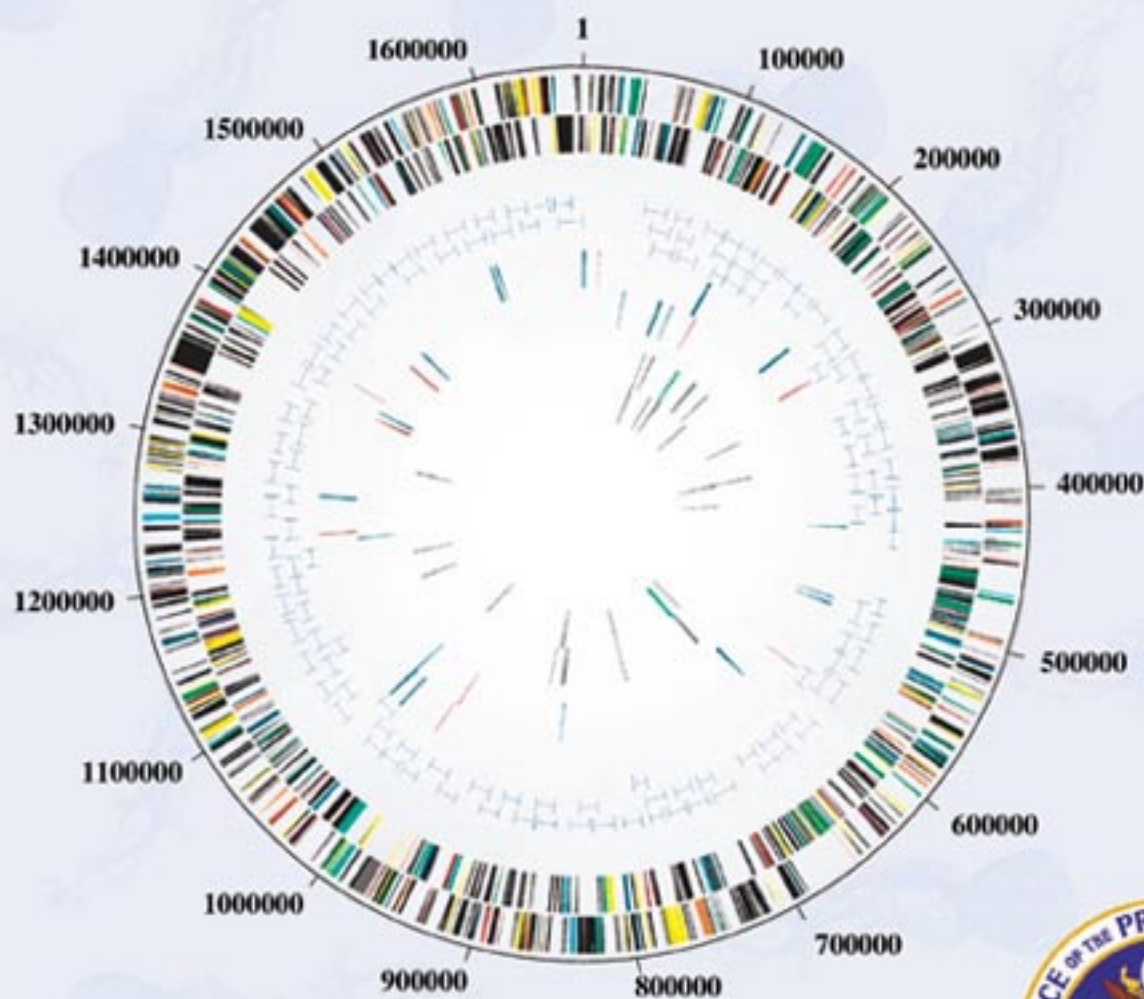


Interagency Report on the Federal Investment In *Microbial Genomics*



Methanococcus jannaschii



This is a report from the
Biotechnology Research Working Group
Subcommittee on Biotechnology
Committee on Science
National Science and Technology Council

Cover figure reprinted with permission from "Complete Genome Sequence of the Methanogenic Archaeon, *Methanococcus jannaschii*" C. J. Bult et. al., *Science* 273: 1058-1073.
Copyright 1996 American Association for the Advancement of Science.

Foreword

In April of 1999 the Subcommittee on Biotechnology charged a task group of the Biotechnology Research Working Group to prepare a report that would:

- **Summarize the activities of relevant federal agencies in microbial genomics.**
- **Identify each federal agency's areas of interest in microbial genomics** to identify gaps that could benefit by interagency collaboration.
- **Identify opportunities for and limitations to research in microbial genomics.**

Further elements of the charge were:

The report should contain summaries of the activities and estimates of the investment of federal agencies in the area of microbial genomics in the context of their investment in the larger area of microbial biology. Genomics, in this report, should include both microbial sequencing projects and post-sequencing/functional genomic projects.

The report should outline each agency's areas of interest in microbial genomics, e.g. plant or animal pathogens, environmentally interesting microbes, technology development, etc. The goal in this portion of the report should be to identify gaps and areas of potential interagency collaboration.

The report should present areas of opportunity for microbial genomics and current and potential limitations to the progress or scope of this work. Issues of education, including short- and longer-term challenges to the genomic workforce; access to information, such as genomic or EST sequences or microarray data; and access to technology such as microarrays and computational capacity/bioinformatics should be addressed.

The report of the task group follows.

Mary E. Clutter, Chair
Subcommittee on Biotechnology

Report Contributors Include:

Margaret Werner-Washburne, NSF, coordinator
Bart Kuhn, Eric Eisenstadt, DOD
Daniel Drell, DOE
Kathie Olsen, NASA
Judy Vaitukaitis, NIH-NCRR
Michael Gottlieb, NIH-NIAID
Elise Feingold, Bettie Graham, NIH-NHGRI
Dennis Mangan, NIH-NIDCR
Marcus Rhoades, NIH-NIGMS
Eugene Koonin, NIH-NLM-NCBI
Linda Beth Schilling and Gregory Vasquez, NIST
Mike Perdue, Caird Rexroad, USDA-ARS
Peter Johnson, Michael Roberts, Ann Lichens-Park, Sally Rockey, USDA-CSREES

With special help from Lynn Fletcher, NSF

EXECUTIVE SUMMARY

The age of microbial genomics began four years ago with the publication of the complete sequence of *Hemophilus influenzae*. Since then the number of completed genomes has increased exponentially and, with the growth of high-throughput DNA sequencing facilities, there is no end in sight. The federal investment in microbial genomics has been instrumental in the growth of this field. It has led to the sequencing of model organisms, such as *Escherichia coli* and *Saccharomyces cerevisiae*, moving researchers who work on these organisms into the stage of functional genomics, where issues of access to technology, reagents, and computational analyses become significant. Support for sequencing of human pathogens has led to the identification of potentially new strategies that *Plasmodium falciparum* uses for synthesizing antigenic variants. The investments in sequencing microbes like *Thermotoga maritima* and *Deinococcus radiodurans* have led to changes in our understanding of the extent of horizontal gene transfer among prokaryotes and the role of horizontal gene transfer in the evolution of genome structure. This has led to deeper questioning of the nature of bacterial species and has driven fundamental changes in the analysis of prokaryotic evolution.

The current effort in microbial genomics in each of the federal agencies is based on the mission of the specific agency. As a result, there are clear gaps or opportunities for cooperation in the area of microbial genomics. Some of the gaps or opportunities were identified by one or two agencies, but most were seen as important by all of the agencies that participated in this report.

Gaps and opportunities that apply specifically to microbial genomics include:

- genomic analysis of microbes whose genomes are of scientific interest or practical importance but are not well represented in publicly funded sequencing projects
- development of techniques for culturing novel microbes
- development of techniques for *in situ* genomic analysis of microbial ecologies appropriate to field work in extreme environments, including the space environment

Gaps and opportunities that apply to genomics more broadly include:

- infrastructure to increase the U.S. capacity in computational biology and bioinformatics and to provide access to genomic technology and reagents
- technology development for high-throughput genomic research, especially in the areas of computational biology, bioinformatics, and functional genomics
- technology transfer in the area of genomics to the broader industrial community and more rapid access to these technologies by academic researchers
- training in genomics and computational biology at all levels
- development of uniform policies on issues such as data release and reagent sharing

TABLE OF CONTENTS

| | |
|--|------------|
| INTRODUCTION | 1 |
| DOD- DEPARTMENT OF DEFENSE..... | 5 |
| DOE – DEPARTMENT OF ENERGY..... | 7 |
| NASA- NATIONAL AERONAUTICS AND SPACE ADMINISTRATION..... | 9 |
| NIH – NATIONAL INSTITUTES OF HEALTH..... | 11 |
| NCRR – NATIONAL CENTER FOR RESEARCH RESOURCES | 11 |
| NHGRI - NATIONAL HUMAN GENOME RESEARCH INSTITUTE | 11 |
| NIAID - NATIONAL INSTITUTE OF ALLERGY AND INFECTIOUS DISEASES | 13 |
| NIDCR – NATIONAL INSTITUTE OF DENTAL AND CRANIAL RESEARCH..... | 14 |
| NIGMS – NATIONAL INSTITUTE OF GENERAL MEDICAL SCIENCES | 15 |
| NATIONAL LIBRARY OF MEDICINE – NATIONAL CENTER FOR BIOTECHNOLOGY INFORMATION | 17 |
| NIST – NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY..... | 18 |
| NSF – NATIONAL SCIENCE FOUNDATION | 21 |
| USDA – UNITED STATES DEPARTMENT OF AGRICULTURE..... | 24 |
| GAPS/OPPORTUNITIES..... | 26 |
| REFERENCES..... | 29 |
| APPENDIX 1 – FUNDING AGENCIES AND ORGANISMS OF RESEARCH | A-1 |
| APPENDIX 2..... | A-7 |
| APPENDIX 3..... | A-8 |

INTRODUCTION

The DNA sequence that comprises the genome of a living creature encodes the directions for growth and development of the organism. It provides insight into how that organism is built, responds to stimuli, grows, and reproduces, and, sometimes, even how it dies. The field of genomics developed from the desire to know how differences in sequence of four deoxynucleotides lead to synthesis of RNA and protein and eventually the development and regulation of an entire organism. The analysis of genomic DNA is providing new insights into evolution, pathogenesis, cell organization, structure, and function, and interactions among organisms in the environment- ultimately leading to the development and testing of models to explain complex systems from host-pathogen interactions to nutrient cycling on Earth.

Microorganisms have played a critical role in the development of genomics. Publication of the genome of *Hemophilus influenzae* in 1995, the first complete genomic sequence from any single or multi-cellular organism, heralded the advent of microbial genomics (1). Since then, 20 additional microbial genomes have been completely sequenced (including the eukaryotic microbe, *Saccharomyces cerevisiae*) and many more microbial genomes are in progress and nearing completion. The ever-increasing ability of researchers to study genomes and the way they are organized has changed the way biologists look at living cells and fueled a revolution that is reaching far beyond the life sciences. It has also led to the realization that we know much less about the microbial life on earth than previously thought.

Microbes are an amazingly heterogeneous but relatively unknown group of organisms. Over approximately 3.8 billion years of evolution, microorganisms have managed to develop the ability to survive and grow in an astonishing range of “extreme” environments, including the extremes of pressure, nutrient availability, temperatures, salt, oxygen potential, pH, ionizing radiation, and water potential. They have developed the ability to grow as pathogens and symbionts that exhibit many different and intricate connections with their hosts. They may live as single cells, in colonies, or in intricately structured communities. Microbes have found ways to make almost every type of environment “hospitable”. Although we do know about microbes from a variety of environments, genomic and molecular research has led to the estimate that 99.99% of microbes resident in the environment have yet to be discovered. Some explanations for our lack of familiarity with most microbes are the challenges and difficulty of culturing them and the fact that the vast majority have never been associated with any human, plant, or animal diseases (2). Because of the unique properties of the microbes that have been studied and the almost incomprehensible number of microbes on earth that are yet to be studied, these organisms represent an untapped and extremely valuable resource for the basic sciences, medicine, biotechnology, and bioengineering.

While microbes have been known mostly for their association with human, animal, and plant disease or their role in production of wine, beer, bread, and cheese, the truth is that microbes, both known and unknown, are essential to life on Earth. Microbes, resident in

the soil, in the oceans, in and on most life forms, are increasingly thought to participate in global carbon and nitrogen management, biotransformation and biodegradation reactions involving metals and other elemental substances suggesting they play a major role in the global "metabolism" of Earth. Paradoxically, the microbes, particularly the bacteria and fungi, that comprise a relatively invisible biological background are, in terms of their total mass, species diversity, and metabolic range, among the dominant life forms on the earth.

Genomics and the improvements in microscopic technology are leading to remarkable new discoveries about microbial community structure, intercellular communication in nature, and strategies for survival. An example of this is the relatively recent discovery that microorganisms in extreme environments (including prosthetics within living organisms and in space stations such as Mir) are found in biofilms (3). For microorganisms, biofilms offer protection from environmental stresses and allow them to withstand even killing doses of antibiotics. It has been estimated that biofilms play a significant role in transmission and persistence of human disease. Another example is the realization, after the genomic sequence was complete, that a quarter of the genome of the bacterium *Thermotoga maritima* has come, apparently by lateral transfer, from a different domain of life, the Archaea (4). A third discovery was the remarkable ability of the bacterium *Deinococcus radiodurans* to withstand normally lethal doses of ionizing radiation by fragmenting its genome and repairing it over a period of time after the exposure.

Genomic analysis has made scientists refine their ideas of microbial evolution. Recent discoveries, including the finding that a quarter of the *Thermotoga maritima* genome appears to have been derived from horizontal gene transfer from Archaea (4) have led to the realization that our current algorithms for understanding microbial evolution are not adequate for deciphering evolutionary relationships in prokaryotes (5). By developing new algorithms and attempting to culture new microbial species, computational biologists and microbiologists, respectively, seek to address these new challenges to our understanding of evolution of prokaryotic life on earth. Further analysis of sequenced genomes may make it possible to distinguish linear inheritance from horizontal gene transfer through the genetic and computational analysis and modeling of metabolic pathways. Thus, the ability to analyze entire genomes has allowed scientists to recognize and begin to address in fundamentally new ways some of the surprising intricacies of microbial evolution.

Functional genomic analysis of microbes is a tool for understanding life in the universe. Understanding the origin and evolution of life on earth is paramount for the search for life elsewhere. Given that the most primitive surviving lifeforms are microbial, and that microbes have played a dominant role in the evolution of the terrestrial biosphere for the past 4 billion years, it is reasonable to assume that microorganisms exist in other locations within the space environment. Genomic analysis provides the framework to develop models for understanding the limits of life and the nature of habitable environments and thus, is essential for understanding the potential for life elsewhere in the universe and developing the strategies to detect it.

Genomic analysis of pathogenic microbes may lead to new ways to treat disease. Not only will knowing all of the genes in a pathogen allow the identification of novel therapeutic and vaccine targets, but this information can be used to identify novel types of antibiotics or growth regulators that normally function to control the growth of organisms in mixed microbial communities. Having the genomic sequence of pathogens makes it possible to examine the diversity and evolution of pathogens, identify mechanisms of drug resistance, understand how pathogens interact with their hosts, and characterize pathogen virulence factors. Genomic analysis has also contributed to the realization that more needs to be known about the diversity of microorganisms living within and on humans, plants, and animals and the significance of that diversity in maintaining health (2). Genomic information is beginning to provide insights into the natural history and epidemiology of infectious diseases and, thereby, into disease prevention. Because of the clear significance of genomic analysis to human health, over half of the microbial sequencing projects to date have involved human pathogenic microbes. By comparison, little publicly funded genomic research has been done on agriculturally relevant pathogens, although the genomic analysis has the same potential value for plant and animal health.

Functional genomic analysis of microbes has contributed to our understanding not only of microbial but also human cellular function. Because pathways have been conserved through evolution, the ability to study a simple eukaryote such as *S. cerevisiae* at the genomic level, has allowed the identification and study of a variety of genes known to be associated with human diseases (6). Additionally, functional genomics tools, such as the yeast two-hybrid systems, open up the possibility of identifying important interacting proteins from many different organisms, including plants and human cells (7).

The potential use of microbes in biological warfare underscores the need for genomic analysis. The threat that microbes, perhaps supplemented by modern genetic engineering technologies, might be used in terrorist acts is one that must be taken very seriously. Analysis of the genomic sequence of these microbes can greatly aid detection, characterization, forensic "attribution" (e.g. where did a biological agent come from and what are its modifications), and specific response to exposure. Public access to the genomic sequence data for these microorganisms is critical for integration of this information with available medical and epidemiological information and the most rapid development of technology to address the needs in this area.

Although genomic analysis of microbial genomes has yielded surprising discoveries, there is a great deal more to be learned. In each microbial genome that has been sequenced, 40 – 50 % of the putative open reading frames encode proteins of unknown function and 20 to 30% encode proteins seen only in that species. These results suggest that the number of genes (and, therefore, species) yet to be identified is likely to be enormous. The availability of genomic sequence has led to a dramatic change in how biology is done. Now, instead of making single mutants to study complex pathways, biologists can start with the complete set of genes and end with an understanding of the structure and function of whole cells. Finally, the rapid development of functional genomics technologies, such as DNA microarrays and mass

spectrophotometric analyses of protein complexes are making it possible to build a multidimensional and integrated picture of living cells in real time, a task that only a short time ago would have been impossible to imagine.

It is an understatement to say that microbial genomics has contributed to major scientific discoveries. Through genomics, we have new ways of looking at the world and asking questions of enormous value to research in areas such as evolution, metabolism, microbiology, agriculture, bioengineering, and medicine. Genomics has turned scientists into explorers as never before and enabled them to discover new worlds in the microscopic organisms that have been living beneath our feet, over our heads, and within and on our bodies all this time.

DOD- DEPARTMENT OF DEFENSE

A. Agency interests in microbial genomics. The DOD investment in microbial genomics is driven by both biomedical and non-biomedical interests and by military operational requirements. In the biomedical area DOD is interested in developing technologies that would provide health support and services to military personnel and that would counter the threat of endemic infectious diseases and biological warfare (BW) agents. A major focus of the DOD investment in microbial genomics is, therefore, directed at developing genomic-based information about infectious agents, including potential BW agents, that can be exploited for the rational design of therapies, vaccines, detection, and medical diagnostic strategies.

In the non-biomedical area, DOD is interested in biotechnological approaches for developing new materials and managing the impact of DOD operations on the environment. Information emerging from functional genomics research should enable new technologies for development of novel biosynthetic schemes for producing materials of interest to DOD and should provide a better understanding of processes governing the fate and effects of contaminants in marine and terrestrial sediments. Genomic information about novel, naturally occurring plasmids could enable the development of new biotechnology-based tools for manipulating microbes.

To benefit fully from the information provided by genomic sequence analysis requires tools that enable prediction of the structure, function, regulation, and physiological impact of gene products. To this end, DOD has programs that fully integrate genomic sequence and functional genomics research.

Past and current programs/research supported by DOD.

Parasites: The DOD plays a significant role in a consortium of agencies dedicated to sequencing the entire genome of the malaria parasite *P. falciparum*. Such sequence information is being exploited to determine whether there are genes coding for parasite-specific proteins that might be (a) inhibited by drugs specifically designed to prevent or cure malaria infections, and (b) the basis of DNA-based malaria vaccine being developed by DOD.

Microbes: DOD initiated the effort to sequence the genome of *Bacillus anthracis* and has recently begun a new effort to determining the sequences of nine, novel plasmids found in marine sediment microbes. Other microbes and pathogens are sequenced as needed to assist in the development of vaccines, drugs, and/or diagnostic tests.

Functional Genomics: These efforts include interdisciplinary projects between experimentalists and theorists that focus on developing analytical tools to model and simulate the dynamic behavior of genetic (transcriptional) regulatory networks. A related approach is to assemble simple artificial regulatory circuits to establish proof of principle capabilities for designing

transcriptional networks and evaluating our understanding of how naturally occurring networks operate. An additional effort in "molecular field biology" is focused on developing a taxonomy for protein structural domains and folds that will enhance the ability to convert genomic sequence information into a structural and functional database. In both the BW-defense program and infectious disease research program, functional genomics projects are essential for both developing vaccines, drugs and diagnostics, and for supporting the safety and efficacy information required by the FDA for product licensure.

DOE – DEPARTMENT OF ENERGY

A. Agency interests in microbial genomics. DOE has played a significant leading role in the development of the field of microbial genomics. DOE's interest in microbial sequencing focuses on elucidating microbial processes that affect the environment (carbon management), produce energy (methanogenesis, etc.) and participate in the remediation of sites contaminated with heavy metals, radionuclides, and other legacies of a 50 year history of nuclear weapons production. This interest developed as an offshoot of the Human Genome Program initiated by the Office of Biological and Environmental Research (OBER) in 1986. The DOE Microbial Genome Program was begun late in Fiscal Year 1994 with awards to The Institute for Genome Research (TIGR), the University of Utah, and Genome Therapeutics Corp. The OBER program has supported the completion of genomic sequencing for 11 microbes and 12 more are in various states of progress (see Appendix 1). The impact of these projects on microbiology has been dramatic, with no discovery more indicative than the repeated observation that 25-40% of the newly discovered genes have unknown functions and thus presumably participate in as-yet unknown processes within the cell. In addition, the availability of the complete genomic sequence has led to a drastic paradigm shift in the way microbial biology is done: primarily, moving from studying one gene or one protein to being able to quantitate the response of every gene to changing conditions. This shift has allowed rapid identification of genes that may be important in a wide range of responses.

B. Past and current programs/research supported by DOE. The microbial sequencing projects supported by OBER have had a broad impact. OBER has been careful to chose microorganisms for sequencing that fit with their mission while also having a modest genome size (less than 8 Mb), readily obtainable DNA, and either tractable genetics or the potential for molecular genetic approaches. In addition, the organisms chosen were non-pathogens and scientifically interesting, i.e. because of their phylogenetic placement. Although OBER selected organisms based on the above criteria, the genomic analysis of these organisms has had a much broader impact in the areas of evolution, molecular and structural biology, and genetics than could have been predicted. Most recently, the demonstration of a high degree of horizontal gene transfer in *Thermotoga maritima* (Nelson et al., 1999) has contributed to a broad discussion of the notion of species in the bacterial lineage.

The process by which OBER chooses and supports microbial genomes occurs in several steps. Because there is still a vast repertoire of candidate organisms that satisfy all these criteria, OBER has convened a series of "Which Bugs" workshops to solicit external expert guidance to identify the most promising microbes for sequencing. The core process in the DOE Microbial Genome Program is to invite pre-applications, which following a review for programmatic relevance are usually then returned with an invitation to submit a formal application. Formal applications are peer reviewed at panels and funding is supplied to the most meritorious, following an overall program relevance review and within the limitations imposed by the available budget. Although about \$18 million has been spent directly on microbial genome projects through FY 1999, the

OBER during this time has spent about \$31 million for microbial sequencing and research at academic institutions and non-profit organizations. Just over \$6 million (about 1/6 of the total) has been allocated to scientists at DOE National Laboratories.

DOE also supports bioinformatics relevant to microbiology. DOE, along with the NSF and the NIH, supports the Protein Data Bank (PDB) at Rutgers University; PDB stores 3-D structures for proteins from a variety of sources including microbes. DOE also supports a Comprehensive Microbial Resource at (TIGR) which is a database and set of tools that enables comparative genomic and functional genomic analyses. DOE also supports the Ribosomal Data Base at the Michigan State University Center for Microbial Ecology as well as computational efforts at the Lawrence Berkeley National Lab to recalculate the rDNA phylogenetic tree.

Other offices within DOE have also supported microbial research. When all the expenditures for microbial related research within DOE are summed, about \$70-75 million is spent each year on microbial research. The Office of Basic Energy Sciences (BES) has, for a number of years, supported research on fermentation microbiology, extremophiles and their metabolism, and biomaterials and biocatalysis, particularly energy-related enzymes. BES is also supporting the sequencing of one microbe involved in carbon and nitrogen fixation. The Office of Energy Efficiency has supported research on microbial renewable energy production, hydrogen and ethanol production, and cellulose and lignin degradation. The DOE Office of Environmental Management and Waste Remediation has supported work on microbial degradation of organic contaminant wastes. More recently, the DOE Office of Nonproliferation and National Security has begun to support research on microbes that could be biowarfare or bioterrorist threat agents.

NASA- NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

A. Agency interests in microbial genomics. NASA seeks to understand the nature of life in the universe and to assure crew health and productivity for increasing periods of time and at increasing distances beyond Earth. NASA's primary interests are in functional genomics of extreme environments, including the space environment, computational biology, bioinformatics, *in situ* genomic analyses, medical genomics, and genomics as a basis for bioengineering. A high priority activity is to develop the tools that enable correlation of environmental changes with changes in gene expression, and correlate these with resultant gene products, metabolic effects, and structural changes over multiple generations.

NASA's astrobiology program seeks to understand life in the universe—its origins, evolution, distribution and destiny. Microbial genomics plays an important part in all of these endeavors. Through genomic archaeology, clues to the nature of the earliest life on Earth can be gleaned, and insights into the process by which life and the environment co-evolved on the early Earth can be obtained. These studies provide important scientific information in their own right, but also develop a basis for comparing and interpreting information obtained from other worlds in our solar system and beyond it. Most of Earth's biological evolution was dominated by microbes. If life exists beyond Earth, there may be many more worlds dominated by microbial life than by higher life forms. Functional genomic studies of microbes in extreme environments on Earth, especially those conducted *in situ*, provide models for understanding the limits of life and the nature of habitable environments. This information is essential in understanding the potential for life elsewhere in the universe and for developing strategies to detect it.

Microbial genomics is essential to supporting human exploration beyond Earth. Microbes represent a health hazard for exploration crews, either in their natural state or through possible mutations brought about by novel selection pressures, including the closed environment of the spacecraft, microgravity and space radiation. For the future, habitable artificial ecologies designed to operate beyond Earth over decadal time periods will almost certainly employ microbes as part of their life support strategies, including those that are bioengineered for specific functions. Finally, microbial genomics plays an important role in understanding whether terrestrial life can establish a successful evolutionary future beyond Earth by natural or engineered means and how that may be achieved.

B. Past and current programs/research supported by NASA.

Astrobiology: NASA's investment in genomics within the discipline of astrobiology is centered on discovering phylogenetic relationships to determine our last common ancestor, to investigate life's earliest metabolic capabilities, and to infer Earth's earliest environments. These studies are required to understand the origin and evolution of life on this planet as a beacon for the search for life elsewhere, as described in the NASA Astrobiology Roadmap. NASA has played a major

role in supporting Carl Woese's research leading to the definition of a phylogenetic tree based on genomic analysis. This work, with DOE support, culminated in the recent complete sequencing of thermophilic *Methanococcus jannaschii*, the first sequenced archaea, proving that archaea are the third domain of life (bacteria and eukarya being the other two). Work in this area is continuing in areas such as sorting out the role of horizontal gene transfers in evolution and defining early branches of the eukaryotes.

Bioastronautics: NASA's Fundamental Biology Program and Biomedical Research and Countermeasure Program, within the Life Sciences Division, have a strong interest in supporting genomics research, focused on integrated and functional genomics to understand complex biological pathways and systems, and their interactions, in support of human spaceflight. Such an approach is critical, for example, to ultimately understand the myriad physiological changes that occur in bacteria, plants and animals during space travel. Recent examples of such work include: 1) the use of micro-array DNA chip technology to identify sets of genes which respond to altered gravity exposure in plants and animals, 2) targeted gene modification in animals to determine the role of these genes in plant and animal physiology during microgravity exposure, 3) the use of a functional genomics approach to sequence and characterize the expression and regulation of genes involved in key evolutionary events and to test the function of these genes in both same-species and cross-species transient expression assays, and 4) the use of functional genomics to investigate changes in bacterial gene and protein expression with microgravity exposure, and to assess the virulence potential of these pathogens in microgravity.

NIH – NATIONAL INSTITUTES OF HEALTH

NCCR – National Center for Research Resources

A. Agency interests in microbial genomics. NCCR provides a broad array of technologies, tools, and materials to carry out research in microbial genomics. Biomedical Technology Centers support research, development, and access to advanced technologies and techniques needed to analyze microbial genomes, such as synchrotron radiation, mass spectrometry, and nuclear magnetic imaging and spectroscopy. The Shared Instrumentation Grant Program (SIG) funds grants for the acquisition of state-of-the-art instrumentation to groups of NIH-supported investigators. NCCR also provides support for the exploration of new technologies and new approaches in microbial genomics as well as support for shared resources that supply organisms and information to investigators for studies in microbial genomics.

B. Past and current programs/research supported by NCCR. Two NCCR-supported mass spectrometry centers are developing new techniques for directly identifying proteins from complexes that bypass the potential limitations of gel electrophoresis. One process, multidimensional liquid chromatography and tandem mass spectrometry, has recently been applied to analyze all the ribosomal proteins in *Saccharomyces cerevisiae*. Another technique, whole-cell stable isotope labeling, which can be applied on proteome-wide basis, has been applied to high abundance proteins from yeast. The Yeast Resource Technology Center has been established to exploit the yeast genome. This comprehensive approach should serve as the model for future efforts on other eukaryotes as their genome sequences become available. Specifically, the Center integrates a set of state-of-the-art analytical technologies, including mass spectrometry and two-hybrid analysis and microscopy, to analyze protein complexes.

The SIG program provides key instruments needed to analyze microbial genomes including high-throughput protein and DNA sequencers, sequence detector systems, and DNA chip technologies (microarray systems). Three ABI 377 DNA sequencers, funded by SIG, were used to complete the genome of *E. coli* K12. Gene-chip scanning equipment has been used in conjunction with an Affymetrix collaboration to develop an *E. coli* Gene Chip. Currently, a SIG-supported microarray technology is being used to mass produce an *E. coli* microarray for the research community.

NHGRI - National Human Genome Research Institute

A. Summary of NHGRI's interests in microbial genomics: NHGRI supports the development of a wide range of technologies that are applicable to the study of microbial genomes, including improved technologies for DNA sequencing and for large-scale functional analysis of eukaryotic genomes. NHGRI's specific interest in microbial genomics is in the

analysis of the *S. cerevisiae* genome, including support of large-scale functional analyses and the *Saccharomyces* Genome Database (SGD).

B. Past and current microbial genome programs sponsored: NHGRI identified five non-human organisms in which to invest in the development of genomic resources, primarily the complete genomic sequence. Two of these, *E. coli* and *S. cerevisiae*, are microbial organisms. Through the efforts of an international consortium, the genomic sequence of *S. cerevisiae* was completed in 1996; NHGRI supported the generation of 20% of this sequence. The complete sequence of *E. coli*, generated entirely through NHGRI support, was published in 1997. There are no other microbial sequencing projects currently supported by NHGRI.

With the completion of whole genome sequences, over the past few years NHGRI has begun to develop programs in the analysis of genomic sequences. While NHGRI is no longer actively supporting any significant level of research on *E. coli*, it currently supports research projects on genome-scale functional analysis of *S. cerevisiae*. In FY 1998, NHGRI invested \$4.5 million to support a mix of technology development (\$1.0 million), pilot and production-scale (\$3.5 million) activities, approximately half of which were funded through specific solicitations. NHGRI also maintains a significant investment (\$1.3 million in FY1998) in the *Saccharomyces* Genome Database (SGD) at Stanford.

The research projects being supported include: 1) Gene disruptions: U.S. component of an international consortium to generate disruptions in each of the ~6000 genes; development of conditional mutations in essential genes; 2) Large-scale analysis of gene expression (RNA and protein) using a variety of methods including DNA microarrays, kinetic RT-PCR, and mass spectrometry; 3) Other large-scale studies exploiting microarrays, including the function of intergenic regions and analysis of complex traits; 4) Construction of portable libraries; and 5) Cross-referencing of yeast with human and mouse genes.

In addition to these projects that are focused primarily on *S. cerevisiae*, NHGRI supports a myriad of technology development activities that could be applicable to the study of microbial genomes. Most notable is the major emphasis on research to reduce the cost of DNA sequencing. Other relevant activities include efforts to develop or improve technologies for functional analyses, including analysis of RNA and protein expression, protein interactions, genetic mapping and sequence variation, and mutagenesis. Many of these activities include the development of informatics tools.

C. Plans for the investment in microbial genomics and collaborations: NHGRI recently published its new set of five year goals for 1998-2003 (<http://www.nhgri.nih.gov/98plan/>). One of the new goals is directly relevant to microbial genomics: Goal 4 - Technology for Functional Genomics. NHGRI's primary interest is in the area of technology development, where the emphasis is on technologies that can be used on a large scale, are efficient and are capable of generating complete data for the genome as a whole. These technologies should be either broadly applicable to any genome, or applicable to one or more of the eukaryotic organisms that are the

focus of NHGRI support. Application of these technologies to large-scale studies has been and will continue to be, for the most part, considered on a limited basis, as resources allow. In addition, given the broad interest and applicability of these studies across the NIH, coordination and collaboration with other NIH Institutes/Centers, will be maximized. This having been said, NHGRI has had particular interest in supporting large-scale functional studies in *S. cerevisiae* and this interest is likely to continue. These studies serve as a model of these types of projects in other eukaryotic organisms, both from a technical point of view and because of the challenge that analyzing this type and scale of data poses.

NIAID - National Institute of Allergy and Infectious Diseases

A. Agency interests in microbial genomics. The NIAID supports a large amount of research on microbial pathogens, including bacteria, fungi, protozoa and helminthes, that are responsible for diseases of public health importance both domestically and globally. Consistent with the Institute's mission, the goal of this research is to enhance our understanding of the etiology, epidemiology and pathogenesis of infectious diseases and to translate this research into effective therapeutic and prophylactic approaches to control these disease threats. The Institute has recognized the enormous potential of genomics in accelerating the pace of research on infectious agents, including for studies leading to the identification and role of infectious agents in chronic diseases (neurological, cardiovascular, gastrointestinal; neoplasms *etc.*). Having access to a microorganisms entire genome sequence will provide the sequence of every biochemical pathway, every virulence factor, every drug target and every protein antigen. The Institute also recognizes the challenge and opportunity posed by handling and decoding all of this information. The Institute is committed to sustaining its support of projects to sequence disease causing pathogens as well as to increasing its support of resources for functional genomics studies.

B. Past and current programs/research supported by NIAID. To date, the NIAID has funded the projects to sequence the complete genomes of approximately 20 bacterial pathogens, two of which have been completed, viz. *Chlamydia trachomatis* and *Treponema pallidum*. In addition, the Institute funds projects to partially sequence the genomes of parasitic protozoa. With NIAID support, one chromosome each from the malaria parasite *Plasmodium falciparum* and from *Leishmania major* has been completely sequenced. A listing of the genome projects supported by the NIAID is available on the Institute's Web site <http://www.niaid.nih.gov/dmid/genome.html>. The NIAID is also funding work in the areas of bioinformatics and functional genomics of pathogens. The Institute currently supports, through an Inter Agency Agreement with DOE, a database for sexually transmitted pathogens (<http://www.stdgen.lanl.gov/>). This database builds on the genome sequence information derived from the large-scale sequencing of *Treponema pallidum*, *Chlamydia* and other STD organisms. A number of NIAID-supported investigators are using the genome sequence information to prepare microarrays for gene expression analyses.

C. Future investments and collaborations in microbial genomics. The Institute has recently reviewed its support of large-scale genome sequencing projects. As a result of this review, the NIAID has drafted a policy statement that indicates a number of changes in the mechanism of support and the terms of award for these projects. An NIAID-convened Blue Ribbon Panel on Genomics, held on May 12-13, 1999, endorsed the policy. The policy was approved by the National Advisory on May 25, 1999. In brief, the policy indicates: 1) large-scale genome sequencing projects will be supported under the cooperative mechanism; 2) applicants are required to obtain prior approval from NIAID before submission; and 3) Institute organism priorities for large-scale sequencing projects. The new policy is intended to stabilize funding for large-scale genome sequencing projects at least at the current level of \$12M/year. The Institute will publish the Council-approved policy on mechanisms of support for large-scale genome sequencing projects in the NIH Guide in the very near future. In addition, the Institute will announce its list of prioritized organisms for future projects.

The NIAID is developing its plans for providing resources to enable investigators to take advantage of the large-scale genomic sequence information. These resources will provide at least some of the following:

- reagents derived from the sequencing projects (e.g. tiling set of plasmid clones; PCR primers or amplicons for each ORF);
- databases;
- microarrays and chips;
- emerging technologies (e.g. proteomics).

It is anticipated that a contract resource will be funded beginning in FY2000.

The NIAID collaborates with a number of other federal and international agencies as well as private foundations in support of microbial genomics. Examples of these collaborations include: co-funding of the malaria genome sequencing project with DOD, the Wellcome Trust and the Burroughs-Wellcome Fund, coordination of genome sequencing projects with the Wellcome Trust, discussions with the World Health Organization's Programme for Research and Training in Tropical Diseases for genome projects on parasites and invertebrate vectors of disease.

NIDCR – National Institute of Dental and Cranial Research

A. Agency interests in microbial genomics. Despite remarkable scientific advances in the control of infectious diseases in general medicine, infections of the oral cavity, such as periodontitis, caries, infections of dental implants, and viral stomatitis remain a formidable problem in contemporary dentistry. This challenge is further magnified by sub-populations who are immunocompromised because of drug use, systemic viral diseases such as AIDS or inadequate saliva production. To meet these challenges, the NIDCR supports basic, clinical, translational, epidemiological and developmental research on infectious diseases of the oral cavity.

Sequence analysis of the entire genome promises to yield a comprehensive picture of the structure and function of oral microorganisms. In this regard, genome analysis may be able to elucidate previously unrecognized pathogenic mechanisms that can be blocked by drug therapies, and immunogenic components ideal for vaccine development. In addition, data from these studies will enable extensive comparisons to be made between bacterial genera and species, thereby identifying the genetic basis for virulence and ability to survive in the oral cavity.

B. Past and current programs/research supported by NIDCR. The NIDCR currently supports the complete sequencing of four oral pathogenic bacteria and a yeast. Three of these bacteria, *Porphyromonas gingivalis*, *Treponema denticola*, and *Actinobacillus actinomycetemcomitans*, are associated with periodontitis (gum disease). The other bacterium, *Streptococcus mutans*, is a principle cause of dental caries. The yeast, *Candida albicans*, is a common opportunistic pathogen that causes a painful and debilitating oral mucositis in patients with AIDS/HIV infection or who are otherwise immunosuppressed.

The total NIDCR/NIH amount invested in these projects in FY1998 was approximately \$1.8 million and in FY1999 is approximately \$4M. The NIDCR is partnering with The Wellcome Trust, United Kingdom, to support the sequencing of *Candida albicans*.

C. Future investments in microbial genomics. The NIDCR plans to support complete genomic sequencing of several other bacteria and important pathogens of oral infectious diseases in the next three years. These include the periodontitis pathogens *Fusobacterium nucleatum*, *Bacteroides forsythus*, and *Prevotella intermedia*, and *Streptococcus gordonii* and *Streptococcus sanguis*, which are oral bacteria involved in the formation of dental plaque and, endocarditis, a serious infection of the heart. The genomic sequences of these bacteria may illuminate their biology and ways to control the pathogenesis of disease caused by such endogenous organisms.

NIGMS – National Institute of General Medical Sciences

A. Agency interests in microbial genomics. Within NIGMS, the Division of Genetics and Developmental Biology (GDB) supports research in the area of microbial genomics. NIGMS supports basic biomedical research that is not targeted to specific diseases or disorders. The Institute places great emphasis on the support of individual, investigator-initiated research grants and, in general, does not solicit proposals for specific research initiatives. The Division of Genetics and Developmental Biology (GDB) is responsible for support of virtually all projects that deal with the organization, transmission, and function of the hereditary material. The research grants currently supported by GDB range from studies of fundamental genetic mechanisms, such as DNA replication and gene expression, to studies of population genetics and the more complex regulatory systems that underlie cell growth and development. The majority of these grants support conventional, hypothesis-driven research projects. Consequently, proposals submitted to the NIH for global sequencing or functional analysis of a microorganism are usually assigned to other institutes. However, as described below, GDB does support

projects in the area of genomics when the genomics components are directed towards answering fundamental genetic questions. NIGMS also supports projects aimed at developing novel technologies that will benefit many investigators. Some of these projects, including several in the SBIR Program, are directed towards improving the methods used for yeast functional genomics.

B. Past and current programs/research supported by NIGMS. NIGMS is currently supporting genomic sequencing of the facultative methylotroph, *Methylobacterium extorquens* (see Appendix 1). The project's overall goal is to identify the genes responsible for the unusual phylogenetic amalgam of metabolic pathways that exist in this organism by sequencing 95% of its DNA.

NIGMS-sponsored genomic-scale functional analyses are currently underway in *Escherichia coli* and *Pyrobaculum aerophilum*. The goal of the *E. coli* project is to determine the function of all open reading frames in the genome. NIGMS support for this project reflects the major role that *E. coli* has played as a model for the study of fundamental genetic mechanisms. The goal of the *P. aerophilum* project is to develop tools for functional genomics in a thermophilic archeon. This organism is of particular interest because of its unusual physiology and its position on the phylogenetic tree of hyperthermophiles.

NIGMS also supports a variety of functional genomics studies in *S. cerevisiae*, including those aimed at application of the two-hybrid system to yeast genomics. Major efforts are being made to specify all of the protein-protein interactions in yeast as well as to develop the three-hybrid system in order to detect all RNA-protein interactions. Smaller scale Small Business Innovation Research (SBIR) projects are directed towards establishment of a database for yeast functional genomics and development of a yeast antibody-display system. Other yeast SBIR projects that terminated during the past year have included efforts to develop a dual-bait two-hybrid system and to produce reporter genes for simultaneous monitoring of expression levels of several genes at once. An exploratory study is also underway to develop a genetic system for cloning yeast genes of unknown function.

C. Future investments in microbial genomics. NIGMS recognizes that training the future scientists to work in the broad area of genomics is critical. Research training at NIGMS is funded primarily by institutional pre-doctoral training grants. These awards are intended to provide broad, interdisciplinary training in areas such as molecular and cellular biology and genetics. Thus, support has not been provided for training in more focused areas such as microbial genomics although training of this type could occur as one component of a genetics pre-doctoral award. However, the Institute has recently decided to initiate a new training program in Bioinformatics and Computational Biology. The goal of the program will be to enhance the quantitative skills of students in the life sciences. The initial awards are targeted for July 1, 2001. Although its success remains to be determined, it is very likely that the new program will include training in bioinformatics that will be relevant for individuals intending to pursue careers in microbial genomics.

National Library of Medicine – National Center for Biotechnology Information

A. Agency interests in microbial genomics. NCBI's interests lie in the computational analysis of microbial genomes. In addition to curating and maintaining GenBank, NCBI supports an in-house effort in computational biology and bioinformatics focused on the analysis of microbial genomes.

B. Past and current programs/research supported by NCBI. Comprehensive analysis of microbial genome is a long-term, multi-step project. The first phase involves characterization of all the genes and the encoded proteins and requires, as a pre-requisite, a non-redundant database of microbial protein sequences. The core of the NCBI's effort in this direction is the database of **Clusters of Orthologous Genes (COGs)** that was developed at the NCBI beginning in 1997 and currently includes about 50% of the genes from 20 complete genomes of unicellular organisms (bacteria, archaea and eukaryotes). Each COGs consists of genes from at least 3 different genomes that have been deduced to be likely orthologs, i.e. genes are directly related by vertical evolutionary descent. These orthologous genes, by inference, are likely to be functionally equivalent.

The 4 inter-related research aims in NCBI's work on microbial genomes are to:

- create representative sets of proteins for complete genomes and classify them using the COG approach
 - use COG analysis to refine the representative set (e.g., to include genes that have been missed in original analyses or have frame-shifts due to sequencing errors)
- analyze proteins not falling into COGs
- identify distant relatives to COGs (This ties in with another major project at the NCBI, namely the development of a Conserved Domains Database)
- identify organism-specific families of paralogs, i.e. genes related by gene duplication and not by vertical evolutionary descent
 - work with "expert groups" to provide curated information on specific genes, protein and functional systems, e.g. virulence factors, restriction/modification systems
 - provide computable structural/functional features for all proteins, including signal peptides, transmembrane regions, regions of low complexity, and coiled-coil domains

NIST – NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

A. Agency interests in microbial genomics. The Biotechnology Division at NIST is mandated to develop the generic measurements, models, data, and standards needed to accelerate the commercialization of biotechnology, serving industry, academia, the federal government and international commerce. Within the scope of this mandate, the Biotechnology Division is providing national traceability for measurements in human identification, forensics, and DNA diagnostics, for example, and accurate and reliable data and predictive models. The Biotechnology Division will anticipate and address the next generation measurement needs of the nation by performing cutting-edge research in this area.

B. Past and current programs/research supported by NIST. NIST efforts in microbial genomics derive from its support of research in functional genomics as well as bioinformatics and technology development related to these areas. An example of this is the CARB-TIGR Proteome Project, (supported in part by funding from NIH) to determine the crystal structures of microbial open-reading-frames of unknown function with the intention of gaining insight into the function of the unknown proteins. It is expected that as an outcome of this study, proteins that are expressed in low concentrations or have short lifetimes in these organisms will be identified, potentially providing new drug targets, new industrial catalysts, or novel protein folds and functions.

NIST also supports research in the area of protein detection. Practical proteomics systems that detect and characterize the effects of pharmaceuticals on protein synthesis and accumulation in animal test systems or that can provide the increased sensitivity and dynamic range needed for protein detection in humans are being developed by Large Scale Biology/BioSource through support by the Advanced Technology Program (ATP). Two-dimensional gel electrophoresis is the tool being used to identify and quantitate proteins. This technology will enable the creation of genomic-scale protein databases.

Many industrial firms are taking advantage of the unique capabilities of microbial metabolic pathways to carry out chemical transformations that add value to existing products or synthesize completely new products. Examples include the synthesis of chiral precursors for use in pharmaceutical industry, or manufacture of L-amino acids for feed supplements in agriculture. In many cases, these pathways are not very well characterized. The Laboratory program at NIST, mostly through base-funded in-house activities in the Biotechnology Division, focuses on developing measurement methods, databases, and generic technologies that facilitate the use of microbial metabolic capabilities by industry. Internal funds and resources are leveraged in this activity through participation in the Interagency Working Group on Metabolic Engineering, which supports extramural work mostly in academic laboratories, and by ATP intramural funds

that are obtained by establishing linkages to generic aspects of the extramural work funded by ATP. Several industry-led ATP projects include a focus on novel approaches to metabolic engineering for industrial products such as organic acids and polymers.

Data and Model Systems for Metabolic Engineering – this intramural ATP project supports research in metabolic engineering of organisms aimed at producing industrially relevant products, such as chemical feedstock replacements, polymers and chemical products. In this project, the enzymes in the aromatic amino acid biosynthetic pathway, or chorismate pathway, are being studied using a variety of physical, biophysical and chemical techniques, such as calorimetric, kinetic, x-ray crystallographic and molecular modeling methods, to establish the role of each enzyme in the pathway and the details of each catalytic mechanism. This information will be used to identify areas for enzyme modification in order to optimize the production of product, such as increasing the catalytic turnover, minimizing feedback inhibition, and maximizing the lifetime and reaction rates of the enzymes. This project began in FY 1998 with about \$500K/yr support from ATP and an additional investment of about \$800K/yr from the NIST Biotechnology Division.

Hand-held Devices for DNA Analysis – The development of powerful, low cost devices for convenient DNA analyses is supported by ATP. Affymetrix, Inc and Molecular Dynamics are using advanced manufacturing technologies to fabricate DNA microarrays and electrophoresis instruments. The resulting devices should contribute to the speed and convenience of DNA analysis, facilitate the study of human genetics, aid in managing diseases, such as cancer and AIDS, contribute to new drug discovery and provide DNA diagnostics to the health and other industrial sectors, and, in general, help to reduce costs in the trillion dollar U.S. health care industry.

Immediate DNA Detection – Development of a cartridge-based method of detecting specific DNA sequences at Molecular Innovations was supported by the ATP, in which DNA in a sample is extracted, probed for medically relevant genetic sequences, and only then made visible via a series of reactions that agglutinate dyed beads into colored particles. In another ATP project, Third Wave Technologies was supported for the development of a rapid method for mutation screening that is already saving scientists time and money in the laboratory. Large numbers of samples can be scanned to identify those that are most likely to have mutations. Both of these low-cost fast detection technologies can be instrumental in screening for specific microbial DNA traits. The ATP “Tools for DNA Diagnostics” program supports these and other programs expected to benefit microbial genomics research.

Other technology development funded by NIST – ATP supports rapid sequencing and integrated systems for real-time analysis. Technology for rapid DNA sequencing based on Matrix Assisted Laser Desorption Ionization-Time of Flight (MALDI-TOF) mass spectrometry was designed by GeneTrace Systems, Inc. with support from the ATP. The development combines automated DNA sequencing reactions with MALDI-TOF mass spectrometry. New chemistries, robotics, instrumentation and software reduces the analysis time of a sample from hours to seconds, and

the cost from \$300-\$5000 to a few dollars. Integrated systems for real-time analysis are being developed by Perkin-Elmer Corp. – Applied Biosystems. This project is aimed at advancing real-time detection of biologically-based contamination of plants, food, water, and other environmental samples by using probes that fluoresce only in the presence of specific sample DNA.

Finally, NIST supports bioinformatics through its involvement with the Protein Data Bank. The Protein Data Bank (PDB) is an international repository of three-dimensional macromolecular structural data, funded by NSF, DOE and NIH. NIST is a member of a consortium, the Research Collaboratory for Structural Bioinformatics (RCSB), comprised of Rutgers, the State University of New Jersey, the San Diego Supercomputer Center of the University of California, San Diego, and the National Institute of Standards and Technology, who manage the PDB. NIST has the responsibility of maintaining data uniformity and standards of the data in order to facilitate more complex and comprehensive querying of the exponentially increasing data. The RCSB took over the management of the PDB in FY 1999. NIST Biotechnology Division invests an additional \$500K/yr in this project.

NSF – NATIONAL SCIENCE FOUNDATION

A. Agency interests in microbial genomics. NSF supports microbiological research in a broad range of areas, including environmental and evolutionary biology, metabolic engineering, genetics, mathematics, oceanography, computer sciences, and chemistry to name a few. A recent search of award abstracts for microbial-related research resulted in the identification of almost 5400 grants awarded since 1989. NSF's interest in microbial genomics parallels its support of microbiological research, including plant-associated microbes, microbes of interest in basic research, microbes that occupy critical or compelling environmental or evolutionary niches, microbes developed for metabolic engineering, and microbes that are models for the higher eukaryotic systems supported by NSF, such as plants.

B. Past and current programs/research supported by NSF. NSF supports microbial genomics through a number of established programs. The Microbial Genetics program in the division of Molecular and Cellular Biology supports genomic sequencing and functional genomics of *Halobacterium*, *Neurospora crassa*, *Synechococcus elongatus*, and *E. coli*. The Bioengineering program supports functional genomics projects related to metabolic engineering in *E. coli*. Other programs such as Science and Technology Centers (2 of which are microbiology-based) and Biological Databases and Informatics, the Office of Polar Programs, and Oceanography also fund research related to microbial genomics. In an effort to support broad collaborative research across science, engineering, and education, NSF has developed several cross-directorate programs. These include Life in Extreme Environments (LEn), Microbial Observatories and Biocomplexity programs, which support research in microbial systems and communities, while also fostering genomic-level analysis. Finally, programs in Education and Human Resources (EHR) have supported laboratory research, curriculum development, and education in the area of microbiology and genomics, including the Genome Radio Project.

NSF's funding of large-scale sequencing efforts began in the early 1990's with support of the *Neurospora* genome project, an undergraduate training program and EST-sequencing project based in the EHR-Research Improvement in Minority Institutions program. *Neurospora crassa* is a filamentous ascomycete with complex developmental stages whose genome appears to be very different from that of *S. cerevisiae*. This project is currently supported by Microbial Genetics. Microbial Genetics also supported the sequencing of the largest *Halobacterium* plasmid and supports the sequencing of *Halobacterium* sp. *Halobacterium* is a halophytic archeon that is of interest because of its phylogenetic position within the archaea, its subcellular structure, and its ability to grow in high salt environments. The *Synechococcus* sequencing and functional genomics project is a collaborative project at Texas A&M and the University of Idaho. *Synechococcus* is a cyanobacterium whose circadian rhythms have been studied extensively both at the molecular genetic and physiological levels and whose genome is smaller than the related, sequenced cyanobacterium *Synechocystis*. NSF's investment in microbial genomics overall is \$2.6M with \$2.1M in currently funded grants.

NSF-funded *E. coli* comparative and functional genomics projects include: the comparison of 11 different but related bacterial strains to evaluate the molecular evolution of *E. coli*-related genomes; a study of the integration of two different computational platforms to enable the identification of conserved regulatory elements across whole bacterial genomes; a collaborative project studying metabolic pathways aimed at an *in silico* prediction of phenotypes from genotype leading to the development of a model of genetic circuits; and a genomic-scale screen for membrane proteins in *E. coli*. NSF's current investment in functional genomics is approximately \$900K.

NSF has a major investment in the ever-increasing computational and database needs that provide infrastructure not only for microbial genomics but for many areas of life science. Examples include projects dealing with genomic mapping data, the Ribosome Database Project (RDP), a microbial biodegradation database, and the Protein Data Bank (PDB). All seek to accommodate the rapid growth in storable information, allow complex querying, and facilitate access to data as well as linkage to and integration with other databases. The latter program is funded jointly with NIH and DOE. The current NSF investment in this area is about \$4.5M.

NSF has a strong interest in and commitment to education and has funded approximately \$4.7M in the area of genomics curriculum development, with about \$3.3M in current funding, including an interdisciplinary program for graduate training in bioinformatics and the Genome Radio Program. Again, these apply not only to microbial genomics but to genomics more broadly.

NSF funds two large centers in the area of microbiology, including the Biotechnology Center at the University of Washington that is sequencing the *Halobacterium* genome and the Center for Microbial Ecology at Michigan State University. The NSF investment in these centers since fiscal year 1992 is \$38M. Novel strategies for increasing microbial isolation capabilities are supported by a recent NSF award through the Major Research Instrumentation program. Researchers at the Advanced Microbe Isolation Laboratory at Oregon State University will develop automated approaches for culturing and identifying novel microorganisms from natural ecosystems.

C. Future investments and collaborations in microbial genomics. Microbial genomics is an important component of NSF's future plans. Linking the genome sequence to function of microbes in natural ecosystems is an emerging theme. Currently, programs such as Microbial Genetics, LExEn, Biocomplexity, and others will be responsible for these proposals. A new microbiology postdoctoral program whose goal is to develop a cohort of scientists trained in non-model microbial systems and microbial systematics will be announced for FY2000. NSF also has a post doctoral fellowship program in Biological Informatics to address the need for scientists trained in computational biology and bioinformatics.

NSF collaborates with a number of agencies in microbial genomics, and bioinformatics, and infrastructure projects. For example, NSF, DOE, and NIH jointly provide funding for the Protein Data Bank to a consortium that includes NIST. NSF, USDA, DOE, ONR, and EPA support Metabolic Engineering, which funds genetic circuit research.

USDA – UNITED STATES DEPARTMENT OF AGRICULTURE

A. Agency interests in microbial genomics. The USDA recognizes that microbial genome sequencing and subsequent functional genomics will provide enormous benefits to the agricultural sector. As a mission-linked agency, the USDA supports research in the biological, environmental, physical, and social sciences on regional and national problems relevant to agriculture, food, forestry, and the environment. Genomics is, and will continue to be, a tool of growing importance for studying and solving agricultural-related problems; it is expected to be the driving force for research in the life sciences over the next decade. The USDA anticipates dedicating increased resources to microbial genomics. Without a substantial investment, U.S. agriculture risks losing its competitive edge in the global economy.

B. Past and current programs/research supported by USDA. The Agricultural Research Service (ARS) funds intramural research at USDA laboratories across the country. Except for characterization of novel emerging organisms (primarily viruses), ARS has no programs dedicated to complete sequence analysis for discovery of novel genes or gene products. ARS does however support gene sequence analysis and post-sequencing functional genomics for problem-related research. About 25 viruses or virus groups are being studied in sequencing/genomics programs of various types supported by ARS. They include animal, plant and insect viruses and viroids. Fifteen bacteria or related organisms were identified in programs utilizing genomics. Members of some of these, such as *Salmonella* and *Campylobacter*, have already been sequenced in their entirety (through other funding sources); existing ARS programs will focus on the functional genomics of these organisms. At least 14 genera of fungi were identified in various projects utilizing genomics approaches.

The Cooperative State Research, Education and Extension Service (CSREES) funds extramural projects in microbial genomics. Proposals are funded through the National Research Initiative Competitive Grants Program (NRICGP), which is the USDA's major extramural competitive research grants program. In the NRICGP, investigators that propose whole-genome sequencing currently can submit proposals to two program areas: Animal Health and Well-Being, Plant Pathology. To date, fiscal constraints, rather than a lack of need or desire have limited CSREES' funding of microbial genome projects. In Fiscal Year 1998, the NRICGP's Animal Health and Well-Being Program provided partial support (\$200,000; 1 year) to sequence an avian strain of *Pasteurella multocida*. In the current Fiscal Year 1999, the NRICGP will support a gene discovery project for *Neospora caninum* (\$300,000; 3 years). Approximately 10,000 EST's and 15,000 Genomic Sequence Tags (GST's) will be sequenced and annotated. This effort is synergistic with ongoing genomics work with a related apicomplexan parasite, *Toxoplasma gondii* (supported by the NIH).

The USDA is continuing to sponsor an International Agricultural Microbes Genome Conference to meet the growing needs of this field. The conference will promote communication between and among researchers conducting plant, animal, and soil microbial genome research. Plenary lectures and a poster session will cover: sequencing; technology and bioinformatics; and functional genomics/applications. The meeting will take place in San Diego on January 13-14, 2000. Additional information is available at: <http://www.ag-microbial.org/>.

C. Future investments and collaborations in microbial genomics. Genomics research will steadily increase in ARS. Funding requests have been substantial each year to increase research in all areas of genomics. Twelve new high-throughput sequencers have been purchased for various ARS sites to increase gene sequencing and other genetic analyses. ARS plans to identify the best candidates for genomics analysis among the multitude of agricultural microbes. Promising candidates for complete sequencing projects from the plant bacteria include members of the *Rhizobium/Bradyrhizobium* and related soil organisms, the fruit pathogens, *Xanthomonas* and *Xylella* and the *Phytoplasmas* (mycoplasma-like plant pathogens). In animal health programs, *Anaplasma marginale*, *Pasteurella haemolytica* and *Serpulina hyodysenteriae* would be attractive candidates currently under study in ARS. Among the genomics-related programs of fungi, several strong candidate strains were identified. These included *Rhizoctonia*, *Fusarium*, *Aspergillus* and *Phytophthora*, all of which include members of significant agricultural importance. ARS-funded programs on the parasites in the genera *Neospora* and *Eimeria* also have strong genomics components, making these potential focus organisms. ARS will coordinate the selection of any target organisms with other agencies to avoid unnecessary duplication.

For Fiscal Year 2000, there is strong interest within CSREES to launch a competitive microbial genomics initiative for microorganisms relevant to U.S. agriculture. CSREES envisions initially supporting high-throughput sequencing of the genomes of microorganisms (including viruses, bacteria, fungi and protozoa). Pathogenic and beneficial microbes of animals, plants and soils would be included. Evaluation criteria for relevance include: economic importance; unique biological or environmental features; genetic tractability; and evolutionary significance. The biological question to be addressed will be of critical importance. In studies of pathogenicity, for example, two or more closely related genomes may need to be sequenced. Larger genomes of protozoa and fungi may be partially funded if future plans for completing the work are outlined.

The USDA is very interested in pursuing interagency efforts, similar to the successful collaborations established for genome sequencing of *Arabidopsis thaliana* and rice, for a microbial sequencing initiative and post-sequencing/functional genomics. Collaborations are sought from all interested Federal agencies whose mission areas intersect and complement those of the USDA.

GAPS/OPPORTUNITIES

The contributors to this report have identified a number of areas that, based on the current federal investment in microbial genomics, represent gaps or opportunities for significant contributions to the development of this field. A few of these gaps and opportunities are more salient for some agencies than others, however, the need for infrastructure support, training in genomics, computational biology, and increased integration in policy development was common to all agencies. While some of the gaps/opportunities are specific to microbial genomics, many of them are common issues for all genomic research.

Gaps and Opportunities Specific to Microbial Genomics

Microbes whose genomes are of scientific interest or practical importance that are currently not well represented in public sequencing efforts

- Agriculturally important microbes, both plant and animal pathogens and microbes that are neutral or beneficial to plants and animals, such as nitrogen-fixing bacteria.
- Mollicutes (mycoplasmas, phytoplasmas or spiroplasmas), some of which cause economically important plant and animal diseases, are not part of planned or ongoing sequencing projects.
- Phylogenetically important microbes at evolutionary branch points, such as *Planctomyces* and *Verrucomicrobium* that will help understand the evolution of microbial life.
- Fungi
- Algae
- Difficult-to-culture microbes
- Functional genomics of microbes in extreme environments, including the space environment

Computational Biology and Informatics for Microbial Genomics

- Development and maintenance of user-friendly, organism-based, relational databases.
- Coordination of databases to enhance connectivity, translatability, and interaction among microbial genome and other databases.

- Development of a microbiology and microbial genomics site that coordinates genomics and educational activities, such as the DOE or NHGRI Human Genome web sites (see Appendix)

Training to lay the groundwork for Microbial Functional Genomics

- Microbial biology and genomics at all levels
- Microbial systematics.

Technology and Reagents for Microbial Genomics

- Development and distribution of community-wide resources, e.g. knockout or conditional mutants of model organisms; chip and microarray analyses that are available to the academic community
- Development of a national cDNA library
- Development of novel culture techniques for currently unculturable microorganisms
- Development of techniques for *in situ* genomic analysis of microbial ecosystems to characterize community diversity and gene expression in natural populations

Policy issues for Microbial Genomics

- Policy on uniform data release for microbial sequencing projects.
- Policy on sharing reagents, e.g. novel strains, clones, primer sets or sequences of primers used in published articles, etc.
- Ongoing coordination of microbial genome activities, including targeted microbes, database support, training programs, etc.

General Infrastructure Gaps and Opportunities for Genomics

- Development of high-throughput technology for sequencing and functional genomics, including development of software for analysis of complex data sets. Priority should be given to development of technology that can be made accessible to the broad range of scientists.

- Access to genomic resources, including reagents, microarrays, mass spectrometers (MALDI-TOF), etc.
- Continued development of computational biology and bioinformatics
- Training in the broad area of genomics at all levels
- Training in bioinformatics and computational biology at all levels.
- Training at all levels for students, teachers, and researchers to access, use, and develop new technologies
- Enhancement of technology transfer in the area of genomics to the broader industrial community and more rapid access to these technologies by academic researchers

REFERENCES

1. Whole-genome random sequence of *Haemophilus influenzae* Rd. R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, K. McKenney, G. Sutton, W. Fitzhugh, C. Fields, J. D. Gocayne, J. Scott, R. Shirley, L. I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith, J. C. Venter. *Science* 269: 496-512, (1995)
2. The search for unrecognized pathogens. D. A. Relman. *Science* 284: 1308-1311, (1999)
3. Bacterial biofilms: a common cause of persistent infections. J. W. Costerton, P. S. Stewart, & E. P. Greenberg. *Science* 284: 1318-1322, (1999)
4. Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. K. E. Nelson, R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter & C. M. Fraser. *Nature* 399: 323 – 329, (1999)
5. Is it time to uproot the tree of life? E. Pennisi. *Science* 284: 1305-1307, (1999)
6. Action of BTN1, the yeast orthologue of the gene mutated in Batten disease. D. A. Pearce, T. Ferea, S. A. Nosel, B. Das, & F. Sherman. *Nat. Genet.* 22: 55-58, (1999)
7. Hepatitis C virus core protein binds to a DEAD box RNA helicase. N. Mamiya, H. J. Worman. *J Biol Chem* 274: 15751-6, (1999)

APPENDIX 1 – FUNDING AGENCIES AND ORGANISMS OF RESEARCH

| Funding Agency | Organism | Phylogenetic Branch | Genome Size | Project Description | Notes | URL¹ |
|-----------------------|--|--------------------------------------|--------------------|--|---|------------------------|
| DOD | <i>Bacillus anthracis</i> | Bacteria Gram-positive | 4.5 Mb | Genome sequencing | Pathogen; causes Anthrax; used in biological warfare | a |
| DOD | <i>Plasmodium falciparum</i> Chr10,11 (isolate 3D7) | Eukaryote Slime molds | 2.1 Mb | Genome sequencing | Pathogen; malaria parasite | d |
| DOD | <i>Pyrobaculum aerophilum</i> | Archaea Thermoproteus | 2.2 Mb | Functional genomics | Hyperthermophile | a |
| DOE | <i>Aquifex aeolicus</i> | Bacteria Aquifex | 1.6 Mb | Genome sequenced, functional genomics | Chemolithoautotrophic thermophile | a |
| DOE | <i>Archaeoglobus fulgidus</i> | Archaea Methanosarcina | 2.2 Mb | Genome sequenced, functional genomics | Sulfur-metabolizing hyperthermophile | a |
| DOE | <i>Carboxydotherrnus hydrogenoformans</i> | Bacteria Thermodesulfobacterium | 1.8 Mb | Genome sequencing | Hydrogen production | |
| DOE | <i>Caulobacter crescentus</i> | Bacteria Purple bacteria | 3.8 Mb | Genome sequencing | Used to study reg. of cell cycle events; bioremediation | a |
| DOE | <i>Chlorobium tepidum</i> | Bacteria Green sulfur bacteria | 2.1 Mb | Genome sequenced | Thermophilic photoautotroph, role in global carbon cycling | a |
| DOE | <i>Clostridium acetobutylicum</i> | Bacteria Gram-positive | 4.1 Mb | Genome sequenced | Industrial biotechnology & waste remediation | a |
| DOE | <i>Dehalococcoides ethenogenes</i> | Bacteria Green nonsulfur bacteria | | Genome sequencing | Anaerobic chemoautotroph useful for bioremediation | a |
| DOE | <i>Deinococcus radiodurans</i> | Bacteria Deinococci | 3.3 Mb | Genome sequenced | Chemoorganotroph, highly resistant to desiccation and gamma radiation | a |
| DOE | <i>Desulfovibrio vulgaris</i> | Bacteria Purple bacteria | 1.7 Mb | Genome sequencing | Sulfate-reducing bacteria; bioremediation | a |
| DOE | <i>Geobacter sulfurreducens</i> | Bacteria Purple bacteria | 1.0 Mb | Genome sequencing | Bioremediation | |

Appendix 1 continued

| Funding Agency | Organism | Phylogenetic Branch | Genome Size | Project Description | Notes | URL¹ |
|-----------------------|--|-----------------------------|--------------------|-----------------------------------|---|------------------------|
| DOE | <i>Methanobacterium thermoautotrophicum</i> | Archaea Methanobacterium | 1.8 Mb | Comparative & functional genomics | Methane-producing, moderate thermophile | a |
| DOE | <i>Methanococcus jannaschii</i> | Archaea Methanococcus | 1.7 Mb | Comparative & functional genomics | Autotrophic, methane-producing thermophile (also listed under NASA) | a |
| DOE | <i>Mycoplasma genitalium</i> | Bacteria Gram-positive | 0.6 Mb | Sequenced, functional genomics | Free-living organism with smallest known genome | a |
| DOE | <i>Nitrosomonas europaea</i> | Bacteria Purple bacteria | 1.6 Mb | Genome sequencing | Carbon/nitrogen management | |
| DOE | <i>Nostoc</i> spp. | Bacteria Cyanobacteria | <10.0 Mb | Genome sequencing | Carbon management | |
| DOE | <i>Prochlorococcus marinus</i> | Bacteria Cyanobacteria | | Genome sequencing | Carbon management | |
| DOE | <i>Pseudomonas putida</i> | Bacteria Purple bacteria | 5.0 Mb | Genome sequencing | Opportunistic plant and animal pathogen; bioremediation | a |
| DOE | <i>Pyrobaculum aerophilum</i> | Archaea Thermoproteus | 2.2 Mb | Sequenced, functional genomics | Hyperthermophile; model for high temperature growth | a |
| DOE | <i>Pyrococcus furiosus</i> | Archaea Thermococcus | 2.1 Mb | Genome sequenced | Heterotrophic, sulfur-reducing thermophile | a |
| DOE | <i>Rhodobacter capsulatus</i> | Bacteria Purple bacteria | 3.7 Mb | Genome sequencing | Carbon and nitrogen fixation | k |
| DOE | <i>Rhodospseudomonas palustris</i> | Bacteria Purple bacteria | <5.0 Mb | Genome sequencing | Hydrogen production | |
| DOE | <i>Shewanella putrefaciens</i> Strain: MR-1 | Bacteria Purple bacteria | 4.5 Mb | Genome sequencing | Model bacterial system for reductive dehalogenation reactions; bioremediation | a |

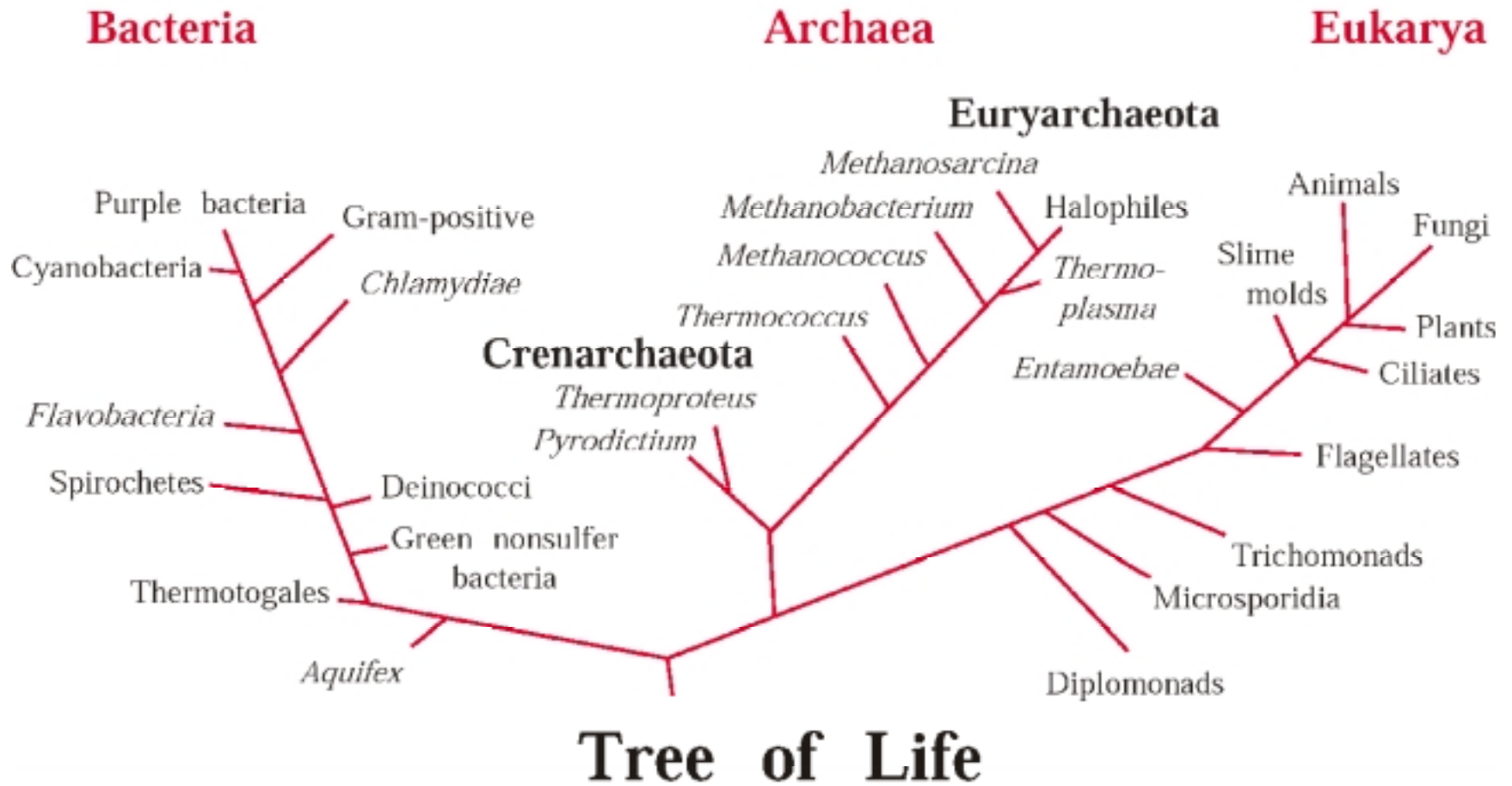
Appendix 1 continued

| Funding Agency | Organism | Phylogenetic Branch | Genome Size | Project Description | Notes | URL¹ |
|-----------------------|---|-----------------------------|--------------------|---|---|------------------------|
| DOE | <i>Thermotoga maritima</i> | Bacteria Thermotogales | 1.9 Mb | Genome sequenced, functional genomics | Hyperthermophile; model for molecular mechanisms of protein thermostability | a |
| DOE | <i>Thiobacillus ferrooxidans</i> | Bacteria Purple bacteria | 2.9 Mb | Genome sequencing | Obligate chemolithoautotroph; carbon management, bioremediation | a |
| NASA | <i>Methanococcus jannaschii</i> | Archaea Methanococcus | 1.7 Mb | Comparative & functional genomics | Autotrophic, methane-producing thermophile (also listed under DOE) | a |
| NHGRI | <i>Escherichia coli</i> K-12 | Bacteria Purple bacteria | 4.6 Mb | Genome sequencing | Enteric bacteria | b |
| NHGRI | <i>Saccharomyces cerevisiae</i> | Eukaryote Fungi | 12.1 Mb | Genome sequencing, functional genomics | Yeast; model system for genetics & molecular bio. | c |
| NIAID | <i>Chlamydia trachomatis</i> Strain: serovar D | Bacteria Chlamydiae | 1.0 Mb | Genome sequencing | Obligate intracellular pathogen, STD | d |
| NIAID | <i>Enterococcus faecalis</i> Strain: V583 | Bacteria Gram-positive | 3.2 Mb | Genome sequencing | Pathogen | d |
| NIAID | <i>Escherichia coli</i> Strain: O157:H7 | Bacteria Purple bacteria | 4.6 Mb | Genome sequencing | Enteric bacteria; can cause severe foodborne disease | d |
| NIAID | <i>Giardia lamblia</i> Strain: WB 5 chrom. | Eukaryote Diplomonads | | Genome sequencing | Pathogen | d |
| NIAID | <i>Leishmania major</i> Strain: Friedlin 36 chrom. | Eukaryote Slime molds | 0.3 Mb | Genome sequencing | Pathogen; causes cutaneous leishmaniasis | d |
| NIAID | <i>Mycobacterium tuberculosis</i> Strain: CSU 93 | Bacteria Gram-positive | 4.4 Mb | Genome sequencing | Pathogen; virulent clinical isolate | d |
| NIAID | <i>Neisseria gonorrhoeae</i> Strain: FA1090 | Bacteria Purple bacteria | 0.2 Mb | Genome sequencing | Pathogen; causes gonorrhea (STD) | d |

Appendix 1 continued

| Funding Agency | Organism | Phylogenetic Branch | Genome Size | Project Description | Notes | URL¹ |
|-----------------------|---|-----------------------------|--------------------|----------------------------|--|------------------------|
| NIAID | <i>Plasmodium falciparum</i> Strain: 3D7 chr2,10,11 | Eukaryote Slime molds | 3.1 Mb | Genome sequencing | Pathogen; malaria parasite | d |
| NIAID | <i>Salmonella typhimurium</i> | Bacteria Purple bacteria | 4.5 Mb | Genome sequencing | Pathogen; enteric bacteria | d |
| NIAID | <i>Staphylococcus aureus</i> Strain: 8325 | Bacteria Gram-positive | 2.8 Mb | Genome sequencing | Pathogen; coagulase-positive staphylococci | d |
| NIAID | <i>Staphylococcus aureus</i> Strain: COL | Bacteria Gram-positive | 2.8 Mb | Genome sequencing | Pathogen; coagulase-positive staphylococci | d |
| NIAID | <i>Streptococcus pneumoniae</i> Strain: 23F and 4 | Bacteria Gram-positive | 2.1 Mb | Genome sequencing | Pathogen; common cause of bacterial pneumoniae | d |
| NIAID | <i>Streptococcus pyogenes</i> Strain: M1 serotype class 1 rheumatogenic strain | Bacteria Gram-positive | 2.0 Mb | Genome sequencing | Pathogen; facultative anaerobe | d |
| NIAID | <i>Treponema pallidum</i> Strain: Nichols | Bacteria Spirochetes | 1.1 Mb | Genome sequencing | Pathogen; causes syphilis (STD) | d |
| NIAID | <i>Trypanosoma brucei</i> Strain: TREU 927/4 | Eukaryote Slime molds | | Genome sequencing | Pathogen | d |
| NIAID | <i>Ureaplasma urealyticum</i> Strain: Serovar 3 | Bacteria Gram-positive | 0.8 Mb | Genome sequencing | Pathogen | d |
| NIAID | <i>Vibrio cholerae</i> Strain: N16961, Serotype 01 | Bacteria Purple bacteria | 2.5 Mb | Genome sequencing | Pathogen; causes cholera | d |
| NIDCR | <i>Actinobacillus actinomycetemcomitans</i> | Bacteria Gram-negative | 2.2 Mb | Genome sequencing | Oral pathogenic bacteria; causes periodontitis | g |
| NIDCR | <i>Candida albicans</i> | Eukaryote Fungi | 16.0 Mb | Genome sequencing | Opportunistic yeast; causes oral mucositis | h |
| NIDCR | <i>Porphyromonas gingivalis</i> | Bacteria Gram-negative | 2.2 Mb | Genome sequencing | Oral pathogenic bacteria; causes periodontitis | b, i |
| NIDCR | <i>Streptococcus mutans</i> | Bacteria Gram-positive | 2.2 Mb | Genome sequencing | Oral pathogenic bacteria; causes dental caries | b, j |

APPENDIX 2



Tree of Life, from National Center for Genome Resources (<http://www.ncgr.org/graphics/microbe/purple.gif>)

