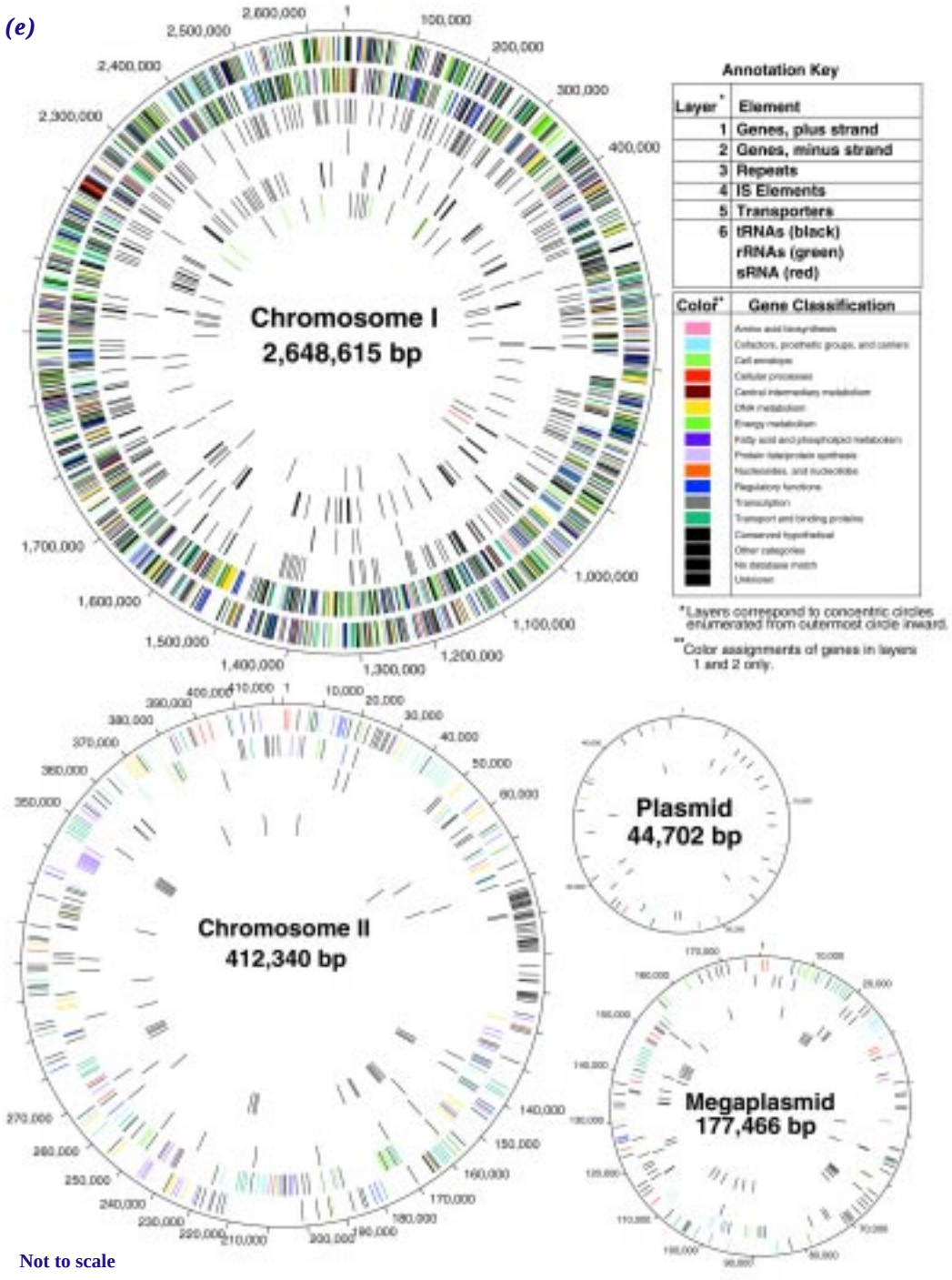


(c) The upper panel depicts DNA fragments extracted from* D. radiodurans *cells after high doses of radiation. The lower panel shows an intact, repaired DNA molecule hours later. Both panels depict "optical maps" of molecules viewed on a slide through an optical light microscope. Understanding the remarkable DNA-repair mechanisms of* D. radiodurans *may offer insights into some human cancers caused by DNA damage.

***(d) Ordered restriction map (colored circle) and optical map of a single circular 415,000-base (415-kb) DNA molecule snipped apart using a special DNA-cutting protein, the restriction enzyme* Nhe *I. Circular DNA elements are difficult to identify using nonoptical approaches, since these molecules break and become linear elements. Optical mapping generates a picture of the entire genome's architecture, revealing the number of chromosomes and the existence of extrachromosomal elements. This technique was critical to the discovery that* D. radiodurans *has four chromosomal elements rather than just one.* [Science *285, 1558–62]**

(c, d) Photos provided by David Schwartz (University of Wisconsin, Madison)

(e)



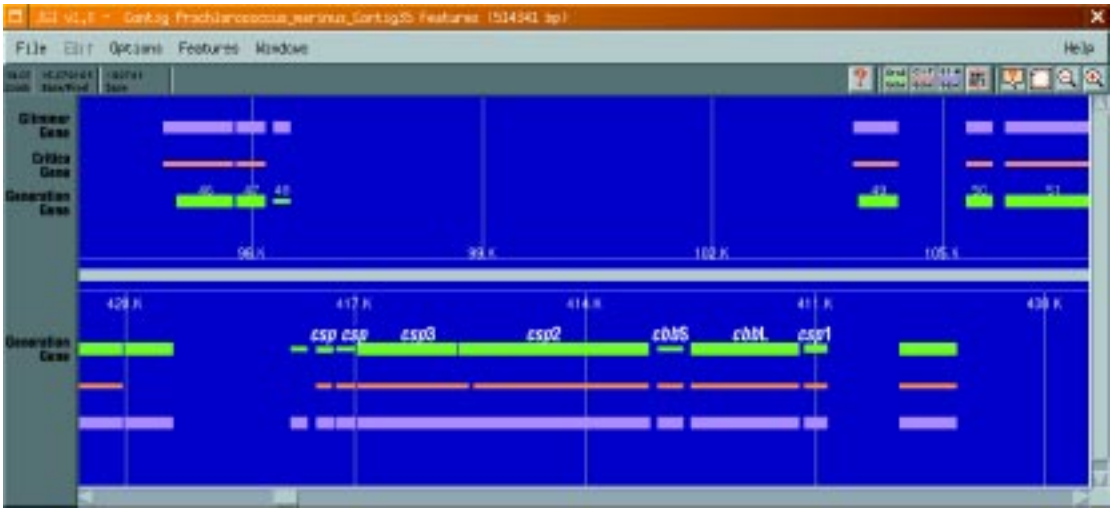
Not to scale

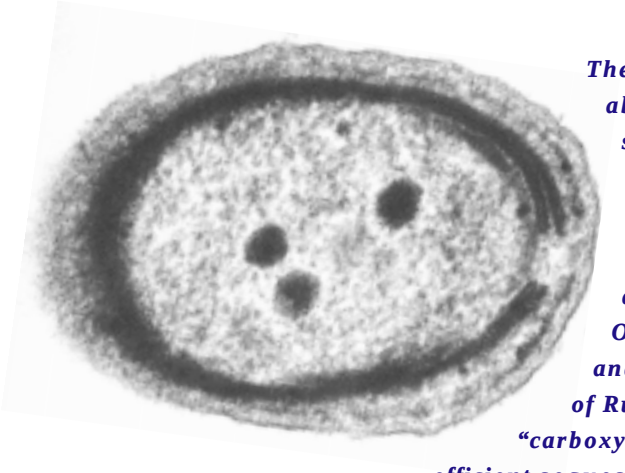
(f) Graphic representations of the four DNA molecules that compose the D. radiodurans *genome. Each circular molecule contains the genes (depicted as colored spikes in the two outermost layers of each circle) that encode all the functions enabling* D. radiodurans *to replicate, metabolize nutrients, and withstand such environmental insults as radiation exposure. [Reprinted, with permission, from White et al.,* Science *286, 1571–77, © 1999, American Association for the Advancement of Science]*

Diagram provided by Owen White (The Institute for Genomic Research)

See also "Superbug Survives Radiation," p. 6.

Designed for Success: DNA Analyses Reveal Details





The microscopic marine* Prochlorococcus *may be the most abundant photosynthetic organism on Earth. Analyses of DNA sequence data now offer some clues as to how the genomic structures of these organisms contribute to their success.

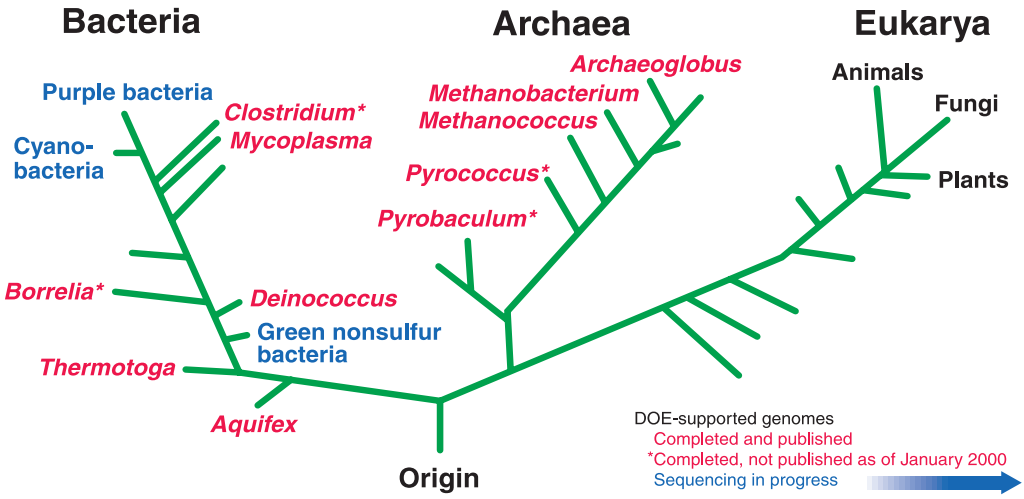
Prochlorococcus marinus MED4 ***(electron photomicrograph at left) uses the enzyme Rubisco to help convert or "fix" carbon dioxide to complex sugars in carrying out photosynthesis. Oxygen, a byproduct of photosynthesis, competes for Rubisco and could potentially interfere with this process. The organization of Rubisco and other proteins into tiny structures called "carboxysomes" (three dark-colored bodies in photo), however, promote efficient sequestering of carbon dioxide and help promote its fixation to the complex sugars used as chemical energy.***

Analyses indicate that the genes encoding Rubisco [*cbbL *and* cbbS *above the green line in the screen shot from the Genome Channel (*genome.ornl.gov*)] are organized into a complex in the same genomic region that also encodes components of the carboxysome shell proteins (*csp*). This genetic organization can ensure that the required amounts of each protein are produced at the proper time.

Researchers use multiple systems for identifying genes in microbial DNA; the systems perform well in combination to generate a consensus gene model. The screen shot depicts the result of using three gene-finder programs on* P. marinus *DNA sequences: Glimmer (pink line), Critica (orange line), and Generation (green line). Conceptual translations of these gene models are used to search for similarities to other gene sequences in public databases and to determine their relationships to particular protein families. Researchers use these results to construct metabolic frameworks and assign particular functional roles to genomic sequences.

Genome Channel draft contig view and annotation by Frank Larimer (Oak Ridge National Laboratory); original DNA sequence provided by Jane Lamerdin (Joint Genome Institute and Lawrence Livermore National Laboratory); electron photomicrograph of* Prochlorococcus *(isolate* MIT 9313*) courtesy of Sallie Chisholm (Massachusetts Institute of Technology)

Shaking Up the Tree of Life



Genomes in Progress

Purple bacteria
- *Caulobacter*
- *Dehalococcoides*
- *Desulfovibrio*
- *Geobacter*
- *Methylococcus*
- *Nitrosomonas*
- *Pseudomonas*
- *Shewanella*
- *Rhodopseudomonas*
- *Thiobacillus*

Cyanobacteria
- *Nostoc*
- *Prochlorococcus*
- *Synechococcus*

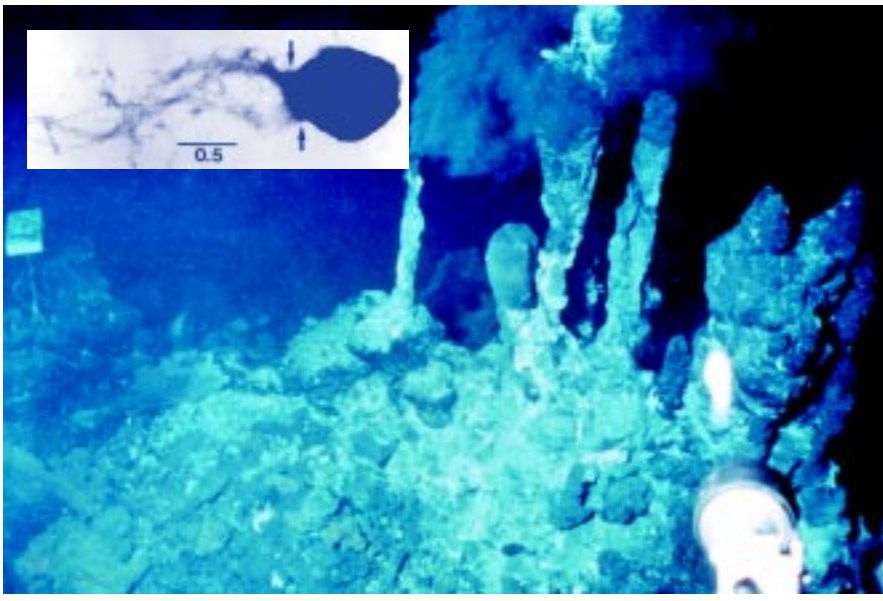
Green nonsulfur bacteria
- *Chlorobium*

Microbe positional data provided by Frank Larimer (Oak Ridge National Laboratory)
For details on specific species supported by DOE, see p. 12.

This tree of life—or phylogenetic tree—traces the pattern of descent of all life over millions of years into three major branches: Bacteria, Archaea, and Eucarya. Until 1996, however, scientists had confirmed the existence of only two of those branches. And although Earth's biomass is largely microbial, most previous studies focused primarily on a tiny portion at the tip of the Eucarya branch, the region containing animals and plants. Newly available microbial DNA sequences on the other two branches, many supported in the DOE Microbial Genome Program (MGP), have enabled large-scale genomic comparisons among different organisms for the first time, and the information is changing some long-held views of the history of life.

In 1996, for example, comparisons of genomic sequences from the microbe* Methanococcus jannaschii *(electron photomicrograph below) with those of other organisms confirmed the existence of the archaeal branch of life.* M. jannaschii *was first isolated in 1983 in the area of a "smoker," a hydrothermal vent on the floor of the Pacific Ocean. Thriving at pressures that would crush a conventional submarine, this heat-loving, methane-producing microbe lives without sunlight, oxygen, or organic carbon. DOE MGP researchers hope to exploit these unusual properties, along with the unique characteristics of other microbes studied. Increasing knowledge about microbial life and its enormous range of capacities will have far-reaching implications for environmental, energy, health, and industrial applications.

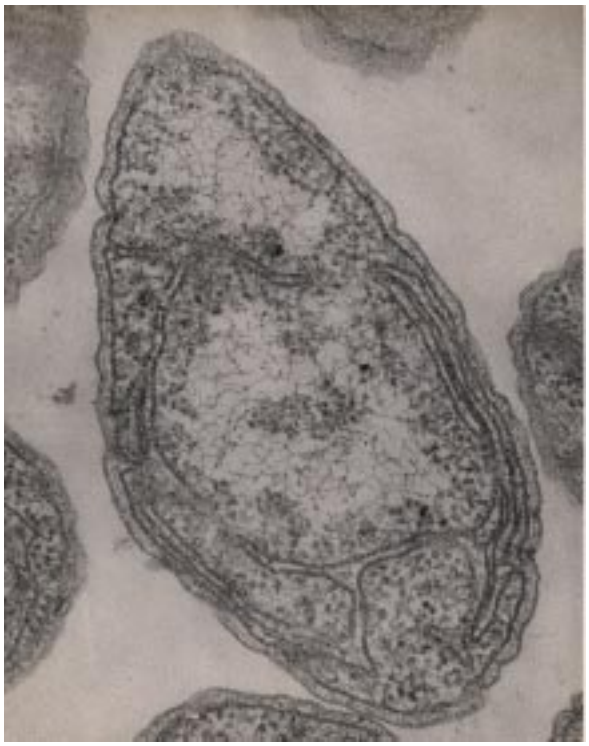
New revelations also are showing some surprising connections among microbes previously thought to have diverged evolutionarily long ago. Researchers expect that, in addition to providing intriguing new details on phylogenetic trees, the continued explosion of genome data and analysis tools will lead to new insights into how biological systems function.



Photomicrograph of M. jannaschii *reprinted with permission from Springer-Verlag; photograph of hydrothermal vent, © Woods Hole Oceanographic Institution*

Microbes and Climate Change

In the past 60 years, the amount of carbon dioxide emitted, mainly though expanding use of fossil fuels for energy, has risen dramatically and is thought to contribute to global climate change. Unless we make major alterations in the way we produce and use energy, predictions for the next century suggest a continued increase in emissions as well as rising concentrations of carbon dioxide in the atmosphere. In addition to controlling fossil fuel emissions, other methods must be explored for stabilizing or decreasing carbon dioxide and other greenhouse gases.

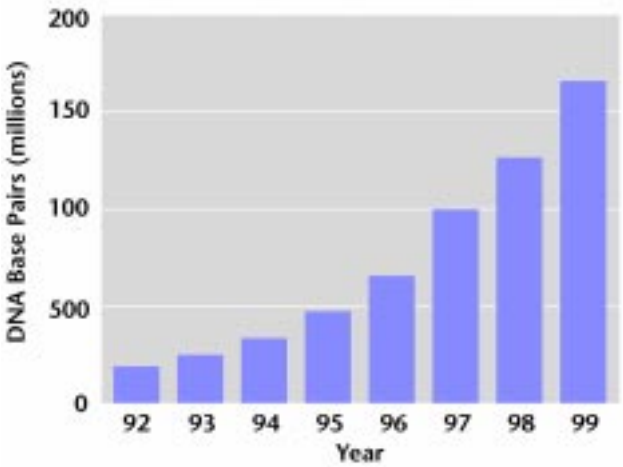


Electron photomicrograph of **Nitrosomonas**
© Stan Watson (Woods Hole Oceanographic Institution)

***One of the many poorly understood aspects of the global warming phenomenon is the role that specific microorganisms play in the natural carbon cycle on Earth. As part of its recently launched Carbon Management Science Program, DOE hopes to stabilize or decrease atmospheric carbon dioxide concentrations by identifying ways to manage carbon levels in the terrestrial biosphere. One avenue of exploration is to sequence the genomes of microbes that use carbon dioxide as their sole carbon source. These organisms include* Nitrosomonas europaea *(pictured),* Prochloro-coccus marinus, Rhodopseudomonas palustris, Nostoc punctiforme, *and a marine* Synechococcus.

Progress in Bacterial and Archaeal Sequencing

The establishment in 1995 of the DOE Microbial Genome Program helped jump-start the explosion of microbial genomic data entered into the public databases. Researchers estimate that more than 200 million additional microbial DNA bases will be sequenced in the next 2 to 3 years, with some 200,000 predicted genes. This is about twice the number of genes expected from the completion of the Human Genome Project. The biological functions of at least half of these genes will be unknown, underscoring the need for further explorations to discover and apply microbial capabilities in new ways to benefit humankind.



Graph data by Frank Larimer (Oak Ridge National Laboratory)

Abstracts of Research Projects

Sequencing and Analysis

Functional and Computational Analysis

Resources for Genomic Comparison

Ethical, Legal, and Social Issues

Sequencing the Genome of *Nitrosomonas europaea*, an Obligate Lithoautotrophic, Ammonia-Oxidizing Bacterium

Daniel J. Arp, **Alan B. Hooper**,[1] **Jane E. Lamerdin**,[2] David Arciero,[1] Andre Arellano,[2] Karolyn Burkhart-Schultz,[2] Anne Marie Erler,[2] Norman Hommes, Martin G. Klotz,[3] Jenny M. Norton,[4] Warren Regala,[2] Luis Sayavedra-Soto, and Stephanie Stilwagen[2]
Botany and Plant Pathology; Oregon State University; 2082 Cordley; Corvallis, OR 97331-2902
541/737-1294, Fax: -3573, *rpd@bcc.orst.edu*
[1]University of Minnesota
[2]Lawrence Livermore National Laboratory
[3]University of Louisville
[4]Utah State University

As part of the DOE initiative to explore the role of microorganisms in global carbon sequestration, the Joint Genome Institute intends to obtain the complete genomic sequence of the autotrophic nitrifying bacterium *Nitrosomonas europaea*. This organism is the most studied of the ammonia-oxidizing bacteria that are participants in the biogeochemical N cycle. Nitrifying bacteria play a central role in the availability of nitrogen to plants and hence in limiting CO_2 fixation. The reaction catalyzed by these bacteria is the first step in the oxidation of ammonia to nitrate. These bacteria also are important players in the treatment of industrial and sewage waste in the first step of oxidizing ammonia to nitrate. Evidence suggests that ammonia-oxidizing bacteria contribute significantly to the global production of nitrous oxide (produced by the reduction of nitrite). *N. europaea* also is capable of degrading a variety of halogenated organic compounds, including trichloroethylene, benzene, and vinyl chloride. The ability of nitrifying organisms to degrade some pollutants may make these organisms attractive for controlled bioremediation in nitrifying soils and waters.

N. europaea is a Gram-negative member of the β-proteobacteria subdivision, possessing a genome size of at least 2.2 Mb. The microbe can be transformed and deletion mutants engineered, allowing the study of genotype-phenotype relationships. To complete the sequence of the *N. europaea* genome, a whole-genome shotgun strategy is being used similar to that employed successfully for many tens of bacterial organisms. The 8× genome coverage generated by the shotgun data is being supplemented with a scaffold of paired end sequences from clones in the low-copy-number fosmid vector. Shotgun data from this organism were assembled with PHRAP (Phil Green, University of Washington) and will progress through "auto-finishing" using software written by Matt Nolan (JGI-LLNL) and David Gordon (University of Washington) prior to human intervention in the assembly. Fingerprinting of a minimal spanning path of fosmids will be used to aid verification of the final assembly. A sequence-analysis pipeline, devel-

oped by Manesh Shah and Frank Larimer of Oak Ridge National Laboratory, is being used to define open reading frames (ORFs) and query public databases for protein-nucleotide similarities. Periodic lists of putative ORFs will appear on the Web site (*http://bbrp.llnl.gov/jgi/microbial/nitrosomonas_homepage.html*) as the genomic coverage continues to grow. The raw sequence data also are directly queryable through the accompanying BLAST server or can be downloaded from the JGI ftp server.

This will be the second member of the β-subdivision to have been sequenced. The most-studied gene products in this organism are those involved in the oxidation of ammonia, principally the hydroxylamine oxidoreductase (HAO), ammonia monooxygenase (AMO), and the accompanying cytochromes that make up the electron-transport chain. We hope the genome sequence will reveal strong candidates for as-yet-unidentified proteins specific to the N-oxidation pathways unique to this organism. The nature and regulation of enzymes in the nitrite-to-nitrous oxide pathway also are of interest. The operon encoding the subunits of AMO is duplicated and the amino acid sequences of the two operons differ by only a single nucleotide. The gene that codes for HAO is present in three copies. The extent to which other genes are duplicated in the genome is not known but is one anticipated outcome of generating the genomic sequence of *N. europaea*. As one of the few strictly autotrophic bacteria currently being sequenced, *N. europaea*'s genome sequence is expected to reveal the identity and number of genes required for and suited to autotrophy and possibly provide an indication of the basis for obligate autotrophy. The sequence will allow direct comparison to genes identified in another lithoautotrophic organism, *Thiobacillus ferroxidans*, which derives its energy from the oxidation of iron or sulfur compounds. Comparison of the metabolic capabilities of this organism with those of photoautotrophs and other lithoautotrophs may reveal the range of capabilities that were lost or gained as *N. europaea* descended from its evolutionary ancestors.

Nostoc Genome Sequencing

Ronald M. Atlas
Department of Biology; University of Louisville; Louisville KY 40292
502/852-3957, Fax: -0725, *r.atlas@louisville.edu*

An expert advisory panel met with Jane Lamerdin of the Joint Genome Institute (JGI) to select a strain of the heterocystous cyanobacterium *Nostoc* for genome sequencing. Based upon its relevance to carbon sequestration and the likelihood of providing significant new scientific information, the panel selected *Nostoc punctiforme PCC 73102, ATCC 29133*. This strain fixes nitrogen and carbon dioxide, forms symbiotic relationships, exhibits cell differentiation with the formation of motile hormogonia (a diagnostic characteristic of the genus *Nostoc*), has a complex life cycle, has established genetic transfer systems, and

is divergent from other cyanobacteria being sequenced. DNA from *N. punctiforme* is being prepared by Jack Meeks for submission to JGI for sequencing. The advisory panel will work with JGI during the annotation phase and will participate in publication of the data. The panel consists of Ronald M. Atlas (Department of Biology, University of Louisville), Jack Meeks (Division of Biological Sciences, University of California, Davis), Malcolm Potts (Department of Biochemistry and Nutrition, Virginia Polytechnic Institute); Jeff Elhai (Department of Biology, University of Richmond), and Theresa Thiel (Department of Biology, University of Missouri, St. Louis).

The Complete Genome Sequence of *Prochlorococcus*

Sallie W. Chisholm
Departments of Civil and Environmental Engineering and Biology; Massachusetts Institute of Technology; 15 Vassar St. 48-425; Cambridge, MA 02139
617/253-1771, Fax: /258-7009, *chisholm@mit.edu*
http://web.mit.edu/chisholm/www

Prochlorococcus is a unicellular cyanobacterium that is very abundant in the temperate and tropical oceans. It has been shown to contribute 32 to 80% of the total photosynthesis in the world's oligotrophic oceans, the higher values being found in the Pacific. Thus, *Prochlorococcus* plays a significant role in the global carbon cycle and the regulation of the earth's climate.

Molecular phylogenies have shown that *Prochlorococcus* is closely related to marine *Synechococcus*, forming a single lineage within the cyanobacteria. Unlike *Synechococcus*, *Prochlorococcus* lacks phycobilisomes and contains divinyl chlorophyll a (8-desethyl, 8-vinyl chlorophyll a, or "chl a2") and divinyl chlorophyll b (chl b2) as its major photosynthetic pigments. These pigments enable it to absorb blue light more efficiently than *Synechococcus* at the low-light intensities and blue wavelengths characteristic of the deep euphotic zone.

We recently demonstrated that there are at least two ecotypes of *Prochlorococcus*, each of which is distinguished by its photophysiology and molecular phylogeny. One is capable of growth at irradiances, and the other is not. We hypothesize that multiple ecotypes of *Prochlorococcus* coexist in all oceanic environments, alternating in dominance according to light gradients and seasonal mixing dynamics. We would expect to find, for example, that ecotypes adapted to low light are dominant at the base of the euphotic zone in stratified waters and those adapted to high light dominate at the surface. The ecotypes differ in other physiological properties besides light-harvesting efficiencies, and these too will play a role in regulating their distributions. Ultimately, a comparison of the complete genomes of these two ecotypes will provide valuable insights into the regulation of microdiversity in marine microbial systems.

Prochlorococcus is an ideal candidate for complete genome sequencing for a variety of reasons: (1) it is the smallest known phototroph and has a relatively

small genome size (1.8 Mb); (2) it is widespread and abundant and is easily identified and enumerated in its environment using flow cytometry; (3) its unique photosynthetic pigment (divinyl chlorophyll a) makes its contribution to total photosynthetic biomass in natural communities easily assessed; (4) different ecotypes have been identified that are very closely related according to their 16S rRNA sequences but are physiologically distinct; and (5) we have an extensive culture collection of isolates from different oceans and environments.

We plan to work with scientists at the DOE Joint Genome Institute (*http://bbrp.llnl.gov/jgi/microbial/prochlorococcus_homepage.htm*) to obtain the entire genomic sequence of *Prochlorococcus marinus* (*MED4*), one of the ecotypes adapted to high light. Our role in the project is to supply *Prochlorococcus* DNA and to be a general source of information on the ecology and biology of the organism.

Sequencing the Large Linear Chromosome of *Borellia burgdorferi* and a Strain of *Clostridium*

John J. Dunn and **F. William Studier**
Biology Department; Brookhaven National Laboratory; Bldg. 463, 50 Bell;
P.O. Box 5000; Upton, NY 11973-5000
631/344-3012, Fax: -3407, *jdunn@bnl.gov*
631/344-3390, Fax: -3407, *studier@bnl.gov*
www.genome.bnl.gov

In a program to explore possible improvements in the accuracy, speed, and efficiency of genome sequencing, we sequenced the large linear chromosome of *Borrelia burgdorferi*, the spirochete that causes Lyme disease. This 909,275-bp sequence is available on our Web site, along with a comparison of the same sequence determined independently by The Institute for Genomic Research (TIGR).

The Brookhaven National Laboratory (BNL) sequence was determined by random first-end and directed second-end sequencing of plasmid libraries of random chromosomal fragments, followed by primer walking using 12-mer primers generated by ligation of two hexamers on hexamer templates. The sequence assembly was confirmed and contigs were aligned by end sequencing a framework of ~35-kb fesmid clones, which spanned the entire sequence. The few remaining gaps were filled by polymerase chain reaction amplification from fesmid clones or genomic DNA. The sequence extends to the ends of the clones we obtained (which did not include the covalently closed ends of the chromosome) and lacks 404 bp at the left end and 249 bp at the right end of TIGR's sequence, which extends to the ends. The entire BNL sequence was determined at least once on each complementary strand.

The BNL and TIGR sequences are very similar, but there are some differences. The TIGR sequence contains seven copies of a 162-bp imperfect tandem repeat that occurs only twice in the BNL

sequence. There are 86 other discrepancies, only some of which are in a few remaining areas of relatively low quality in the BNL sequence. In addition, the BNL sequence contains 65 ambiguities (reflecting different base pairs at the same position in different clones), and the TIGR sequence contains 43 ambiguities. For each ambiguity in either sequence, one of the ambiguous bases matches the base at that position in the other sequence. It seems likely that each DNA preparation used for cloning and sequencing has polymorphisms at the 0.01% level, with a similar level of polymorphism between the two DNA preparations.

We are currently sequencing the genome of a *Clostridium* strain being studied at BNL as a possible bioremediation agent. This anaerobic, nitrogen-fixing spore former can convert water-soluble uranyl ion U(VI) to less soluble U(IV). Its circular genome is about 4 Mb, and no plasmids have been detected. More than 500 kb of edited unique sequence has been obtained so far. Clone libraries are being constructed in vectors we developed that allow an ordered set of nested deletions to be generated from either end of cloned fragments at least 10 kb long. These vectors were designed to allow sequencing and ordered assembly of both DNA strands in highly repeated regions such as those encountered in human DNA. In Clostridium, the vectors allow directed sequencing of particularly interesting areas by using nested deletions to fill in the framework generated by end sequencing. We expect to sequence the relevant U and N_2 reductases and identify most genes involved in intermediary metabolism.

This is a completed project.

DOE-Funded Microbial Genome Sequencing at The Institute for Genomic Research

Claire Fraser
The Institute for Genomic Research; 9712 Medical Center Dr.; Rockville, MD 20850
301/838-3500, Fax: -0209, *cfraser@tigr.org*
www.tigr.org

The Institute for Genomic Research (TIGR) is a not-for-profit research institute with interests in structural, functional, and comparative analysis of genomes and gene products in viruses, bacteria, archaea, and both plant and animal eukaryotes, including humans. Microbial genome-sequencing efforts at TIGR supported by the Department of Energy since 1995 have produced complete genome sequences for five organisms: *Mycoplasma genitalium*, *Methanococcus jannaschii*, *Archaeoglobus fulgidus*, *Thermotoga maritima*, and *Deinococcus radiodurans*. In addition, nine other DOE-funded microbial genome projects are in progress at TIGR, with an estimated completion date of 2001 for all work. In total, the DOE-funded microbial genome sequencing projects at TIGR represent nearly 33 million base pairs (Mb) of DNA and an estimated 30,000 microbial genes. The information generated in these projects is available from the TIGR Microbial Database (*www.tigr.org*).

The strategy that we use for whole-genome sequencing is called a "shotgun" method. In shotgun sequencing, the genome is sheared randomly into small pieces that are then cloned, sequenced, and reassembled to form a whole genomic sequence. With the shotgun approach, there is no need to develop a genetic or physical map of the genome before sequencing it; the sequence itself serves as the ultimate map. In large shotgun-sequencing projects, DNA fragments are assembled into a consensus sequence. Key to the success of the shotgun method is the availability of a truly random genomic DNA clone library and a powerful, accurate algorithm for reassembling the fragments into a complete genome. The basic approach for genome assembly is to compare all individual sequences to find overlaps and use this information to build a consensus sequence. Using new software developed at TIGR for large-scale genome sequencing projects, we have assembled the complete genomes of 12 microbial species to date.

The next step in whole-genome analysis is to identify all the predicted genes and search the translated protein sequences against protein sequences available in public databases. Because of the tremendous conservation in protein sequence among organisms throughout evolution, putative genes can be identified by sequence similarities.

The Minimal Gene Complement of *M. genitalium*

The *Mycoplasma* class consists of small wall-less bacteria that parasitize a wide range of hosts, including humans, animals, plants, insects, and cells in culture; they are believed to represent a minimalist life form, having yielded to selective pressure to reduce genome size and eliminate unnecessary genes. *M. genitalium* was selected as one of the first to be sequenced because it has the smallest genome of any known free-living organism. *M. genitalium* lives in a parasitic relationship with its primate hosts in ciliated epithelial cells of genitalia and respiratory tracts. Examining the makeup of the *M. genitalium* genome reveals much about the metabolic and biochemical capacity of this organism.

All genes necessary for life in *M. genitalium* are packaged in a 580,070-base (bp) circular chromosome. Genome analysis suggests that the *M. genitalium* genome contains about 470 genes (average size, 1040 bp), which make up 88% of the genome (on average, a gene every 1235 bp). This value is similar to that found in other microbial genome sequences. These data indicate that *Mycoplasma*'s reduction in genome size has not resulted in increased gene density or decreased gene size.

A complement of genes involved in DNA maintenance, repair, transcription, translation, and cellular transport is present; however, no complete pathways for amino acid, fatty acid, purine, or pyrimidine biosynthesis were identified in *M. genitalium*. Comparison of the minimal *M. genitalium* genome to that of more complex organisms suggests that differences in genome content are reflected as profound differences in physiology and metabolic capacity. The reduction in *M. genitalium*'s genome size is associated with a marked reduction in the number and components of biosynthetic pathways, thereby requiring the pathways to use metabolic products from their hosts.

Perhaps one of the most surprising findings from whole-genome sequencing and analysis of *M. genitalium* is that about one-third of the predicted proteins identified in this organism displayed no sequence similarity to known genes from any other organisms. This means that, even for this simplest of free-living organisms, we still do not understand a considerable amount of its biology. Determining whether the unknown genes in *M. genitalium* are species specific or exhibit a more widespread phylogenetic distribution will be of interest.

Comparing the *M. genitalium* genome with those of other microorganisms from diverse habitats will provide insights into what constitutes a minimal set of genes necessary for a self-replicating organism as well as the mechanisms associated with changes in genome organization and content in nature. This information, in turn, will be useful for modifying and engineering organisms to perform specific biochemical tasks in the laboratory or the environment.

Genome Sequence of the Archaeon *M. jannaschii*

The archaea were discovered as a unique phylogenetic domain of life by Carl Woese in the 1970s using sequence data from the small subunit of ribosomal RNA as a biosystematic marker. *M. jannaschii* was the first representative of the archaeal domain to be completely sequenced. Isolated in 1982 from a deep-sea hydrothermal vent, *M. jannaschii* fixes carbon dioxide to methane as its primary energy-producing biochemical pathway. Because this organism thrives at deep-sea pressures and temperatures of 85°C and above, its genome should provide insights into how genomes and gene products survive and function under these extreme conditions. Understanding the genetic basis of methanogenesis biochemistry in the thermophilic, barophilic *M. jannaschii* will bring us closer to harnessing the unique biochemistry of methanogens as a source of renewable energy.

Analysis of the *M. jannaschii* genome sequence reveals that between 50 and 60% of its genes or gene products have no match to any other currently known gene sequence. In addition, initial attempts to map database-matched genes onto known biochemical pathways suggest that *M. jannaschii's* biochemistry and physiology are quite unique among cellular organisms. For example, certain enzymes associated with gluconeogenesis and the synthesis of pentose sugars for nucleotide biosynthesis, such as fructose 1,6-biophosphate aldolase and fructose 1,6-biophosphate phosphatase, are not found among the predicted genes in *M. jannaschii.* Whether other gene products have been recruited to serve the function of these missing genes or the genes cannot be detected by standard sequence similarity methods is not yet known.

Most genes involved in *M. jannaschii's* cellular-information processing (replication, transcription, and translation) are more similar to functionally equivalent counterparts in eukaryotes, not bacteria. On the other hand, *M. jannaschii* genes that are involved in energy production, cell division, and basic cellular metabolism are more like genes in bacteria. Further analysis of the *M. jannaschii* genome sequence, together with sequence from other members of the archaeal domain of life, will give additional insights into the

evolutionary relationship among the prokaryotes.

Complete Sequence of the Thermophilic Archaeon *A. fulgidus*

Biological sulfate reduction is part of the global sulfur cycle, ubiquitous in the earth's anaerobic environments and essential to the workings of the biosphere. Growth by sulfate reduction is restricted to relatively few groups of prokaryotes; all but one of these is bacteria, the exception being the archaeal sulfate reducers in the archaeoglobales. These organisms are unique in that they are unrelated to other sulfate reducers and they grow at extremely high temperatures, between 60 and 95°C. They can grow both organoheterotrophically (using a variety of carbon and energy sources) or lithoautotrophically on hydrogen, thiosulfate, and carbon dioxide. The known archaeoglobales are strict anaerobes, most of which are hyperthermophilic marine sulfate reducers found in hydrothermal environments and in subsurface oil fields. High-temperature sulfate reduction by *Archaeoglobus* species contributes to deep subsurface oil well "souring" by producing iron sulfide, which causes corrosion of iron and steel in oil- and gas-processing systems.

The genome of the type-strain of the archaeoglobales *A. fulgidus* was sequenced to better understand the biology of this group of organisms. Genome analysis reveals a total of ~2400 genes; these include genes for sulfate reduction, a great diversity of electron transport systems, a large number of transporters with specificity for both organic and inorganic molecules, and β-oxidation of fatty acids. The information-processing systems and the biosynthetic pathways in *A. fulgidus* have counterparts in the archaeon *M. jannaschii.* However, the genomes of these two archaea indicate dramatic differences in the way these organisms sense their environment, perform regulatory and transport functions, and gain energy. Another interesting feature revealed by genome analysis is that *A. fulgidus* displays extensive gene duplication in comparison with other fully sequenced prokaryotes. This suggests that gene duplication has been an important evolutionary mechanism for increasing physiological diversity in the archaeoglobales.

About 25% of the *A. fulgidus* genome encodes conserved genes with unknown biological function, two-thirds of which are shared with *M. jannaschii.* Another 25% of the *A. fulgidus* genome represents genes that are unique to this organism, indicating that there is substantial diversity among members of the archaea. As additional archaeal and bacterial genome sequences are completed, we may begin to define a core set of genes that are shared among prokaryotes and those that are unique to bacterial or archaeal species.

Thermotoga maritima

The thermotogales are a group of nonsporeforming rod-shaped bacteria that represent the most thermophilic of the known organotrophic bacteria. The type strain *Thermotoga maritima MSB8*, isolated originally from geothermal-heated marine sediment at Vulcano, Italy, has an 80°C optimum temperature for growth. *T. maritima* metabolizes many simple and complex carbohydrates including glucose, sucrose, starch, xylan, and cellulose. Xylan is a complex plant polymer that represents the most abun-

dant noncellulosic polysaccharide in angiosperms, where it accounts for 20 to 30% of the dry weight of wood tissues. Cellulose is the most abundant biopolymer occurring in nature, estimated to account for 75×10^9 tons of dry plant biomass annually. Both cellulose and xylan, through conversion to fuels (e.g., H_2), have major potential as renewable carbon and energy sources.

T. maritima is of evolutionary significance because small subunit ribosomal RNA (SSU rRNA) phylogeny has placed the bacterium as one of the deepest and most slowly evolving bacteria. To further elucidate its unique metabolic properties and evolutionary relationship to other microbial species, we sequenced the genome of *T. maritima MSB8* using the whole-genome random sequencing method. The 1,860,725-bp *T. maritima* genome contains 1872 predicted coding regions, 54% (1005) of which have functional assignments and 46% (867) of which are of unknown function. Almost 7% of the predicted coding sequences in the *T. maritima* genome are involved in the metabolism of simple and complex sugars, a percentage more than twice that seen in other bacterial and archaeal species sequenced to date. Biosynthetic pathways for nine amino acids were identified in *T. maritima*, but the bacterium has an extensive system for the uptake of peptides from the environment.

Phylogenetic analysis of genes in the *T. maritima* genome has demonstrated that gene evolution may not give a true picture of organismal evolution; gene duplication, gene loss, and horizontal gene transfer probably account for many inconsistencies in single-gene phylogenies. The complete genome of *T. maritima* has, however, revealed a degree of similarity with the thermophilic archaea in terms of gene content and overall genome organization that was not previously appreciated. Of the sequenced bacteria, *T. maritima* has the highest percentage (24%) of genes that are most similar to archaeal genes. Some 81 of these genes are clustered in regions of the genome that range in size from 4 to 20 kb. Five of these regions have a composition substantially different from the rest of the genome, suggesting that lateral gene transfer has occurred between the thermophilic archaea and bacteria. In addition, repeat structures in *T. maritima* have been identified only in thermophiles, and 108 genes on the *T. maritima* genome have orthologues only in the genomes of other thermophilic bacteria and archaea. One explanation for the relatedness between thermophilic organisms seems to be the occurrence of lateral gene transfer.

Deinococcus radiodurans

Deinococcus radiodurans, originally discovered in food samples exposed to severe gamma irradiation, is the most radioresistant organism ever isolated. An important component of this resistance is the ability to repair damage to its own chromosomal DNA. D. *radiodurans* cultures exposed to 1.5 Mrad of radiation display a reduction in size of genomic DNA fragments corresponding to about 100 double-stranded breaks per genome. Typically, most prokaryotic and eukaryotic organisms cannot tolerate more than five double-stranded breaks per genome without reduced survival.

Within 8 to 10 hours after radiation exposure, the *D. radiodurans* genome is fully restored with no evidence of double-stranded breaks. During this repair time,

cellular replication of *D. radiodurans* is arrested; after this 8- to 10-hour interval, the cells display 100% survival with no detectable mutagenesis of their completely restored genome. DOE's interest in *D. radiodurans* includes understanding its ability to withstand radiation, particularly as it relates to the possibility of this organism's potential for bioremediation of toxic waste sites that contain radioactive isotopes.

The genome sequence of *D. radiodurans* is complete, and we have determined that the genome is composed of three chromosomes and a small plasmid. Inspection of the set of genes with similarity to DNA-repair enzymes has so far been inconclusive regarding radiation resistance; *D. radiodurans* does not appear to contain repair genes that would make it unique among other bacteria. However, a number of unique sequence elements have been identified that are being tested for their role in radiation resistance. These experiments, coupled with the high-throughput analysis of gene expression using microarray technology, should lead to a more complete understanding of this bacterium's gamma radiation resistance in the near future.

Shewanella putrefaciens: A Model Organism for Bioremediation

Shewanella putrefaciens is a bacterium involved in microbiologically influenced corrosion, anaerobic consumption of toxic organic pollutants, removal of toxic metals by sulfide precipitation, and removal of toxic metals and radionuclides by conversion to insoluble reduced forms. Whole-genome sequencing of *S. putrefaciens* will furnish the bioremediation community with detailed knowledge of metabolic pathways involved in all these processes, providing an excellent model system for manipulating organisms for remediation or control.

In addition, a complete genome sequence for *S. putrefaciens* will furnish important information on engineering specific regulatory mutants for bioremediation. For example, mutants that continue to metabolize anaerobically, even in the presence of oxygen, could be used to remove uranium (U^{6+}) in dilute environments where oxygen is still present. *S. putrefaciens* grows both aerobically and anaerobically. In its anaerobic phase, it acts as a metal reducer. The potential of metal-reducing bacteria in pollutant removal is very high for both the short and long terms, especially for those iron reducers that are not inhibited by oxygen.

Two separate reports suggest that *Shewanella* spp. can donate electrons to chlorinated hydrocarbons, thus reductively dechlorinating toxic compounds by converting tetrachloromethane to trichloromethane. In addition, organisms such as *S. putrefaciens*, which can produce Fe^{2+}, have potential to catalyze the reduction of toxic nitrates. Metals can be removed from solution via direct reduction by metal-reducing bacteria such as *S. putrefaciens*.

While iron and manganese are solubilized, other metals are converted to insoluble forms upon reduction. Of note are chromium (Cr^{6+}) and uranium (U^{6+}), both of which are soluble in the oxidized form but insoluble as the respective species reduced by Cr^{3+} and U^{3+}. Reduction of U^{6+} has been demonstrated for *S. putrefaciens* and has been proposed as a mechanism for concentrating and thus

removing radionuclide waste. As with uranium, the removal of toxic chromium should be possible using either intact cells or cell-free systems of the metal-reducing bacteria.

Complete genome sequences for all these metabolic processes would accelerate bioremediation efforts in metal and radionuclide reduction, chlorinated hydrocarbon pollutants, and toxic nitrates. We are midway through the closure process in the complete genome sequencing of *S. putrefaciens*. Random sequencing was completed in July 1998, and closure began in August 1998. Analysis of the assemblies suggests that the completed genome size will be about 5 Mb.

Preliminary observation of the gene content of this organism has shown similarities between *S. putrefaciens* and *Vibrio cholerae* in some role categories (small molecule biosynthesis, central intermediary metabolism) but differences in others (sugar metabolism). It will be interesting to examine these similarities and differences in light of the different ecological niches occupied by these organisms.

Chlorobium tepidum

The taxonomic group of green sulfur bacteria (Chlorobiaceae) are formally classified as Gram-negative organisms. Members of this genus are photoautotrophs that can generate chemical energy through an electron transport chain in the cytoplasmic membrane that is associated with a light-harvesting complex housed in a specialized organelle called the chlorosome. The components of this light-harvesting apparatus and some of its organizational structure are reminiscent of photosystems found in plant chloroplasts and, therefore, the evolutionary relationship of these prokaryotes to eukaryotic organelles is of interest. *Chlorobium* species also can fix CO_2, although the biochemical pathway used by these prokaryotes is distinct from the Calvin cycle found in higher plants.

C. tepidum initially was identified from a hot spring in New Zealand. This species is thermophilic with an optimum growth temperature of about 47°C. It has a genome size of 2.1 Mb with a G+C content of 56.5 mol%. *C. tepidum* was nominated for sequencing by DOE because of its photosynthetic capacity and its interesting phylogenetic position in the bacterial kingdom.

C. tepidum sequencing and closure has been completed. Genome annotation is under way and soon will be completed.

Caulobacter crescentus

Caulobacter crescentus is placed in the alpha-purple bacteria that also include *Rickettsia*, *Rhizobium*, *Agrobacterium*, and *Brucella* species. It is the most prevalent nonpathogenic bacterium in nutrient-poor freshwater streams. It is also found in marine environments. To facilitate location of nutrient sources, *C. crescentus* is motile and chemotactically competent during the swarmer phase of its life cycle. In its nonswarmer phase *Caulobacter* adheres to solid substrates such as rocks. It is a component of the organisms responsible for sewage treatment. *Caulobacters* are being modified for use as bioremediation agents for removing heavy metals from wastewater streams.

Caulobacter crescentus exhibits a well-studied developmental pattern, independent of environmental stress, with morphologically defined stages of the cell cycle. It has easily observable

physical structures that define these specific cell cycle stages. Two major events in *C. crescentus* cell cycle are used by researchers to elucidate fundamental processes required for development. These are the tight regulation of chromosomal replication and the temporally and spatially regulated biogenesis of the flagellum. The two processes are linked by a common transcriptional regulator that orchestrates the response of multiple cellular processes to the progression of the cell cycle.

The genome was electronically annotated at the end of the random sequencing phase; the data, along with the assembly files, was sent to Dr. Lucy Shapiro (Stanford University), Dr. Bert Ely (University of South Carolina), and Dr. Janine Maddock (University of Michigan), who are collaborating with us on final assembly and annotation of the genome. The project is now in the closure phase.

Pseudomonas putida

Sequencing of *Pseudomonas putida KT2440* began in January 1999 as a joint effort between TIGR and a German consortium consisting of groups from MHH (Medizinische Hochschule Hannover, Hannover, Germany); GBF (Gesellschaft für Biotechnologische Forschung mbH, Braunschweig, Germany); DKFZ (Deutsches Krebsforschungszentrum, Heidelberg, Germany); and QIAGEN (QIAGEN GmbH, Hilden, Germany). The study is supported by grants from BMBF of Germany and the U.S. Department of Energy.

The genome sequence will be used for in-depth functional analyses including comparisons of genome structure and function with the related organism *P. aeruginosa*. Understanding structure and function of the *P. putida* genome will allow for its increased use in biotechnological areas, including the production of natural compounds, remediation of polluted habitats, and the use of strains to fight plant diseases.

The *P. putida* genome sequence is expected to be closed in the next few months. The number of libraries for scaffolding the genome, access to the genome sequence of *P. aeruginosa*, and the complementary functional studies being conducted by the German consortium should reduce chances of major assembly problems in the genome.

Geobacter sulfurreducens

The complete genome sequence of *Geobacter sulfurreducens* is being determined to better understand its genetic potential. *G. sulfurreducens* is an important member of a family (Geobacteraceae) of delta proteobacteria capable of oxidizing organic compounds including aromatic hydrocarbons to carbon dioxide with Fe(III) or other metals and metalloids including U(VI), Tc(VII), Co(III), Cr(IV), Au(III), Hg(II), As(V) and Se(VII) serving as the terminal electron acceptor. It is the dominant group of iron-reducing microorganisms recovered from a wide variety of aquifer and subsurface environments when both molecular and traditional culturing techniques are used. *Geobacter* plays a critical role in the biogeochemical cycling of carbon, iron, and other metals. Its genetics and physiology are a subject of intense study in part due to the importance that these processes can play in the remediation of contaminated anaerobic subsurface environments. The determination of the *G. sulfurreducens* genome is being

accomplished using a random shotgun cloning approach to provide at least sixfold coverage of a 1-Mb genome followed by closure of remaining physical or sequence gaps. Searches of sequences and contigs from the early random phase of sequencing using the BLAST algorithm and database have produced high scores with low expect values indicating significant homologies to proteins contained in the database. These include enzymes considered important to basic housekeeping functions such as tRNA synthases and amino acid synthesis as well as those essential to other metabolic processes known to occur in *G. sulfurreducens* including nitrogen fixation. A number of sequences have produced no significant alignments indicating the likelihood of genes encoding for novel functions. Of further significance has been the extension of N-terminal sequences previously obtained from cytochromes known to be important in dissimilatory iron reduction. Thus, the genome will provide information crucial to the further understanding of this important metabolic process.

The Comprehensive Microbial Resource

One of the challenges presented by large-scale genome sequencing efforts is the effective display of information in a format that is accessible to the laboratory scientist. Conventional databases offer the scientist the means to search for a particular gene, sequence, or organism but do little to display the vast amounts of curated information that are becoming available. TIGR has developed methods to effectively "slice" the vast amounts of data in the sequencing databases in a wide variety of ways, allowing the user to formulate queries that search for specific genes as well as to investigate broader topics such as genes that might serve as vaccine and drug targets.

The Comprehensive Microbial Resource (CMR) is a facility for annotation of TIGR genome sequencing projects, a Web presentation of all fully sequenced microbial genomes, curation from the original sequencing centers, and further curation from TIGR (for those genomes sequenced outside TIGR). The Web presentation of CMR includes the comprehensive collection of bacterial genome sequences, curated information, and related informatics methodologies. The scientist can view genes within a genome and also can link to related genes in other genomes. This allows construction of queries that include sequence searches, isoelectric point, GC-content, GC-skew, functional role assignments, growth conditions, environment, and other questions and the isolation of genes of interest. The database contains extensive curated data as well as prerun homology searches to facilitate data mining. The interface allows the display of the results in numerous formats that will help the user ask more accurate questions. This resource should be of value to the scientific community to design experiments and spur further research. Resources of this type are an essential tool to make sense of bacterial genome information as the number of completed genomes continues to grow.

Rhodopseudomonas palustris Genome Project

Caroline S. Harwood
Department of Microbiology; University of Iowa; 3-432 Bowen Science Bldg.;
Iowa City, IA 52242
319/335-7783, Fax: -7679, *caroline-harwood@uiowa.edu*

Rhodopseudomonas palustris is a common soil and water bacterium that makes its living by converting sunlight to cellular energy and by absorbing atmospheric carbon dioxide and converting it to biomass. This microbe can also degrade and recycle components of the woody tissues of plants (wood is the most abundant polymer on earth). Because of its intimate involvement in carbon management and recycling, *R. palustris* has been selected by the DOE Carbon Management Program to have its genome sequenced by the Human Genome Program's Joint Genome Institute (JGI).

R. palustris is acknowledged by microbiologists to be one of the most metabolically versatile bacteria ever described. Not only can it convert carbon dioxide gas into cell material but nitrogen gas into ammonia, and it can produce hydrogen gas. It grows both in the absence and presence of oxygen. In the absence of oxygen, it prefers to generate all its energy from light by photosynthesis. It grows and increases its biomass by absorbing carbon dioxide, but it also can increase biomass by degrading organic compounds—including such toxic compounds as 3-chlorobenzoate—to cellular building blocks. When oxygen is present, *R. palustris* generates energy by degrading a variety of carbon-containing compounds (including sugars, lignin monomers, and methanol) and by carrying out respiration.

R. palustris undergoes two major developmental processes. The first is cell division by budding. This process of asymmetric cell division results in two different kinds of daughter cells—one a motile swarmer cell and the other a stalked nonmotile cell. The second is the differentiation of an elaborate system of intracytoplasmic membrane vesicles when cells run out of oxygen and are placed in light. The membranes are used to house photosynthetic pigments and associated proteins. Budding division and differentiation to photosynthetically competent cells both require a temporally regulated program of gene expression followed by a pattern of precise localization of protein products.

The diverse metabolism and the developmental cycles of *R. palustris* are a large part of what makes this bacterium such a seductive target for genome sequencing. With the entire genome sequence in hand, determining how *R. palustris* can coordinate and appropriately express its many metabolic capabilities in response to changing environmental conditions will be possible, as will devising strategies to maximize this bacterium's carbon-recycling capabilities.

R. palustris has a genetic system; genes can be moved in and out of this bacterium easily, and specific genes thus can be targeted for mutagenesis. This is of great value because it will allow researchers to rapidly apply information gained from genome sequencing to the developing area of functional genomics.

This work will supply the JGI with sufficient *R. palustris* genomic DNA for genome sequencing as well as any information needed about the biology of *R. palustris*.

Sequencing Microbial Genomes of Environmental Relevance

Jane E. Lamerdin
Joint Genome Institute; Lawrence Livermore National Laboratory; 7000 East Ave.; Livermore, CA 94550
925/423-3629, Fax: /422-2282, *lamerdinl@llnl.gov*
http://bbrp.llnl.gov/jgi/microbial

The DOE Joint Genome Institute (JGI) has established a new program to obtain the complete genome sequence of microorganisms that may significantly impact global climate. This program supports the new DOE Global Carbon Management and Sequestration initiative, which funds basic research aimed at understanding factors that contribute to global warming and effective ways to manage carbon (particularly carbon dioxide) in soil and ocean ecosystems. The goal of JGI's effort is to explore the role of diverse microorganisms in carbon cycling by elucidating their genetic content to identify metabolic pathways that allow these organisms to adapt to their respective niches. These specialized processes include nutrient-uptake systems, pathways that contribute to nitrogen fixation and carbon cycling in soils, and pathways that regulate photosynthesis. JGI's work is focused initially on five microorganisms: *Nitrosomonas europaea*, *Rhodopseudomonas palustris*, *Nostoc punctiforme*, and two marine cyanobacteria, *Prochlorococcus marinus* and *Synechococcus*. The common trait shared by these microbes is that all are autotrophic (i.e., they fix $C0_2$ as their sole carbon source), are fairly numerous within their respective ecosystems, and contribute materially to carbon cycling or biomass production (with the exception of *N. europaea*).

N. europaea is a soil-dwelling chemolithoautotroph that oxidizes ammonia to nitrite, a process that often depletes nitrogen available to plants, thereby limiting $C0_2$ fixation. Significantly, when oxygen concentrations in soils are low, *N. europaea* oxidizes nitrite to N_20, a catalyst of ozone breakdown and greenhouse gas production. We expect that the genome sequence of *N. europaea*, one of the few obligately autotrophic bacteria currently being sequenced, will allow us to catalog the identity and number of genes required for autotrophy. The genome sequence also should uncover special redox enzymes that allow *N. europaea* to adapt to the narrow niche it occupies.

R. palustris is a purple nonsulfur phototrophic bacterium commonly found in soils and fresh water. This species is of particular interest to the Carbon Management program because it is able to degrade and recycle components of woody tissues of plants (wood is the most abundant polymer on earth). It also possesses a large repertoire of metabolic capabilities, including the ability to fix $C0_2$ into cellular material, fix nitrogen gas into ammonia, and produce and use hydrogen gas. In the absence of oxygen, it grows phototrophically; in the presence of oxygen, it can generate energy by degrading sugars, organic acids, and methanol and can carry out respiration.

Nostoc punctiforme is a cyanobacterium that enters into symbiotic associations with fungi and lichens; these relationships are relevant to carbon cycling and sequestration in tundra. *Nostoc* species also have complex life cycles, fix nitrogen, and are capable of chromatic adaptation. *Prochlorococcus* and *Synechococcus* are unicellular picoplankton, which are major biomass producers in the world's temperate and tropical oceans. *Synechococcus* species are abundant in surface waters, while *Prochlorococcus* is found to exist in the layer 100 to 200 m deep. *Prochlorococcus* possesses an unorthodox pigment composition of divinyl derivatives of chlorophyll *a* and *b*, alpha carotene, zeaxanthin, and a type of phycoerythrin. The last has not yet been shown to function in light harvesting. By contrast, the highly related *Synechococcus* contains chlorophyll *a* and phycobilins that are more typical of cyanobacteria. *Prochlorococcus*, the only photosynthetic organism known to contain this particular combination of pigments, could be a model for the ancestral photosynthetic bacterium that gave rise to cyanobacteria and chloroplasts. Sequence analysis of the *Prochlorococcus* genome may shed more light on this hypothesis, and a comparison of the two genomes should provide additional insights into cyanobacterial radiation in general.

In part due to the lack of physical maps and mapping resources for these particular organisms, we have employed a whole-genome shotgun strategy to determine the complete sequence of each microbe. To aid our assembly, we are supplementing our six- to eightfold genome coverage in plasmid paired ends with a large-insert scaffold of paired ends in the low-copy-number fosmid vector. As the genome size increases (e.g., in *Nostoc*), we will shift to BAC clones for this scaffold. These scaffold clones are being fingerprinted to aid in verification of the final sequence assembly. We also will obtain optical maps of several of the larger organisms, *Nostoc* in particular, through a collaboration with David Schwartz at the University of Wisconsin.

JGI has completed the initial data-generation phase for *N. europaea* and *P. marinus*, which produced >95% of the genomic sequence for each microbe. (Progress towards completion can be monitored through our Web site: *http://bbrp.llnl.gov/jgi/microbial*.) A similar level of coverage is anticipated for *R. palustris* by mid-March. Finishing is under way on the first two organisms, and we expect closure of both by spring of 2000. With the level of coverage achieved by the initial data-generation phase, we can readily generate a rough inventory of the types of genes present in each organism. Preliminary or draft analyses have been performed on *N. europaea* and *P. marinus* by Frank Larimer and his team at Oak Ridge National Laboratory. The resulting catalog format provides user scientists with access to the contents of unfinished sequence data in a consumable format, without the need for protracted data manipulations on their part (for example, see *http://compbio.ornl.gov/~fwl/neur_files.html*). This allows them to focus on identifying gene products of particular interest to their research programs. The raw sequence data also are directly queryable through an accompanying BLAST server or can be downloaded from JGI's ftp server.

In summary, JGI's new microbial sequencing program is well under way, with at least three organisms on target to

be completed before the end of FY00. A scientific advisory board has assigned additional organisms for FY00 that continue the theme of relevance to the Global Carbon Management and Sequestration effort. We anticipate generating about 20 to 25 Mb of microbial genomic sequence in FY00 (initially in ~eightfold genome coverage) and ramping to a rate of 60 Mb in FY01.

See also the related abstracts of Ronald Atlas, Daniel Arp, David Schwartz, Caroline Harwood, Frank Larimer, and Sallie Chisholm.

The Genome of *Geobacter sulfurreducens*

B. A. Methe, Linda Banerjei,[1] William C. Nierman,[1] O. Snoeyenbos-West, S. Sciufo, and **Derek R. Lovley**
Department of Microbiology; University of Massachusetts; Amherst, MA 01003
413/545-9651, Fax: -1578, *dlovley@microbio.umass.edu*
[1]The Institute for Genomic Research; Rockville, MD 20850

The complete genome sequence of *Geobacter sulfurreducens* currently is being determined to better understand its genetic potential. *G. sulfurreducens* is an important member of a family (Geobacteraceae) of delta proteobacteria. This family is capable of oxidizing organic compounds including aromatic hydrocarbons to carbon dioxide with Fe(III) or other metals and metalloids including U(VI), Tc(VII), Co(III), Cr(IV), Au(III), Hg(II), As(V) and Se(VII) serving as the terminal electron acceptor. It is the dominant group of iron-reducing microorganisms recovered from a wide variety of aquifer and subsurface environments when both molecular and traditional culturing techniques are used. Geobacter plays a critical role in the biogeochemical cycling of carbon, iron, and other metals. Its genetics and physiology are a subject of intense study in part due to the importance that these processes can play in the remediation of contaminated anaerobic subsurface environments. The determination of the *G. sulfurreducens* genome is being accomplished using a random shotgun cloning approach to provide at least sixfold coverage of a 1-Mb genome followed by closure of remaining physical or sequence gaps. Assembler software and other computer programs developed by The Institute for Genomic Research are used to assemble the genome and aid in gap closing, finishing, and annotation. Searches of sequences and contigs from the early random phase of sequencing using the BLAST algorithm and database have produced high scores with low expect values indicating significant homologies to proteins contained in the database. These include enzymes considered important to basic housekeeping functions such as tRNA syntheses and amino acid synthesis as well as those essential to other metabolic processes known to occur in *G. sulfurreducens*, including nitrogen fixation. A number of sequences have produced no significant alignments, indicating the likelihood of genes encoding for novel functions. Of further significance has been the extension of N-terminal sequences previously obtained from cytochromes known to be important in dissimilatory iron reduction. Thus, the genome will provide information crucial to the further understanding of this important metabolic process.

Optical Approaches for Physical Mapping and Sequence Assembly of the *Deinococcus radiodurans* Chromosome

David C. Schwartz
Biotechnology Center; University of Wisconsin-Madison; 425 Henry Mall; Madison, WI 53706
608/265-0546, Fax: /262-6748, *dcschwartz@facstaff.wisc.edu*
www.chem.wisc.edu/~schwartz

Maps of genomic or cloned DNA frequently are constructed by analyzing the cleavage patterns produced by restriction enzymes. Restriction enzymes are remarkable reagents that consistently cleave only at specific four- to eight-nucleotide sequences, varying according to the specific enzymes.

Restriction enzymes are reliable, numerous, and easily obtainable, and there now are around 250 different sequences represented among thousands of enzymes. Restriction maps characterize gene structure and even entire genomes. Furthermore, such maps provide a useful scaffold for the alignment and verification of sequence data. Restriction maps generated by computer and predicted from the sequence are aligned with the actual restriction map.

Restriction enzyme action traditionally has been assayed by gel electrophoresis. This technique separates cleaved molecules on the basis of their mobilities under the influence of an applied electrical field within a gel-separation matrix (small fragments have a greater mobility than large ones). Although gel electrophoresis distinguishes different-sized DNA fragments (known as “fingerprinting”), the original order of these fragments remains unknown. The subsequent task of determining the order of such fragments is labor intensive, especially when making restriction maps of whole genomes, and, therefore, the procedure is not widely employed despite its obvious usefulness to genome analysis.

Our laboratory developed Optical Mapping, a system for the construction of ordered restriction maps from individual DNA molecules. The mapping substrate consisted of very large, randomly sheared genomic DNA fragments that were bound to derivatized glass surfaces and cleaved with the restriction enzyme *Nhe I*. The resulting fragments were imaged by fluorescence microscopy. Cut sites were visualized as gaps between cleaved DNA fragments that retained their original order. A whole-genome restriction map of *Deinococcus radiodurans*, a radiation-resistant bacterium able to survive up to 15,000 grays of ionizing radiation, was constructed without using DNA libraries, the polymerase chain reaction, or electrophoresis. Very large, randomly sheared, genomic DNA fragments were used to construct maps from individual DNA molecules that were assembled into two circular overlapping maps (2.6 and 0.415 Mb), without gaps. A third smaller chromosome (176 kb) was identified and characterized. Aberrant nonlinear DNA structures that may define chromosome structure and organization, as well as intermediates in DNA repair, were visualized directly by optical mapping techniques after irradiation.

This high-resolution restriction map was used by collaborators at The Institute

for Genomic Research to verify sequence-assembly data from *D. radiodurans* by aligning the restriction map predicted from their sequence. Optical mapping of *D. radiodurans* also rendered insights into the organism's biology by providing a picture of the entire genome's basic organization. The genome was shown to be composed of two rather than one chromosome, and the presence of other extrachromosomal elements was demonstrated.

Whole-genome characterization by optical mapping may facilitate further understanding of the radiation-resistant nature of *D. radiodurans*, which is being used as a vehicle for bioremediation of toxic organic pollutants within radioactive waste dumps.

Whole-Genome Sequence of *Pyrobaculum aerophilum*

Melvin I. Simon and **Sorel Fitz-Gibbon**
Biology Division; California Institute of Technology; 1200 E. California Blvd.; Pasadena, CA 91125
626/395-3944, Fax: /796-7066, *simonm@starbase1.caltech.edu*
www.tree.caltech.edu

Pyrobaculum aerophilum was chosen as a model organism for the study of hyperthermophiles and archaea. This rod-shaped microbe, isolated from a boiling marine vent, has a maximum growth temperature of 104°C, not far from the 113°C maximum known for all life. Unlike most hyperthermophiles, however, *P. aerophilum* is able to withstand exposure to oxygen and thus is amenable to experimental manipulations on the laboratory benchtop. In addition to being an ideal model-organism candidate, *P. aerophilum* warrants further studies because of its phylogenetic position as a member of the crenarchaea-eocytes, which may be the eukaryotes' closest prokaryotic relatives.

The entire *P. aerophilum* genome has been sequenced using a random shotgun approach (3.5× genomic coverage) followed by oligonucleotide primer-directed sequencing guided by our fosmid map. The genome was assembled and edited using the Phred-Phrap-Consed system. The 2.2-Mb genome codes for about 2500 proteins, 30% of which have been identified by sequence similarities to proteins of known function. We have made extensive use of the MAGPIE software for genome annotation and GeneMark and Glimmer for prediction of coding regions. In completing the "polishing" of the genome, we are nearing our goal of no more than 1 error in 10,000 bases. We also are continuing to annotate the genome and attempting to improve our functional predictions by using information on conserved residues, potential 3-D structure alignments, and gene phylogenies.

In our publications early in 1999, we discussed in detail the results of the annotation process. One interesting set of results pertains to genes involved in DNA repair. Two major mechanisms for avoiding mutations during DNA replication are the DNA polymerase's immediate editing of the growing strand and the mismatch-repair system's detection and correction of mismatches soon after replication.

Homologs of the *Escherichia coli* proteins involved in mismatch repair have been found in humans, and damage to them has been implicated in hereditary nonpolyposis colon cancer. However, homologs to mismatch-repair proteins have not been detected in the *P. aerophilum* genome nor in any of the other three completed archaeal genomes. It remains to be seen whether mismatch-repair activities can be detected in these organisms, and, if so, whether different enzymes have been recruited for these functions or the archaeal homologs have diverged too much to be recognized by simple sequence comparisons.

Having the entire genome sequence is an extraordinary tool for research on this organism, and numerous downstream projects already are in progress. The genome sequence has been invaluable in guiding work to develop a laboratory research system that would allow such *E. coli*-like experiments as gene knockouts and homologous overexpression of archaeal proteins. The *P. aerophilum* genome-proteome also is being used by several laboratories worldwide to develop methods for high-throughput 3-D structure determination. Proteins from thermophiles appear to be more stable than their mesophilic homologs and may have higher rates of successful crystallization, thus simplifying the development of high-throughput "structural proteomics."

Completion of microbial genome sequences provides not only a wealth of information on individual species but also allows implementation of new methods for deciphering genomes. For example, it is now possible to predict functionally linked proteins simply by looking for the presence or absence of similar distribution patterns among completed genomes. With perhaps half the proteins in microbial genomes having no clear functional assignments, a good deal of exciting work remains to be done.

This is a completed project.

Whole-Genome Shotgun Sequencing

Douglas Smith
Genome and Technology Development; Genome Therapeutics Corp.; 100 Beaver St.; Waltham, MA 02154-8440
781/398-2378 or /893-5007 (ext. 219), Fax: /893-9535 or /642-0310,
doug.smith@genomecorp.com
www.genomecorp.com

The information in the chromosome of a bacterium (or any other organism) is encoded in the specific sequence of four chemical building blocks called nucleotides. Millions of these nucleotides are polymerized into long strands that stick together in pairs to form the DNA double helix. Genes are encoded in the DNA by specific sequences of nucleotides, much as the words in this paragraph are encoded by sequences of letters. Bacterial chromosomes typically contain 1 to 7 million nucleotide pairs (abbreviated Mb).

Current biochemical methods for determining DNA sequences generate "reads" of about 500 to 700 nucleotides. To sequence an entire bacterial genome,

therefore, a method is needed for accurately piecing together lots of individual reads. To accomplish this, we use a "whole-genome shotgun" approach in which thousands of sequence reads (enough to span a whole genome 7 to 8 times) are generated from random locations in the genome. Using powerful computer programs, investigators then assemble these sequences into overlapping sets that, together with additional information, can be joined to reassemble the entire chromosome.

Methanobacterium thermoautotrophicum

This organism is a member of the archaea, one of the three major kingdoms into which all living things can be classified (the other two are bacteria, which include most of the familiar disease-causing organisms; and eucarya, which include protozoa, fungi, plants, animals, and humans). Archaea are interesting because many of their cellular processes are similar to those of eucarya, while others are more closely related to bacteria.

M. thermoautotropicum, originally isolated from sewage sludge, also is found in the manure of farm animals. In combination with other organisms, *M. thermoautotrophicum* can be used to produce methane from such materials. The organism prefers growth temperatures of about 65°C and is capable of growing and producing methane in the presence of only hydrogen, carbon dioxide, and a few salts. The complete genome sequence provides information that could be used to reengineer the organism to grow more rapidly and to produce larger amounts of methane with fewer by-products. The thermostable proteins may be useful in the chemical industry as reagents for bioconversion or biocatalysis.

Using the whole-genome shotgun approach, we completed the sequence of the entire 1.75-Mb genome of *M. thermoautotrophicum* during 1997. In the shotgun phase, we generated over 36,000 sequence reads (about 13 Mb, or 7.5-fold genome coverage). The reads were assembled, and the resulting sets of overlapping fragments were joined together by using a "primer-walking" technique to generate new sequences extending from the ends of the contigs. Additional biochemical tools and computer programs were used to identify and fix misassembled regions and to confirm the links between the assembled sequences, allowing us to reconstruct the entire circular chromosome.

The resulting sequence was analyzed to identify the encoded genes. Many *M. thermoautotrophicum* genes encode proteins that are more closely related to eucaryal proteins (from higher plants and animals) than to bacterial ones. This is especially true of components involved in transcription and translation, processes by which gene sequences are "expressed" to produce protein products in the cell. Comparisons to the genome of *Methanococcus jannaschii* (another archaeon) revealed many similarities but also many differences. Both organisms contain a significant number of unique genes that are unrelated to any other known genes. This finding underscores the high degree of complexity and genetic diversity present in the biological universe.

Clostridium acetobutylicum

Continued Microbial Genome Program work in our laboratory focused

on the gram-positive, spore-forming bacterium *C. acetobutylicum ATCC 824*. Its 4.1-Mb genome, reflecting its more complex life processes and metabolism, is more than twice the size of *Methanobacterium*. The organism is related to the pathogenic species *C. botulinum*, *C. tetani*, and *C. perfringens*, which cause the diseases botulism, tetanus, and gangrene, respectively.

Isolates of *C. acetobutylicum* were identified before the First World War when rubber shortages stimulated a search for microbes that could produce butanol for synthetic rubber production. Chaim Weizmann (who later became the first president of Israel) developed a process for ABE fermentation (to produce acetone, butanol, and ethanol) using *C. acetobutylicum* and plant starch that was later pursued commercially. Demand for acetone during the Second World War led to the establishment of a molasses-based ABE process, but increases in the cost of molasses, together with advances in the petrochemical industry, led to its eventual abandonment.

Since that time, scientific interest in the solvent-producing *Clostridia* has continued. A great deal of work has been done to elucidate the metabolic pathways by which solvents are produced. Many solvent-overproducing derivatives (strains) have been identified, and it is now possible to pursue a rational approach to develop modified strains with industrially useful properties. Experimental research systems have been developed that allow genes to be manipulated in these organisms, and strains have been altered to grow on cellulose constituents that will not support the growth of natural strains. The complete genome sequence will be immensely useful in further development of these organisms as natural bioconversion factories for the chemical and fuel industries.

C. acetobutylicum ATCC 824 was sequenced by the whole-genome shotgun approach, essentially as described above but including several technological advances. The finishing phase involved exhaustive gap closure and quality enhancement using a variety of biochemical methods and computational tools. Only a few gaps remain, and a publication describing the work is expected during 1999.

The genome sequences of *M. thermoautotrophicum* and *C. acetobutylicum* are freely available in public databases, enabling research scientists throughout the world to access the information to expedite the development of useful derivatives of these and other organisms.

This is a completed project.

The Complete Genome of the Hyperthermophilic Bacterium *Aquifex aeolicus*

Ronald Swanson
Diversa Corporation; 10665 Sorrento Valley Road; San Diego, CA 92121
619/623-5156, Fax: -5120, *rswanson@diversa.com*
www. diversa.com

Diversa Corporation has completed the genome sequence of the most heat tolerant bacterium currently known. This organism, *Aquifex aeolicus*, is capable of growing at up to 95°C (203°F). Isolated and described only recently, *Aquifex* is related to filamentous bacteria first observed at the turn of the century, growing at 89°C in the outflow of hot springs in Yellowstone National Park. Observation of these macroscopic assemblages would later be instrumental in the drive to culture hyperthermophilic organisms.

Aquifex is able to grow on hydrogen, oxygen, carbon dioxide, and simple mineral salts. The complex metabolic machinery necessary to function as a hyperthermophilic chemolithoautotroph is encoded within a 1,551,335-bp genome only one-third the size of *Escherichia coli*; this small size appears to limit metabolic flexibility. The use of oxygen as an electron acceptor is enabled by the presence of a complex respiratory apparatus. Despite the fact that this organism grows at bacteria's extreme thermal limit, only a few specific indications of thermophily are apparent from the genome.

One of the most exciting results of sequence analysis is the lack of coherence in the apparent phylogenies of different genes. It was widely anticipated that, because of the small subunit ribosomal RNA gene's branching position near the bacterial lineage's root, *Aquifex* gene analysis would shed light on the phenotype of bacteria's last common ancestor, including the bacterial domain's hypothesized thermophilic origin. However, protein-based phylogenies do not in many cases support the original rRNA-based placement and show no consistent picture of the organism's phylogeny. This result has fundamental implications for our understanding of the evolutionary mode.

The sequencing strategy used to assemble the complete genome was based on the whole-genome shotgun approach. Shotgun sequencing is characterized by two phases: an initial, completely random phase in which most data are collected, and a closure phase in which directed techniques are used to close gaps and complete the assembly. By pursuing a strategy in which only 97% coverage was achieved initially, we were able to limit the number of random-phase sequences to only 10,500. Sequence fragments were assembled on an Apple Macintosh computer using Sequencher, a commercially available assembly and editing program. Sequences were obtained from both ends of clones randomly chosen from a fosmid library; using Sequencher, we assembled these sequences with consensus sequences derived from the contigs of random-phase sequences. Gaps between contigs were closed by direct sequencing on fosmids not wholly contained within a

contig. The final assembly comprises 13,785 sequences with an average edited read length of 557 bp.

More than half of *Aquifex*'s 1512 open reading frames were assigned a putative function based on similarity to known sequences. The extreme thermostability of *Aquifex* proteins, coupled with their bacterial origins, makes them ideal candidates for over expression, nuclear magnetic resonance imaging, and X-ray crystallographic studies. Consequently, large numbers of researchers are pursuing structures of the thermostable *Aquifex* proteins, and several heterologously expressed proteins are being evaluated in commercial applications.

This is a completed project.

The Genome Sequence of the Hyperthermophilic Archaeon *Pyrococcus furiosus*

Robert B. Weiss, Frank Robb,[1] and James R. Brown[2]
Human Genetics Dept.; Eccles Institute of Human Genetics; 20 South 2030 East, Room 308 BPRB; University of Utah; Salt Lake City, UT 84112-5330
801/585-3435 or -5606, Fax: -7177, *bob.weiss@genetics.utah.edu* or *bob@watneys.med.utah.edu*
[1]Center of Marine Biotechnology; University of Maryland
[2]Microbial Bioinformatics Group; SmithKline Beecham Pharmaceuticals
http://www-genetics.med.utah.edu/index.html
http://comb5-156.umbi.umd.edu

Pyrococcus furiosus is the best-studied member of the unusual class of organisms known as extreme hyperthermophiles because they live at extremes of temperature and pressure. Isolated from geothermally heated marine sediment in the shallow waters off Vulcano Island, Italy, *P. furiosus* grows optimally at 100°C and derives its energy by fermentation of protein, peptide, and sugar mixtures found in its geothermal environment. The organism is fast growing and capable of dividing every 40 min.

Extreme hyperthermophiles play an important role in advancing the fundamental understanding of protein biochemistry, RNA and DNA metabolism, and protein interactions. How has a cell's macromolecular machinery adapted to function at 100°C? Proteins from organisms living at moderate temperatures unfold or denature when heated, but proteins from hyperthermophiles maintain their three-dimensional shapes. The genome sequence provides a resource for beginning to understand why this happens.

Extremely stable proteins have potential biotechnological uses as rugged industrial catalysts. The diverse metabolism of *P. furiosus* provides a wide variety of biocatalysts that are potentially useful as environmentally safe reagents in transforming biomass to derive energy and specialty chemicals and in degrading organic compounds for environmental detoxification.

The *P. furiosus* genome sequence was completed recently. Its circular chromosome is 1,908,253 bp long with a G-C content of 40.8%. The sequencing

strategy tested a variant of whole-genome shotgun sequencing with a new sequencing vector that allows the genome to be subcloned as larger pieces. The genome was pieced together from fewer than 2500 subclones, compared to the more typical number of 20,000. These medium-insert sequencing vectors may help to assemble the larger human and mouse genomes.

Genome analysis and annotation are ongoing. Recently, the complete sequence of the distantly related *P. horikoshii*, which was isolated from a hydrothermal vent at a depth of 1395 m in the Sea of Japan, was determined by a group in Japan (*www.bio.nite.go.jp*). *P. furiosus* and *P. horikoshii* diverged over 100 million years ago, and comparisons between them are providing unique insight into processes that result in changes to genes and genomes by revealing complex gene rearrangements and changes in gene content.

The sequence was completed in late November 1998, and the annotation phase was completed early in 1999. The sequence is available for searching and downloading from the Web (*www.genome.utah.edu*). Library construction, sequencing, and assembly and the production of finished sequence was done at the University of Utah. Dr. Frank Robb's group provided the organism and has assisted in the finishing and annotation stages. Dr. Brown's group is assisting in the gene-finding and annotation stages of the project.

This is a completed project.

Pangenomic Microbial Comparisons by Subtractive Hybridization

Peter Agron, Lyndsay Radnedge, Evan Skowronski, Madison Macht, Jessica Wollard, Sylvia Chin, Aubree Hubbell, Marilyn Seymour, Christina Nocerino, and
Gary Andersen
Biology and Biotechnology Research Program; Lawrence Livermore National Laboratory; 7000 East Ave.; Livermore, CA 94550
Andersen: 925/423-2525, Fax: /422-2282, *andersen2@llnl.gov*

Sequencing of whole genomes is reshaping microbiology. However, as more sequence information is generated, there will be increased sequence redundancy between closely related species or strains. In the course of time, the amount of new sequence information obtained by whole-genome sequencing with current technology will become increasingly less cost-efficient. We are exploring the use of suppression subtractive hybridization (SSH) of total DNA as a means of focusing sequencing efforts on unique regions when a reference strain of known sequence is compared to a different isolate of the same species or genus. To rigorously examine this approach, two sequenced strains of *Helicobacter pylori* (*J99* and *26695*) were used as a model system to allow rapid determination and mapping of difference products based on sequencing alone.

Using high-throughput SSH methods, difference products can be rapidly cloned, sequenced, and then mapped by comparing the data to the *H. pylori* genome database. To increase the likelihood of amplifying difference products from any given region, several restriction enzymes were used in separate SSH experiments. We have obtained data from 2123 clones that reveal 427 (20%) unique sequences. Control subtractions with an *Escherichia coli* strain containing the transposon Tn5 against its isogenic parent showed a 270-fold enrichment for Tn5 sequences, demonstrating that SSH is highly effective. Current efforts are focused on (1) mapping difference products onto the relevant genome using the cross-match algorithm and Percent Identity Plots, (2) assessing coverage of difference regions by subtracted clones, (3) assessing the redundancy of this coverage, and (4) determining the reproducibility of SSH.

The Genome of the Extremely Radioresistant Bacterium *Deinococcus radiodurans*: Comparative Genomics

Kira S. Makarova,[1,2] Eugene V. Koonin,[3] L. Aravind,[2] Kenneth W. Minton,[1]
Roman L. Tatusov,[2] Y. I. Wolf,[2] Owen White,[3] and **Michael J. Daly**[1]
[1]Uniformed Services University of the Health Sciences; 4301 James Bridge Rd.;
Bethesda, MD 20814-4799
301/295-3750, Fax: -1640, *mdaly@mxb.usuhs.mil*
[2]National Center for Biotechnology Information; National Institutes of Health;
Bethesda, MD 20814
[3]The Institute for Genomic Research; Rockville, MD 20850

Extremophiles are nearly always defined with singular characteristics that allow existence within a singular extreme environment. The bacterium *Deinococcus radiodurans* qualifies as a polyextremophile, showing remarkable resistance to a range of damage caused by ionizing radiation, dessication, ultraviolet radiation, oxidizing agents, and electrophilic mutagens. *D. radiodurans* is most famous for its extreme resistance to ionizing radiation; it not only can grow continuously in the presence of chronic radiation (6000 rad/hour), but it can survive acute exposures to gamma radiation that exceed 1.5 Mrad without lethality or induced mutation. These characteristics were the impetus for sequencing its genome and the ongoing development of its use for bioremediation of radioactive wastes.

Although it is known that these myriad resistance phenotypes stem from its efficient DNA repair processes, the mechanisms underlying this repair remain poorly understood. In this work we present an extensive comparative sequence analysis of the *Deinococcus* genome. *Deinococcus* is the first representative with a completely sequenced genome from a bacterial branch of extremophiles—the *Thermus-Deinococcus* group. Phylogenetic tree analysis, combined with the identification of several synapomorphies between *Thermus* and *Deinococcus*, support that it is a very ancient branch localized in the vicinity of the bacterial tree root. Distinctive features of the *Deinoccoccus* genome, as well as features shared with other free-living bacteria, were revealed by comparing its proteome to a collection of clusters of orthologous groups of proteins (called COGs). Analysis of paralogs in *Deinococcus* has revealed some unique protein families. In addition, specific expansions of several protein families including phosphatases, proteases, acyl transferases, and MutT pyrophosphohydrolases were detected. Genes that potentially affect DNA repair and recombination were investigated in detail.

Some proteins appear to have been transferred horizontally from eukaryotes and are not present in other bacteria. For example, three proteins homologous to plant desiccation-resistance proteins were identified; these are particularly interesting because of the positive correlations of resistance to desiccation and radiation. Further, the *D. radiodurans* genome is very rich in repetitive sequences, namely IS-like transposons and small intergenic repeats. In combination, these observations suggest that several different biological mechanisms contribute to the multiple DNA repair-dependent phenotypes of this organism. The genetic

mechanisms underlying the extreme radiation resistance of the organism are now being characterized experimentally using a newly developed system for analyzing gene expression patterns in *D. radiodurans*.

Protein Expression in *Methanococccus jannaschii* and *Pyrococcus furiosus*

Carol S. Giometti, S. L. Tollaksen, H. Lim,[1] J. Yates,[1] J. Holden,[2] A. Lal Menon,[2] G. Schut,[2] M. W. W. Adams,[2] C. Reich,[3] and G. Olsen[3]
Center for Mechanistic Biology and Biotechnology; Argonne National Laboratory; 9700 S. Cass Ave.; Argonne, IL 60439
630/252-3839, Fax: -5517, *csgiometti@anl.gov*
[1]University of Washington; Seattle, WA 98195
[2]University of Georgia; Athens, GA 30602
[3]University of Illinois; Urbana, IL 61801

Complete genome sequences are now available for both *Methanococcus jannaschii* and *Pyrococcus furiosus*. The open reading frame (ORF) sequences from these completed genomes can be used to predict the proteins synthesized, but laboratory methods are needed to verify those predictions. Two-dimensional gel electrophoresis (2DE), coupled with mass spectrometry of peptides isolated from the gels, is being used to determine the constitutive expression of proteins from these two archaea and to explore the regulation of expression of nonconstitutive proteins. The most abundant proteins (i.e., those easily detectable by staining with Coomassie Blue R250) have been isolated and analyzed from cells grown in minimal nutrient media. Using a combination of matrix-assisted laser desorption ionization (MALDI) and tandem mass spectrometry, 100 proteins expressed by *M. jannaschii* and 50 proteins expressed by *P. furiosus* have been related to specific ORFs in the respective genome sequences. The molecular weights and isoelectric points determined by protein positions in the 2DE patterns are compared with the ORF-predicted molecular weights and isoelectric points for each microbe. Numerous instances have been observed of multiple proteins with different molecular weights or isoelectric points being associated with the same ORF. Possible reasons for such multiplicity include the incomplete unfolding of these highly stable proteins prior to electrophoresis, the nondissociation of subunits, posttranslational modifications such as phosphorylation (multiple proteins with the same identity but different isoelectric points), or peptide cleavage (multiple proteins with the same identity but different molecular weights). Preliminary experiments to change the protein expression of these organisms by altering growth conditions have revealed significant quantitative changes in a small number of proteins visible in 2DE patterns. Correlation of proteins expressed with specific ORFs is now focused on proteins showing quantitative changes in expression and on less abundant proteins. The observed protein abundances and changes in abundance from these proteomic studies could be useful for validation of protein-expression predictions based on ORFs.

Microbial Genome Annotation and Display

Frank Larimer, Doug Hyatt, Miriam Land, Richard Mural, Morey Parang, Manesh Shah, Jay Snoddy, and **Edward Uberbacher**
Computational Biosciences; Life Sciences Division; Oak Ridge National Laboratory; 1060 Commerce Park Dr.; Oak Ridge, TN 37830
865/574-1253, Fax: /241-1965, *larimerfw@ornl.gov*
http://compbio.ornl.gov

Once the genome of an organism has been sequenced, portions that define features of biological importance must be identified and annotated. When the newly identified gene has a close relative already in DNA or protein sequence databases, gene finding in microorganisms is relatively straightforward. The genes tend to be simple, uninterrupted open reading frames (ORFs) that can be translated and compared with the database.

The discovery of new genes without close relatives is more problematic. Although identifying gene-like ORFs is easy, it is very difficult to determine which represent real genes and which are merely statistical artifacts of the sequence. This is a serious problem in organisms with a high G+C content where random ORFs can be abundant due a lack of stop codons.

A second issue in modeling microbial genes is accurate prediction of the start codon, which is complicated further by the use of minor start codons in addition to the universal AUG. An accurate accounting and description of genes in microbial genomes is essential in determining the existence of functional metabolic pathways and other aspects of whole-organism function. Compared to simpler gene-prediction methods using ORFs or single-coding measures, recently developed gene-finding systems show excellent performance in predicting coding genes and start sites, even for the shortest microbial genes. Such highly accurate systems are effective across the phylogenetic spectrum of organisms as an essential baseline of analysis from which much biological insight can be obtained.

Microbial genome sequencing is progressing rapidly. Apart from the twenty-odd published genomes, more than 100 are being sequenced, with plans to sequence hundreds or thousands more. Since every new genome informs those that preceded it, updating genome annotation is necessary to keep these resources relevant; and consistent procedures, tools, and methodology must be applied. The unique functions of each individual organism need to be documented as functions are placed in a recognized, consistent scheme.

We are now representing all completed microbial genomes in the Genome Channel and the Genome Catalog, providing comprehensive sequence-based views of genomes from a full genome display to the nucleotide sequence level. We have developed tools for comparative multiple-genome analysis that provide automated, regularly updated, comprehensive annotation of microbial genomes using consistent methodology for gene calling and feature recognition. The visual genome browser represents around 51,000 microbial GRAIL and 45,000 GenBank gene models. Precomputed BEAUTY searches are provided for all

gene models, with links to original source material and additional search engines. Comprehensive representation of microbial genomes will require deeper annotation of structural features, including operon and regulon organization, promoter and ribosome binding-site recognition, repressor and activator binding-site calling, transcription terminators, and other functional elements. Sensor development is in progress to provide access to these features. Linkage and integration of the gene-protein-function catalog to phylogenetic, structural, and metabolic relationships also will be developed.

A draft analysis pipeline has been constructed to provide annotation for the Microbial Genome Program of the Joint Genome Institute. The first two draft sequences in the pipeline, with many more to come, are the *Nitrosomonas europaea* and *Prochlorococcus marinus* genomes. Multiple gene callers (Generation, Glimmer, and Critica) are used to generate a candidate gene model set. The conceptual translations of these gene models generate similarity-search results and protein family relationships; from these results, a metabolic framework is constructed and functional roles are assigned. Simple and complex repeats, tRNA genes, and other structural RNA genes also are identified. Annotation summaries are available through the JGI microbial genomics Web site (*http://bbrp.llnl.gov/jgi/microbial*); in addition, draft results are being integrated into the interactive display schemes of the Genome Channel and Catalog (*http://genome.ornl.gov*).

WIT2: An Integrated System for Genetic Sequence Analysis and Metabolic Reconstruction

Ross Overbeek,[1,2] Gordon Pusch,[1,2] Mark D'Souza,[1] Evgeni Selkov Jr.,[1,2] Evgeni Selkov,[1,2] and **Natalia Maltsev**[1]
[1]Mathematics and Computer Science Division; Argonne National Laboratory, MCS-221; 9700 S. Cass Ave.; Argonne, IL 60439
[2]Integrated Genomics Inc.
Maltsev: 630/252-5195, Fax: -5986, *maltsev@mcs.anl.gov*
http://wit.mcs.anl.gov/WIT2

The WIT2 system was designed and implemented to support genetic sequence and comparative analysis of sequenced genomes as well as metabolic reconstructions from the sequence data. It now contains data from 38 distinct genomes. WIT2 provides access to thoroughly annotated genomes within a framework of metabolic reconstructions connected to the sequence data; protein alignments and phylogenetic trees; and data on gene clusters, potential operons, and functional domains. We believe that the parallel analysis of a large number of phylogenetically diverse genomes can add a great deal to our understanding of the higher-level functional subsystems and physiology of the organisms. The unique features of WIT2 include the following: (1) WIT2 is based on the unique EMP-MPW collection of enzymes and metabolic pathways developed by

E. Selkov and colleagues; this collection contains extensive information on enzymology and metabolism of different organisms. (2) WIT2 allows researchers to perform interactive genetic sequence analysis within a framework of metabolic reconstructions and to maintain user models of the organism's functionality. (3) WIT2 provides access to a set of Web-based and original batch tools that offer extensible query access against the data. (4) WIT2 supports both shared and nonshared annotation of features and the maintenance of multiple models of the metabolism for each organism. (5) WIT2 supports metabolic reconstructions from expressed sequence tag data.

Microbial Proteomics at Pacific Northwest National Laboratory

Richard D. Smith, Ljiljana Pasa-Tolic, Mary S. Lipton, Pamela K. Jensen, Gordon A. Anderson, and Timothy D. Veenstra
Environmental Molecular Sciences Laboratory, MS K8-98; Pacific Northwest National Laboratory; P.O. Box 999; Richland, WA 993522
509/376-0723, Fax: -7722, *dick.smith@pnl.gov*

Bacterial strains such as *Shewanella putrefaciens MR-1* are key organisms in the bioremediation of metals due to their ability to enzymatically reduce and precipitate a diverse range of heavy metals and radionuclides. Additionally, *Deinococcus radiodurans* is an attractive candidate for bioremediation because of its unique ability to survive exceedingly high doses of ionizing radiation. The need to develop an improved understanding of their enzymatic pathways is important in refining the unique capabilities of these organisms for bioremediation. As a first step, an organism's proteome must be characterized completely. The proteome is the name given to the dynamic array of proteins expressed by a genome. A single genome can exhibit many different proteomes depending on the stage in the cell cycle; cell differentiation; response to such environmental conditions as nutrients, temperature, and stress; and the manifestation of disease states. Although the availability of full genomic reference sequences provides a set of road maps of possibilities and the measurement of expressed RNAs tells us what might happen, the proteome is the key that tells us what really happens. Therefore, the study of proteomes under well-defined conditions can provide a better understanding of complex biological processes, requiring faster and more sensitive capabilities for the characterization of microbial protein constituents.

We currently are developing technologies that integrate and refine protein separation and digestion processes with advanced Fourier transform ion cyclotron resonance (FTICR) mass spectrometric methods. In some of these studies, the cell's protein complement will be digested with a protease and the resulting peptides will be analyzed by capillary liquid chromatography-mass spectrometry (LC-MS). The use of tandem mass spectrometry (MS-MS) provides additional

sequence information that, when combined with the mass of the parent peptide, can be used to search existing databases. This results in peptide identification, which in turn is used to identify the parent protein. Additionally, we are extending this mass spectrometric technology to allow precise quantitation of changes in the protein complement upon perturbation of the microbial environment. This technology, based on the use of stable-isotope labeling, allows the creation of "comparative displays" for the expression of many proteins simultaneously. Two versions of each protein are generated and simultaneously analyzed to study changes in expression (i.e., repression or induction) for hundreds to thousands of proteins. These combined technologies are planned to be developed and demonstrated in a *D. radiodurans* pilot project that also would follow changes in the proteome after exposure to ionizing radiation.

Protein Domain Dissection and Functional Identification

Temple F. Smith, Sophia Zarakhovich, and Hongxian He
BioMolecular Engineering Research Center; College of Engineering; Boston University; 36 Cummington Street; Boston, MA 02215
617/353-7123, *tsmith@darwin.bu.edu*

Using various multialignment and conserved pattern tools (e.g., psiBLAST, BLOCKS, pfam, and pimaII), protein domains as "evolutionary modules" generally can be identified. Using a set of 20 completely sequenced microbial genomes (including yeast), we have generated over 1300 profiles representing diagnostic sequence domains. The majority either cover the entire length of the proteins matching the profile or locate a sequence region clearly identifiable in multiple distinct domain contexts. We are addressing the relationship between such sequence domains and structural domains as well as problems involved in associating these domains to a given biochemical function and the cellular role played by that function.

In collaboration with Julio Collado-Vides (CIFN, Mexico), we are investigating the potential for coordinate regulation among neighboring genes in various biochemical pathways. We began with sets of genes in *Escherichia coli* or some other bacteria or archaea organized in operons. Next, each operon set is being examined in yeast and *Caenorhabditis elegans* for shared regulatory sequences. Initial work led to the identification of two different types of eukaryotic operon-equivalent organizations in yeast and to our 1998 publication in *Microbial and Comparative Genomics.*

Genome Sequencing

Carl R. Woese and **Gary J. Olsen**
Department of Microbiology; University of Illinois; B103 Chemical and Life Sciences Laboratory; 601 S. Goodwin Ave.; Urbana, IL 61801
carl@ninja.life.uiuc.edu, gary@phylo.life.uiuc.edu

We prepared a sequencing-quality genomic DNA library for *Methanococcus maripaludis*, an organism that was being considered for sequencing as part of DOE's Microbial Genome Program (MGP). We have done some partial sequencing of clones from this library as part of a project to use comparative analysis to elucidate the differences between related high- and low-temperature proteins (this sequencing was partially supported by funding from the National Aeronautics and Space Administration).

We also prepared a sequencing-quality genomic DNA library for *Giardia lamblia*, a eukaryotic microorganism. This permitted Mitchell Sogin (Marine Biology Laboratory) to generate preliminary genome sequencing data for a successful grant application to the National Institutes of Health.

The sequence data resulting from our participation in MGP have stimulated additional research by our group and others. More specifically:

1. We continue to make new gene identifications through comparative analyses of sequenced genomes.
2. We have experimentally verified the function of some novel RNA methylase genes.
3. We have collaborated in the experimental identification of a novel, archaeal S-adenosyl methionine synthetase.
4. We have cloned and expressed RNA polymerase genes and transcription-initiation factors from archaea and have experimentally identified new protein-protein interactions in the transcription apparatus.
5. We have supplied 27 research groups with genomic DNA and cell mass from organisms sequenced as part of the MGP.
6. We are contributing ideas formulated as part of an MGP proposal into a successful collaboration with Carol Giometti (Argonne National Laboratory) to study the proteomes of *Methanococcus jannaschii* and *Pyrococcus furiosus*.
7. We have worked with the research group of Ross Overbeek (Argonne National Laboratory) on the development of his WIT and WIT2 environments for genome analysis and comparison and have used the WIT2 system to help with our analyses.

A Pilot Study to Develop and Demonstrate a High-Throughput New Approach to Characterizing Total Cellular Proteins Expressed by *Deinococcus radiodurans R1*

Kwong-Kwok Wong, Richard D. Smith, Ljiljana Pasa-Tolic, and Owen White[1]
Pacific Northwest National Laboratory; P.O. Box 999; Richland, WA 99352
509/376-5097, Fax: -6767, *kk.wong@pnl.gov*
[1]The Institute for Genomic Research; Rockville, MD 20850
www.tigr.org

Deinococcus radiodurans, with its exceptional radiation resistance, was once thought to grow within nuclear reactors, but further studies now suggest that the deinococci are soil microorganisms. Besides its resistance to radiation, *D. radiodurans* also has extreme resistance to cellular and genetic damage that occurs in other organisms after exposure to many genotoxic chemicals, oxidative damage, high levels of uv radiation, and desiccation. Thus, *D. radiodurans* is a potential candidate to be engineered for degradation of hazardous chemicals at mixed-waste sites, and it is important to understand at the molecular level how the bacteria can adapt to such stressful environments. The Institute for Genomic Research has completely sequenced the *D. radiodurans* genome, enabling further functional analysis of putative genes encoded by the bacteria.

In a pilot study, we have established a "2-D virtual gel" method and demonstrated that this new methodology is applicable to characterizing proteins expressed by *D. radiodurans*. Although numerous facets of the technology need significant refinement, we have generated preliminary results that are a major step beyond any "proteome" measurements made to date in terms of speed and sensitivity. In a single capillary isoelectric focusing (CIEF) separation with online FTICR mass spectrometry, we have detected at least 800 different proteins (based on the number of discrete molecular weight species above 5 kDa). This single experiment (requiring less than 30 min) uses about 250 ng of total protein, about 20 to 30 times less than that of a typical 2-D polyacrylamide gel electrophoresis experiment. This corresponds to low femtomole quantities for the average detected protein (with some proteins being detected at levels well into the attomole range). The potential exists to greatly improve the methodology's sensitivity, thereby opening up the detection of very low copy number regulatory proteins.

Related to these efforts, we also have developed a general targeted mutagenesis method based on *D. radiodurans* genomic information to define gene function. Using the targeted mutagenesis method, we have shown that both catalase (*katA*) and superoxide dismutase (*sodA*) genes are required for extreme radiation resistance. We are applying the 2-D virtual gel method to analyze proteins expressed by different mutants.

Characterization of expressed proteins by 2-D virtual gel and further targeted mutagenesis analysis will provide a link to the function of the genomic data's predicted open reading frames (ORFs) and is expected to identify new

small genes in the size range at which identifying ORFs is problematic. The resulting information can identify genes of interest and facilitate detailed biochemical and genetic experiments to gain a global understanding of the organism for energy and environmental and industrial applications. The developed 2-D virtual gel method will be applicable to any sequenced organism. This project was funded initially as a pilot study for 2 years but we expect research to continue well beyond that period.

Detection of Noncultured Bacterial Divisions in Environmental Samples using 16S rRNA-Based Fluorescent in Situ Hybridization

Cheryl R. Kuske, Susan M. Barns, and Stephan Burde
Environmental Molecular Biology Group; M888 Life Sciences Division; Los Alamos National Laboratory; Los Alamos, NM 87545
505/665-4800, Fax: -6894, *kuske@lanl.gov*

Microbial genome sequencing projects have focused primarily on species that can be easily cultured. Readily cultured bacteria, however, are only a small fraction of the total bacterial diversity present in the environment. Diverse bacteria representing novel divisions have been identified in many natural environments using 16S rDNA sequence analysis. Microbial processes in these environments are of critical importance to the biosphere, and the noncultured bacteria residing there are a valuable resource for novel genomic information. We have identified novel bacterial divisions from 16S ribosomal RNA gene libraries generated from DNA of a volcanic cinder field and an arid sandstone soil. Using RFLP and sequence analysis, we have analyzed 800 bacterial rDNA sequences obtained from the two arid environments. The majority of sequences were members of recently identified bacterial divisions that have no or very few cultivated members. Using PCR primers specific for two of these divisions (Acidobacterium and OP11) and their subgroups, we have detected both divisions in local hot or warm spring microbial mats and sediments. Analysis of cell abundance of members of these groups is under investigation using fluorescently labeled rRNA probes and fluorescence microscopy. We plan to collect bacterial cells directly from the environmental samples using flow cytometry and cell sorting. The pooled DNA of noncultured bacteria will be a valuable resource of genetic material for comparative analyses of conserved and novel gene families and for targeted genome sequencing.

Phylogenetic Analysis of Hyperthermophilic Natural Populations Using Ribosomal RNA Sequences

Norman R. Pace
Plant and Microbial Biology; University of California; Berkeley, CA 94720-3102
Current Address: Department of Molecular, Cellular, and Developmental Biology; University of Colorado; Boulder, CO 80303-0347
303/735-1864, Fax: /492-7744, *nrpace@colorado.edu*

It has become clear over the past few decades that substantial microbial diversity occurs at very high temperatures. Hyperthermophilic organisms (temperature optima >800°C) promise a wealth of unknown biochemistry and biotechnological potential and challenge our comprehension of biomolecular structure. Nonetheless, relatively little is known about the diversity of life at high temperatures because of a traditional problem in microbial ecology: the inability to cultivate naturally occurring organisms. Molecular techniques recently have been developed, however, that allow the detection and some characterization of organisms without cultivation. Limited surveys of hyperthermophilic communities using such techniques have revealed the existence of an unexpected plethora of organisms, some profoundly different from known ones. This program's main objective was to continue the phylogenetic and quantitative characterization, without cultivation, of ecosystem constituents that are known to be associated with particular high-temperature sites. Main focus was on the Yellowstone geothermal system.

These methods for characterizing organisms in the environment revolve about the use of rRNA sequences for phylogenetic analysis of population constituents. We obtained rRNA genes by directly cloning environmental DNA or by cloning products of polymerase chain reaction (PCR) amplification using primers complementary to universally conserved or phylogenetic group-specific sequences in rDNAs. Comparison of sequences to known rRNA sequences revealed phylogenetic relationships of organisms in the community to known organisms. In a second approach, fluorescently labeled oligonucleotide hybridization probes that bind selectively to rRNA were used for microscopic phylogenetic analysis of single cells. Results were highlighted by ongoing results from Yellowstone hotsprings.

Program results have contributed significantly to the emerging view of microbial diversity. Previous and ongoing studies have revealed a great wealth of archaeal diversity in sediments and scinters of hotsprings 70 to 95°C. These results have revised our understanding of archaea's phylogenetic depth and have allowed the recognition of archaea as a new kingdom. Surveys of bacterial "phylotypes" have expanded substantially our understanding of bacterial diversity; 12 of the current 36 to 38 bacterial divisions were first articulated in this program. Some of the newly discovered evolutionary lineages are sufficiently abundant that they must be significant in this ecosystem. Selected sequence-types were explored further using fluorescently labeled oligonucleotide hybridization

probes to visualize the organisms in their natural setting. Scanning electron microscope investigations showed that a succession of morphotypes forms biofilms on surfaces in hotsprings. Hybridization probes, in concert with available confocal microscopy, eventually will allow the reconstruction of three-dimensional aspects of this geothermal ecosystem.

We and others have found that types of organisms formerly thought to be restricted to high temperatures are in fact abundant at low temperatures and common in our environment. We used PCR primers characteristic of Crenarchaeota (thermophilic in all cultivated instances) to show that such organisms are common in sediments and soils at low temperatures, so they are likely to occur globally; such organisms also have been detected as abundant in the marine environment. Although not yet cultivated, these organisms are sufficiently abundant in the environment that they are likely to have impact on the biosphere's chemistry. Similarly with representatives of bacteria, we encountered many kinds of organisms previously thought restricted to low-temperature ecosystems. Indeed bacteria, not the commonly thought archaea, were found to dominate high-temperature ecosystems.

Overall the program has been contributory and conspicuous in the field of life in extreme environments. The period of performance for this program was July 15, 1995, to September 14, 1997.

This is a completed project.

The Ribosomal Database Project II: Providing an Evolutionary Framework

James R. Cole, Bonnie L. Maidak, Timothy G. Lilburn, Charles T. Parker, Paul Saxman, Bing Li, George M. Garrity, Sakti Pramanik, Thomas M. Schmidt, and **James M. Tiedje**
Center for Microbial Ecology; Michigan State University; East Lansing, MI 48824
Tiedje: 517/353-9021, Fax: -2917, *tiedej@pilot.msu.edu*
www.cme.msu.edu/RDP

The Ribosomal Database Project II (RDP-II) provides rRNA-related data and tools important for researchers from a number of fields. These RDP-II products are used widely in molecular phylogeny and evolutionary biology, microbial ecology, bacterial identification, microbial population characterization, and in understanding the diversity of life. As a value-added database, RDP-II offers the research community aligned and annotated rRNA sequence data, analysis services, and phylogenetic inferences derived from these data. These services are available through the RDP-II Web site (*www.cme.msu.edu/RDP*).

Release 7.1 (September 1999) contained more than 10,000 aligned and annotated small subunit (SSU) rRNA sequences. A special focus of this release was the identification and annotation of sequences from type material. Over 3000 type sequences representing 636 distinct prokaryotic genera were included in release 7.1. These type sequences provide a mechanism for users to place new sequences in taxonomic and phylogenetic frameworks. This release also included the introduction of an interactive assis-

tant to help with the planning and analysis of T-RFLP experiments (TAP T-RFLP).

We are now preparing release 8, scheduled for March 2000. We are enhancing the alignment to match a new set of guidelines for more consistent treatment of secondary structure regions. This release will contain over 20,000 aligned prokaryotic SSU rRNA sequences, including the vast majority of those available through GenBank release 114 (October 15, 1999). Initially, release 8 will be made available without manual curation of annotation information. We are establishing an RDP advisory panel to help us set new annotation standards to better serve our users with available curation resources. Release 8 also will mark a turning point for RDP. It will be the first release since 1994 in which the time has decreased between sequences becoming available through GenBank and being released in aligned format by RDP. We expect both the time and frequency of releases to continue to improve through 2000.

Capturing the Imagination to Capture the Mind: Using the Power of Informal Learning to Advance Science Literacy—A Report from the Microbial Literacy Collaborative

Cynthia A. Needham
Microbial Literacy Collaborative; Mount Mansfield Rd.; P.O. Box 3599; Stowe, VT 05672
802/253-2369, Fax: -6317, *caneedham@aol.com*
www.microbeworld.org

The Microbial Literacy Collaborative (MLC), a partnership of organizations committed to advancing scientific literacy by concentrating on the microbial world, created six major elements: (1) *Intimate Strangers: Unseen Life on Earth*, a science documentary for public television; (2) "Meet the Microbes," a set of hands-on activities; (3) national youth leadership summits for precollege students from traditionally underrepresented communities; (4) *Unseen Life on Earth: An Introduction to Microbiology*, a 12-part video for distance learning; (5) *Intimate Strangers: Unseen Life on Earth*, a companion book to the television series; and (6) an educational Web site for all ages.

Organizations that constitute MLC include the American Society for Microbiology; National Association of Biology Teachers; Oregon Public Broadcasting; and Baker & Simon, Associates. Other organizations include the Association of Science-Technology Centers and the American Association for the Advancement of Science. MLC is funded by the U.S. Department of Energy's Human Genome Program, National Science Foundation, American Society for Microbiology, Annenberg-CPB Project, Corporation for Public Broadcasting, and Arthur Vining Davis Foundations and the Foundation for Microbiology.

Intimate Strangers: Unseen Life on Earth, which received a common carriage designation, was broadcast in November 1999 by 94% of local PBS stations and was viewed by an average of 1.6 million households each week. The four hours of the series include the following:

- "The Tree of Life" delves into our evolutionary past. The key message in this hour is that all living things today evolved from microbes and share fundamental biological properties with them.
- "Dangerous Friends and Friendly Enemies" examines our ancient rivalry with the microbial world.
- "Keepers of the Biosphere" explores the central role that microbes play in sustaining the earth's ecosystems.
- Creators of the Future" examines our present and future use of microbial technologies to solve long-standing problems that affect the way we live.
- Meet the Microbes" is a collection of 17 hands-on activities for use in both informal and formal learning environments. Complementing the major themes of the television documentary, the activities require little or no specialized equipment or knowledge of microbiology. They support open-ended experimental design, help to

address elements of National Science Education Standards, and can be downloaded from the MLC Web site (*www.microbeworld.org*).

National youth leadership summits were week-long experiences designed to introduce youth leaders and their adult sponsors to the microbial world and prepare them to implement the hands-on activities in their local community programs. The first of two summits was held in August 1998 on the St. Paul campus of the University of Minnesota. The training experience was organized with the Association of Science-Technology Centers and their Youth Alive! Program. Participants represented 12 science museums from around the country, with youth leaders drawn primarily from challenged home environments. The second summit was held in Washington, D.C., in July 1999. Participants were drawn from science museums, youth clubs, and school science clubs from 12 different regions in the United States. A third summit is scheduled for summer 2000 in Portland, Oregon.

Unseen Life on Earth: An Introduction to Microbiology is a 12-part telecourse for use in both undergraduate and precollege classrooms. Each 30-minute film focuses on a different aspect of the microbial world. The telecourse, which was designed to address the curriculum standards endorsed by the American Society for Microbiology, is accompanied by teacher and student guides. The telecourse will support a full distance-learning course in microbiology or serve as supporting material for traditional classroom environments. To learn more about the telecourse, visit the MLC website (*www.microbeworld.org*).

Intimate Strangers: Unseen Life on Earth is a richly illustrated book to accompany the PBS science documentary *Intimate Strangers: Unseen Life on Earth*. It combines vivid, descriptive images from the series and original artwork with the compelling story of the world of microbes and their role in the earth's ecosystem. The authors have built upon the series content to offer a more comprehensive view of our relationship with the planet's tiniest inhabitants. Targeted to a general audience, the book's lively style will engage parents and their children and teachers and their students, along with other members of the scientifically interested public. The text puts the vitally important role of the microbial world into stories and terms familiar to the reader. The book is available through ASM Press; for more information, see *www.asmusa.org*.

The MLC's award-winning Web site (*www.microbeworld.org*) provides an Internet-based, interactive venue to introduce all learners to the microbial world. It contains information about the microbial world, educational resources for both teachers and students, and links to other sites with microbial information. The site received 689,620 hits during the month of November 1999 and recorded 29,779 user sessions. *Microbeworld.org* was chosen as a USA Today Hot Site and was featured in NetScape's What's Cool. Education World it an A+ review, and Education Planet gave it the top award.

This is a completed project.

Index of Project Investigators