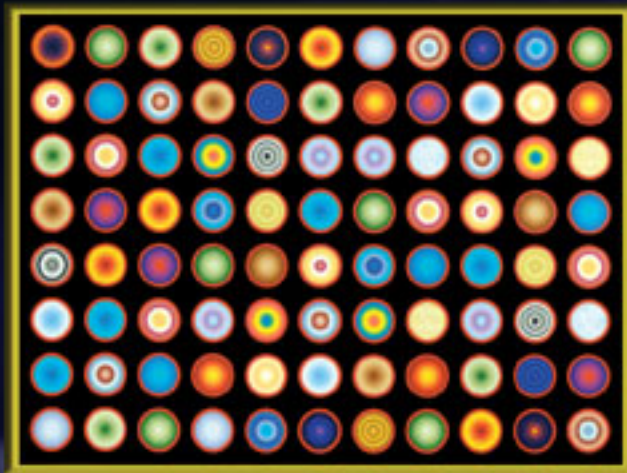
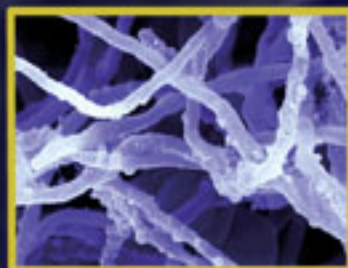
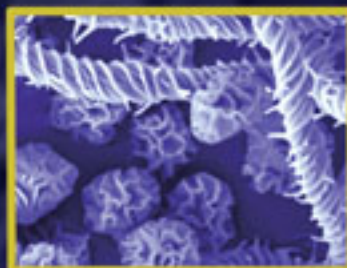


The Microbe Project:

A BIO Advisory
Committee Workshop



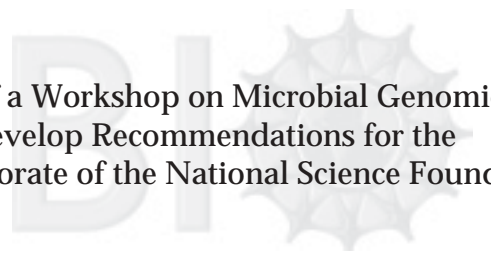
Report of a Workshop on
Microbial Genomics to
Develop Recommendations for the
Biological Sciences Directorate
of the National Science Foundation



The Microbe Project:

A BIO Advisory Committee Workshop

Report of a Workshop on Microbial Genomics to
Develop Recommendations for the
BIO Directorate of the National Science Foundation



Workshop Dates: August 10-11, 2000

Conveners:

Claire M. Fraser

The Institute for Genomic Research
cmfraser@tigr.org

John Wooley

University of California, San Diego
jwooley@ucsd.edu

Table of Contents

About the Workshop	4
Recommendations.....	5
Introduction	7
Background and Challenges	9
Sequencing and Biodiversity.....	9
Bioinformatics and Databases.....	11
Functional Genomics.....	11
Recommendations.....	12
Sequencing and Biodiversity (Goals).....	13
Bioinformatics and Databases (Goals).....	14
Functional Genomics (Goals).....	15
Education and Workforce Issues.....	16
Outreach to Scientific Organizations and the Public.....	17
Appendix	19
Objectives of the Workshop.....	19
Statement of Purpose and a List of Topics.....	20
BIO AC Members, Co-Conveners.....	21
Invited Participants.....	21
NSF Observers.....	22
Location of Meeting and Dates.....	22
References.....	22
Schedule	23

About the Workshop

The Microbe Project: A BIO Advisory Committee Workshop

August 10-11, 2000

Report of a Workshop on Microbial Genomics to develop recommendations for the BIO Directorate of the National Science Foundation.

Conveners:

Claire M. Fraser
The Institute for Genomic Research
cmfraser@tigr.org

John Wooley
University of California, San Diego
jwooley@ucsd.edu

“The purpose of the BIO Advisory Workshop was to provide advice that will help crystallize NSF’s role in the Microbe Project.”

Microbes were the first organisms on earth and predated animals and plants by more than 3 billion years. They are the foundation of the biosphere — both from an evolutionary and an environmental perspective (1). It has been estimated that microbial species make up about 60% of the Earth’s biomass. The genetic, metabolic, and physiological diversity of microbial species is far greater than that found in plants and animals. And yet the diversity of the microbial world is largely unknown, with less than one-half of 1% of the estimated 2-3 billion microbial species identified. Of those species that have been described, their biological diversity is extraordinary, having adapted to grow under extremes of temperature, pH, salt concentration, and oxygen levels.

Perhaps no other area of research has been so energized by the application of genomic technology than the microbial field. It was just five years ago that the first complete genome sequence for a free-living organism was reported (2), and since that first report more than thirty-five microbial genome sequences have been published, with more than one hundred other projects underway (3). This progress has represented, on average, one completed genome sequence every two months and all indications point to this pace continuing to accelerate. Included in the first completed microbial projects are many important human pathogens; the simplest known free-living organism; “model” organisms *E. coli* and *B. subtilis*; thermophilic bacterial species that may represent some of the deepest branching members of the bacterial lineage; five representatives of the archaeal domain; and the first eukaryote, *Saccharomyces cerevisiae*. While this level of progress may seem impressive, it should be stressed that this is only the tip of the iceberg in terms of microbial diversity.

The application of genomics to the study of microorganisms provides a unique opportunity to learn about the unity and diversity of life on this planet. All of the organisms that have been studied to date by whole genome analysis are species that can be grown either in the laboratory or in animal cells. It is important to remember that the vast majority of microbial species cannot be cultivated at all, and these organisms, which live in microbial communities, play essential roles in the overall ecology of the planet. Nevertheless, the study of “laboratory-adapted” microbes has had a profound impact on our understanding of the biology and the evolutionary relationships among microbial species. For example, these efforts have uncovered entirely new metabolic pathways, have accelerated the study of gene regulation in microbial species, have revealed that approximately one-half of all predicted coding sequences are of unknown biological function, and have suggested that lateral gene transfer among organisms has played a significant role in the evolution of microbial species and microbial diversity. None of these insights would have been possible without genome sequencing and analysis. This vast amount of new information has provided an entirely new starting point for investigations in both basic and applied areas of research. The payoffs from these efforts will be significant and will promote advances

Agencies involved in the Microbial Genome Project at time of workshop: NIH, USDA, DOE, NASA, and NSF.

in microbiology and related disciplines, systems biology, drug and vaccine design, and industrial and environmental processes.

Following on from the Interagency Report on the Federal Investment in Microbial Genomics, a new interagency working group is being convened to develop a coordinated, interagency effort, now called "The Microbe Project." NIH, USDA, DOE, and NASA in addition to NSF are involved thus far, and other agencies may join. Each agency's mission will dictate its primary role in the Microbe Project. The NSF role clearly is basic science related to microbial diversity, including microbes in the environment. It seems likely that in future years joint programs may be developed through collaborations among agencies.

Given the rapid progress in the field of microbial genomics, it is perhaps not surprising that a number of challenges have emerged, and certain areas that deserve attention have been overlooked. These include, for example, the need to address priorities for future sequencing projects, to determine the role of small vs. large sequencing groups in the overall enterprise, to define and adopt standards in gene annotation, to develop consistent and fair policies governing data release by sequencing groups and its use by the scientific community, to establish mechanisms for long-term investment in databases and software for data mining and manipulation, to meet the challenges and opportunities in the area of functional genomics, and to train the next generation of genome investigators. To address all these challenges will require expanded cooperation and coordination among the governmental agencies that fund this work.

The purpose of the BIO AC workshop was to provide advice that will help crystallize NSF's role in The Microbe Project. The workshop summarized the accomplishments, challenges, and opportunities in the microbial genomics field that are relevant to the NSF; provided some direct advice such as criteria for selection of microbes that should be sequenced with NSF support and areas for future development in the bioinformatics arena; and identified issues that should be addressed in greater detail in future workshops or other venues including informatics, standards for annotation, and infrastructure needs.

Recommendations

The specific recommendations to the BIO Directorate of the NSF resulting from this workshop can be summarized as follows:

1. Support genomics-based research to elucidate the content and organization of genes in the biosphere. These programs will answer such questions as:
 - How are genes distributed among organisms, and why?
 - How do genes define the interactions of organisms with the environment?

The Microbe Project: A BIO Advisory Committee Workshop



Penicillium. Photo courtesy of Richard Edelman.

- How is biological diversity between and among organisms achieved and maintained?
- How do organisms interact with each other in the environment?

It is anticipated that large-scale DNA sequencing and analysis will play a critical role in these projects. A wide range of research activities were recommended as a means of addressing questions that included:

- complete genome analysis of cultivated prokaryotic species to achieve phylogenetic breadth
- rDNA and genome inventories of multiple environments to understand species distributions and the effect of environment on organism activities such as nutrient and energy cycling

Both viral and lower eukaryotic protists should be included in these activities to gain a comprehensive view of environmental diversity and species relationships.

2. Support development of accurate, automated genome analysis tools.

Support creation of community-specific databases that integrate genomic and experimental information, undergo continuous curation by experts in the field, and interoperate with other databases.

Define and further develop standards for genome annotation that facilitate comparative genomics studies and database interoperability.

- ### 3. Support further development of readily-affordable technologies for genome-enabled science and the requisite tools for handling ever-increasing amounts of functional genomics data. Specific goals include the development of technologies to interrogate every open reading frame in an organism/population, to systematically determine the function of unknown genes/proteins, to measure gene activities in the environment at the level of a single cell, and to study the functional interactions of all organisms in the biome.
- ### 4. Support training programs that foster collaboration and cross-training of scientists from a variety of disciplines including but not limited to microbiology, ecology, genomics, computer science, and bioinformatics. Training programs should be created not only for undergraduates, graduate students, and post-doctoral fellows, but also for established investigators and may consist of training grants, symposia, and intensive summer courses, as examples.
- ### 5. Support educational outreach programs that communicate the excitement and practical importance and benefits of microbial genomics to other scientific communities and to the general public. All too often good science is not funded because the relevance of the research to the non-specialist is not understood. This is a time and an area with potentially great pay-off for society, and it is imperative that this message be effectively communicated.

Introduction



Genome Map

Microbes of agricultural importance, animal pathogens, and microbial extremophiles are largely underrepresented in the more than 120 microbial sequencing projects that have been completed or funded.

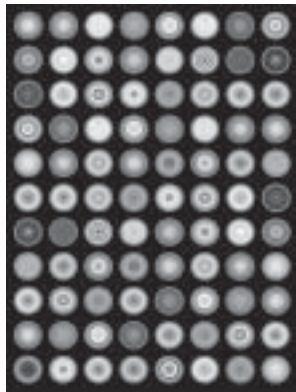
Genomic science is essential for understanding all levels of biological organization from single cells, to the biosphere, and beyond. This will become increasingly apparent as the emphasis in the genomics field shifts from primarily one of generating DNA sequence information to data management and interpretation. There are several ways in which genomics can impact our understanding of microbial life on Earth: (1) Whole genome sequencing of organisms that can be grown in culture has already demonstrated its power in revealing the biochemical and metabolic capabilities of these species; (2) partial genome sequencing can be used to interrogate various environments to provide a preliminary catalog of information on gene identity and function; (3) functional genomics approaches (genome-enabled science) can now begin to elucidate how individual cells and communities work together in an integrated fashion; and (4) databases and bioinformatics tools can allow investigators to fully exploit ever-increasing amounts of genome data to generate new hypotheses about the relationships between organisms and the properties of cells and communities; promote advances in the biotechnology industry, in food production, and in drug design; and understand the origin of life and how it continues to evolve.

More than 120 microbial genome sequencing projects have been completed or funded. The organisms represented in this list are heavily weighted toward human pathogens, with smaller numbers of non-pathogenic bacteria and archaea included. Superficial phylogenetic coverage of the major bacterial groups has been achieved in the selection of species for whole genome analysis; however, microbes of agricultural importance, animal pathogens, and species that inhabit extreme environments have been underrepresented. Moreover, there have been no projects funded to date to tackle unculturable microbial species, which presumably represent the greatest diversity in the microbial world. For example, less than 1% of the bacterial species in the oceans have been cultured, and there are roughly a million of these unknown cells per ml of seawater.

DNA sequencing technology has evolved rapidly over the past ten years and the methods for generating sequence have become faster and more efficient. As a result, the costs for generating a finished DNA sequence has dropped from ~\$1 per base pair in the early 1990s to less than \$0.20 per finished base pair in large sequencing centers. DNA sequencing costs are far less of an obstacle in identifying organisms for genome analysis than they were even three or four years ago, and as a result additional species and organisms with larger genomes (>4 million base pairs) have now been targeted for sequencing.

Identification of genes in prokaryotic genomes has advanced to the stage where nearly all protein coding regions can be identified with confidence. Computational gene finders now routinely find over 99% of protein coding regions and RNA genes. Once the protein coding genes are located, the most challenging problem is determining their function. Typically, about 40-60% of

The Microbe Project: A BIO Advisory Committee Workshop



DNA Microarray
Sequencing data

“There has been little coordinated effort among groups involved in bioinformatics research, and as a result, no standards for genome annotation, curation, and database structure have emerged.”

the genes in a newly sequenced bacterial genome display detectable sequence similarity to protein sequences whose function is at least tentatively known. This sequence similarity is the primary basis for assigning function to new proteins, but the transfer of functional assignments is fraught with difficulties. The remainder of the genes in a completed sequence don't resemble any genes of known function; in those cases, sequence alone doesn't reveal anything about the biological role of the protein product. Better databases and database management tools would go a long way towards maximizing the impact of DNA sequence data that is accumulating in public repositories. The field of bioinformatics – the marriage of biology and computer science – has exploded in recent years out of necessity. However, there has been little coordinated effort among groups that are involved in bioinformatics research, and as a result, no standards for genome annotation, curation, and database structure have emerged. As the amount of genome information continues to grow, it will become increasingly difficult to navigate all of the existing databases that have appeared.

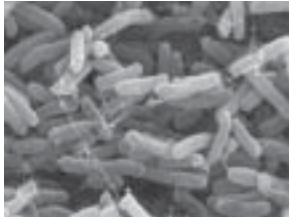
Complete microbial genome sequences represent just the beginning of a new field of genome-enabled science. Comparative genomics and *in silico* studies have begun to reveal insights into protein function and regulation that would not have been possible just a few years ago. It is now possible to think about biological investigations at the whole genome level. For example, with microarray analysis every gene in a microbial genome can be represented on a glass slide. Using this kind of approach, investigators can interrogate the level of expression of all genes in a species under various experimental conditions. This kind of approach harnesses the power of genomic technology in a way that investigators could not dream of only ten years ago. Other approaches such as a two-hybrid analysis or genome-wide knock-out experiments are also providing large-scale biological insights into the workings of microbial cells.

The microbial genomics field is at an important crossroads. Our vision of the future must be built upon today's efforts to generate complete genome sequence information, to develop databases in which to store and make available information from genomics and functional genomics projects, and to develop novel methods for follow-up biological work on a genome-wide scale. While progress in the genomics field has proceeded at a dizzying pace, it is now time to reevaluate where the field is going and how to best achieve the collective goals of the community. Genomic approaches to microbiology have unlocked so many new avenues of investigation. One of the challenges going forward will be how to best leverage available funding to maximize the return on this research investment.

Background and Challenges

Workshop Charge:

To develop a set of recommendations for the BIO Advisory Committee regarding unique research priorities and goals for the NSF in genome-enabled microbiological science in light of the overall federal investment in this area.



Escherichia coli.

Photo courtesy of Shirley Owens. Image reproduced by permission from Microbe Zoo, © 1997 Michigan State University.

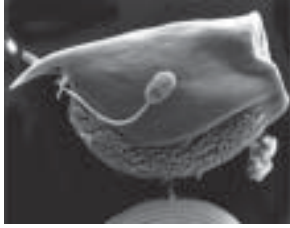
Sequencing and Biodiversity

Genomics is the latest and most exciting tool aimed at unraveling the incredible complexity of microbial life on earth. The field of microbial genomics has made enormous progress during the past five years since the first report on the completion of the genome sequence from a free-living organism in 1995. As of December 2000, the complete sequences of thirty-five microbial species have been published in the scientific literature, with preliminary data from ongoing projects on ~100 other species available via the Internet. The ~135 microbial genome sequencing projects funded to date have focused exclusively on organisms that can be grown in culture or in animal cells. This approach cannot provide a true picture of the diversity of life on earth. For example, culture-independent methods have shown that 30-40% of the cells in ocean waters represent archaeal species and are major biological components of other habitats such as lake sediments and forest soils.

The technology now exists to recover very large fragments of DNA from natural, complex environmental samples. The diversity of the microbes represented in these environmental genome studies appears to mirror that originally identified via 16S rRNA studies. The availability of large insert DNA clones has facilitated a new approach to understanding the genome composition of organisms present in environmental samples, and this provides a link between genomics and microbial ecology studies. For example, De Long and colleagues have recently identified a halophile-like bacteriorhodopsin gene, which had previously only been known in Archaea, in a member of the Bacteria. The proteorhodopsin can be expressed in *E. coli* and binds retinal. This bacterial species represents a new and important type of phototroph in the ocean and the potential exists for the identification of other proteorhodopsin variants in other bacterial species that absorb light at different wavelengths. In addition to the new biology described by this study, these findings have implications in the field of nanotechnology. Bacteriorhodopsins can be used in biofilms as optical switches in optical computers. This is an immediate practical application to come from basic research in microbial genomics.

In addition to identification of new species of microorganisms, genomics also has the potential to provide a comprehensive picture of microbial communities and consortia. As individual microbial genomes are mosaics of genes from mixed heritage, a microbial community is a collection of gene functions distributed among its individual members. No single organism contains all of the genes necessary to carry out the diverse biogeochemical reactions that represent microbial community function. Because microbes mediate and control all major pathways of carbon flow and flux, it will be extremely important to integrate models of biogeochemical response and climatic change with microbial component structure and function. To make this science predictive, we need to understand the microbial component and its

The Microbe Project: A BIO Advisory Committee Workshop



Entodinium caudatum.

Photo courtesy of Mel Yokoyama and Mario Cobos. Image reproduced by permission from Microbe Zoo, © 1997 Michigan State University.

mechanistic interactions. In other words, we need to understand the micro-structure of microbial communities in the sea and in the soil, and genomics is the key to helping us interrogate the systems.

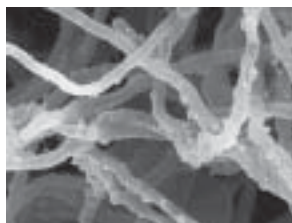
Analysis of complete genome sequences is beginning to provide a great deal of insight into many questions concerning the evolution of microbes. One area of insight has surrounded the occurrence of genetic exchanges between different evolutionary lineages – a phenomenon known as horizontal, or lateral, gene transfer. Prior to the availability of complete genome sequences, studies of horizontal gene transfer were limited because of the incompleteness of the data sets being analyzed. Analyses of complete genome sequences have led to many recent suggestions that the extent of horizontal gene exchange is much greater than was previously appreciated. Although large numbers of genes appear to undergo lateral transfer, analysis of certain sets of genes supports traditional phylogenetic trees. The question then becomes how is lateral gene transfer regulated, and if lateral transfer is common, why don't all genes seem to move in this manner?

The best evidence for there being a “core set of genes” for each evolutionary lineage comes from the construction of “whole genome trees” based on the presence and absence of particular homologs or orthologs in different complete genomes. It is important to note that gene content trees represent averages of patterns produced by phylogeny, gene duplication and loss, and horizontal transfer and, thus, are not real phylogenetic trees. Nevertheless, the fact that these trees are very similar to phylogenetic trees of genes such as rRNA and RecA suggests that although horizontal gene transfer may be extensive, it is somehow constrained by phylogenetic relationships. Other evidence for a “core” of particular lineages comes from the finding of a conserved core of euryarchaeal genomes and the finding that some types of genes may be more prone to gene transfer than others.

Because bacterial “species” result from a combination of linear descent and lateral gene transfer, it is essential that microbial diversity, gene exchange, and phylogeny be studied concurrently. This realization has resulted in a desperate need for additional genome sequence data from both phylogenetically distant and close organisms (as defined by 16S rRNA) from similar habitats. We also need more information from partial genomic sequencing of all organisms in a single environment.

Which genomes should come next? Up until this time, genome projects have focused on organisms that can be maintained in culture and that are easy to identify – marine autotrophs, human pathogens, extremophiles. Our future genomics work needs to look at members of all communities: heterotrophs, symbionts, Archaea, mobile elements (phage), and fungi. What protists would we want to focus on? Anaerobic protist genomes would be a good place to start. One interesting group is amitochondriate protists such as Entamoeba. Did these protists have mitochondria and lose them, or did their mitochondria, in some cases, become hydrogenosome? Genomics can help us answer whether or not there was a single mitochondrial origin, which is very important in trying to understand whether some of these protists are more basal in the eukaryotic tree than others.

“Future genomics work needs to focus on members (microbes) of all communities including heterotrophs, symbionts, Archaea, phage, and fungi.”



Wood degrader --Phanerochaete chrysosporium.

Photo courtesy of Fred Michel.

Current challenges include the lack of standards among groups for genome annotation; the lack of uniformity in data presentation, accessibility, output capabilities, and archiving of software versions; and relatively poor software documentation.

Bioinformatics and Databases

The pace at which the genomics field has moved forward during the past decade has had a profound effect on the emerging field of bioinformatics. The software tools for managing large-scale sequencing projects, for gene identification and annotation, and for database development, for example, are not commercially available, and as a result, both large and small genome centers have developed their own suites of software and tools for handling large sets of genome data. This has led to a series of problems that have become apparent with ~ thirty-five complete genome sequences available and that will only get worse as DNA sequence information and data from functional genomics studies continues to increase. Current challenges include the lack of standards among groups for genome annotation; the lack of funding for ongoing curation of many existing databases; the lack of uniformity in data presentation, accessibility, output capabilities, and archiving of software versions; and relatively poor software documentation. As additional and more complex data come on-line, these challenges will be increased by the lack of database interoperability due to the use of heterogeneous software systems, database schemas that make it difficult to search different types of data, the lack of universal tools for uploading and downloading files or searching databases containing different types of data, and the existence of multiple model organism databases (MODs) that are only poorly linked to each other.

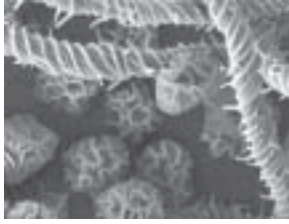
In the near future, many of these deficiencies will need to be addressed if the scientific community is going to be able to fully exploit the information available from a myriad of sources. Some of the needs of diverse communities include development of technology including shared database software tools, a means of querying multiple databases simultaneously, a means of using databases to provide new insight about biological function, a means of using databases for predictive capabilities such as metabolic pathway complement from genome sequence and phenotype of knock-out mutant, a means of carrying out large-scale genome annotation and comparative analyses, and new software for understanding metabolic pathways.

Functional Genomics

The genomics field is a highly disciplinary science, and the need for an interdisciplinary approach is perhaps best exemplified by the current set of opportunities and challenges in the functional genomics arena. Functional genomics is a term that is frequently used but one that has different meanings to different people. It is possible to think about functional genomics in three different but related aspects: scientific, technological, and applied.

The goals of functional genomics from a scientific perspective are to learn how cells and organisms work in an integrated manner and how different taxa have evolved different mechanisms to solve biological problems. The kind of information that we need to address these global questions include such things as the function, localization, movement, post-translational modification, and activity of macromolecules in the cell; the steady-state and non steady-state concentrations of RNA, proteins, and metabolites; and the molecular interactions and integration of these biological molecules and signals.

The Microbe Project: A BIO Advisory Committee Workshop



SEM image of Protist Hermitrichia serpula. Photo courtesy of Shirley Owens. Image reproduced by permission from Microbe Zoo, © 1997 Michigan State University.

The goals of functional genomics from a technological perspective relate to measurement and modeling; development of high throughput, robust technologies to determine the state of a cell at every level of organization; and use of computational tools to allow rapid integration of information and identification of “emergent properties” of a biological system. A number of technologies already exist that can be brought to bear on these questions including knock-out technologies, localization technologies, two-hybrid analysis, microarray and related technologies, 2-D PAGE, mass spectrometry, and X-ray crystallography, for example. Computational analyses are an integral part of this equation and longer-term goals will be to predict function from structure, carry out higher-level analyses of complex datasets, and use this information to model and predict the behavior of biological systems.

The goals of functional genomics from an applied perspective relate to discovery and prediction or intervention in biological systems. This kind of approach has the potential to provide solutions to current problems in medicine, ecology, agriculture, defense, and engineering. No single approach will be sufficient to adequately address any biological question going forward and the need for integration of approaches and disciplines will become ever more urgent.

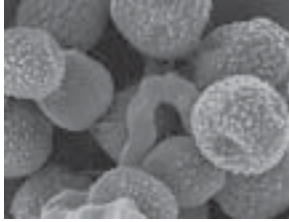
Recommendations

“The National Science Foundation has the opportunity and responsibility to develop a microbial genomics research agenda that is unparalleled in its depth and breadth of coverage.”

Unlike other funding agencies such as the NIH, DOE, USDA, and DOD that fund specific areas of microbial genomics research, the National Science Foundation has the opportunity and responsibility to develop a research agenda in this area that is unparalleled in terms of its depth and breadth of coverage. The NSF can look beyond issues of human health and bioremediation to put in place a series of initiatives that will redefine our understanding of the diversity of life on Earth, that will more fully define the importance of microorganisms in the health and longevity not only of all the species that inhabit the Earth but also that of our planet, and that will reveal the relationships between all species in the continuum of life on Earth. Therefore, a NSF-supported program in microbial genomics should include a number of broad, interdisciplinary projects that address fundamental questions in basic microbiology, evolutionary biology, microbial ecology, population biology, and comparative genomics. The experimental approaches that will be taken include everything from DNA sequencing to functional genomics to studies of populations of uncultured organisms to development of new technologies for genome-enabled science. Selection of projects for funding under this initiative should, for the most part, be driven by biological questions; however, the participants agreed that there is also value in funding genomics efforts that may fall outside of the traditional hypothesis-driven research. While the participants agreed that microbial genomics/genome-enabled science represent a continuum of activities that are interrelated, for the ease of discussion, the following recommendations have been separated into three main areas: sequencing and biodiversity, bioinformatics and databases, and functional genomics.

Sequencing and biodiversity

Genome sciences will play a key role in our understanding of all levels of



Slime mold spores.
 Photo courtesy of Shirley Owens. Photo reproduced by permission from Microbe Zoo, © 1997 Michigan State University.

biological organization from single cells to the biosphere and beyond. Genome sequencing projects provide a biological parts list for an organism or a population that is an essential starting point for understanding how these parts work together to create a living cell or a functional ecosystem and how processes such as photosynthesis and nutrient cycling evolved and operate today. Genome sciences have the potential to change the way in which environmental research is approached, with tremendous benefit to be derived from integration of information about the microbial composition of a given environment with ecosystem models. Genome information will enable investigators to begin to address general questions such as “how are genes distributed among organisms and why?” and “how do genes define the interactions of organisms with the environment?”. As biologists and engineers collaborate in these efforts, genome sciences will begin to support development and discovery of new biomaterials and have a tremendous impact on bioengineering and nanotechnology.

Four primary areas of focus were defined as being essential to understanding microbial life on Earth. The first area focused on gene and genome inventories with long-range research goals of determining (1) patterns of gene distribution among microorganisms; (2) interactions between genes, genomes, and the environment; and (3) the foundation of organism-organism interactions. A second area focused on elucidation of the processes and patterns that govern gene distribution among species. This will require an understanding of phenomena such as vertical gene inheritance vs. lateral gene transfer, gene duplications and coalescence, and invention or recruitment of genes for new activities. A third area was directed toward understanding how genes define interactions with the environment. Relevant questions to be addressed include how environmental parameters affect gene distributions among species vs. how gene distributions among species affect environmental parameters, the effects of organisms on the environment (nutrient and energy cycling), the effect of the environment on organism activities, the development of methods for directed and correlative studies of organisms in their environments, and whether the activities of a group of organisms can be used as a biosensor of environmental conditions. The last area is related to organism-organism interactions and defining the relationships that exist, which are essential for the survival of microbial populations, consortia, symbioses, and endosymbioses.

To be able to address these questions, it will be necessary to expand genome sequencing efforts beyond those that are currently underway. To have a sufficient amount of molecular information for carrying out the above analyses, the following milestones were suggested:

Project Type	2 year goal	5 year goal
Cultivated prokaryotic genomes	100	10,000
Cultivated eukaryotic protist genomes	10	100
Viruses	100	
rDNA-based inventories	1000 environments	10,000 environments
Environmental genomic inventories	10 environments	100 environments

Note: In this table, genomes refer to complete genome sequences whereas inventories refer to partial genome sequences.

One of the most important objectives in initially selecting cultivated prokaryotic species for genome sequencing projects should be to increase the representation of phylogenetic breadth, followed by emphasis on more closely-related species. It is very difficult to make recommendations for time periods beyond 2-3 years given the rapid rate of technological development in large-scale DNA sequencing. There is great potential for continued reduction in the cost of DNA sequencing, making today's seemingly ambitious goals more feasible tomorrow. A tremendous capacity for large-scale DNA sequencing exists in academic and industrial genome centers around the world. The participants agreed that there is no need to fund the development of new infrastructure related to genome sequencing; rather, the goal should be to make best use of existing capacity through scientific collaborations or the establishment of a virtual genomics center that would provide support for some aspects of these projects while leaving other aspects to be carried out in individual research laboratories.

Bioinformatics and Databases

“Bioinformatics and databases are critical to the infrastructure and success of any genomic research.”

Bioinformatics and databases are critical to the infrastructure and success of any genomic research. While it has been possible to automate essentially all aspects of the process for generating large-scale DNA sequences, there are still tremendous bottlenecks in genome annotation, even for large sequencing centers, and too many inconsistencies in standards for genome annotation, which have hampered comparative genome analyses. Moreover, a number of specialized databases now exist that are poorly linked to each other, presenting yet another set of obstacles that must be overcome if this information is to be widely available to the scientific community. As the number of functional genomic analyses conducted increases, the lack of uniformity in data presentation and access are increasingly important issues. Without uniformity, it is difficult or impossible to perform high throughput comparisons of data from different laboratories. The propagation of annotation errors and the non-uniformity of annotation across genomes make discovery of gene and protein function ever greater challenges. Three major goals were defined that address the current limitations in the area of bioinformatics and databases.

The first goal is to develop a set of tools for accurate, automated genome analysis to support genome assembly and gene finding; prediction of biological function; and prediction of metabolic, signaling, and regulatory pathways. Currently, genome annotation and analysis are carried out as a cottage industry, but this will no longer be feasible. Development of a set of tools will necessitate other approaches. Establishment of centers of excellence in genome analysis and/or the development of robust software packages for genome analysis by a commercial organization are two possible examples. Because software and internal parameters change with some regularity, there should be some mechanism for archiving versions of software, and NSF should encourage researchers to publish the version or versions of analysis software used. Additionally, since there are many non-mathematicians using sophisticated analysis software, NSF should take the lead in encouraging/requiring adequate documentation of software so that researchers can understand the algorithms and default parameters to determine the underlying assumptions or limitations in the analysis.

A second goal is to develop a series of community-specific databases (CSDBs) that could be organized based on taxonomic groups or communities of

NSF should take the lead in putting together an international commission to establish standards and recommend measures to ensure compliance.

microorganisms. CSDBs may be geographically distributed, and the workload may be distributed among various groups that have specific expertise in unique areas of research. The advantage of this kind of approach is that it can greatly minimize duplication of effort and duplicate funding of research activities, promote more uniform data quality and representation, and facilitate meaningful linkage between centers involved in these activities. Moreover, it facilitates the adoption of specific standards for genome annotation and curation as well as for database interoperability. Such CSDBs will integrate genomic and experimental information, will undergo continual curation by experts in each respective field, will allow easy access to the stored data by the users, and will be computable so that their use by the scientific community will lead to new insights about biological function.

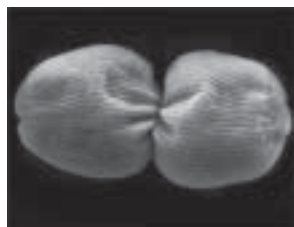
The final goal is to develop a more comprehensive and sophisticated set of standards for genome annotation and comparative analyses and to institute the appropriate sets of controls to ensure better compliance with such standards. This is essential in order to maximize the utility of genome data for end users. There are already too many examples of improper genome annotation in existing databases, and as each new genome project is completed, the quality of the data diminishes due to the continual propagation of errors. Two related issues that must be dealt with as quickly as possible are the need for funding to correct existing errors in genome databases and recognition of the fact that a great deal of work involved in genome annotation and analysis should be viewed as a scholarly activity, rather than just a community service. Progress in this area will likely require the establishment of an international commission to establish standards and recommend measures to ensure compliance. It was suggested that the NSF take the lead in putting this group together perhaps by starting with one or more workshops to better define the problem.

Functional Genomics

The future of functional microbial genomics is absolutely dependent on the development of new, high-throughput, affordable technologies like mass spectrometry and DNA microarray technology that can be widely disseminated to the scientific community. Additionally, for organisms that are genetically tractable, genome-scale reagents such as deletion strains, complete ORF libraries for microarrays as well as suppression studies, and two-hybrid and GFP libraries, need to be made available to researchers in a timely and cost-effective manner. While DNA microarray technology has shown great promise for analysis of gene expression levels for an entire genome, it is not widely available, and the costs for this technology are still too high. Conversely, while 2-dimensional PAGE is widely used, it is not well suited to analysis of proteins on a genome-scale. It is clear that no single technology will suffice for future studies on biological function of genes and proteins, and the integration and robustness of novel technologies is a key component for continued rapid progress.

The data produced in these studies will have to be easy to analyze and visualize and must be interactive with other types of relevant data. Thus, the availability of raw scores as tab-delimited files and careful documentation of normalization methods are critical. As in the case of bioinformatics and database issues, it will be essential to get the input from a cross-section of potential users during the development and planning stages for any new

The Microbe Project: A BIO Advisory Committee Workshop



Entodinium dividing.

Photo courtesy of Mel

Yokoyama and Mario Cobos.

Image reproduced by permission
from *Microbe Zoo*,

© 1997 Michigan State
University.

approaches or technologies. Additionally, because programs are constantly changing and internal parameters may change without notice, some type of archival storage of programs should be carried out so that data collected and analyzed in early studies can be compared with data obtained later. A successful implementation of new functional genomics technologies will require the input from a number of disciplines, and NSF is uniquely poised to take the lead in this arena because it can bring expertise from the biological and physical sciences, engineering, and computer science to bear on overcoming existing technical hurdles.

The development of new technologies for functional genomics in the microbial area can be driven, in large part, by the scientific questions to be addressed. An important short-term goal (1-2 years) is to be able to interrogate every open reading frame (ORF) in every organism in a population under various environmental conditions. Not only is this goal dependent on better technology that would allow for simultaneous measurement of changes in gene expression on a genome level, but it is also dependent on having a catalog of all ORFs for a particular ecosystem, one of the goals that was elaborated under Sequencing and Biodiversity above.

Medium-term goals (3-5 years) include the development of systematic approaches to determine the function of unknown genes. This will require the development of scalable technologies for identifying gene/protein function and may involve combinatorial libraries on chips or beads. Another medium-term goal relates to moving biology from the lab bench to specific environments. Technologies for identification of organisms *in situ* and for assaying gene and protein activity in single cells *in situ* will be required. The concurrent development of integrated, searchable databases and algorithms to enhance integration and analysis of data is essential. These technological breakthroughs will allow accurate studies on how individual organisms, especially unculturable organisms in communities, respond to external stimuli including other organisms in the community. In addition, these increases in technology will allow the organisms themselves to be used as sensors or detectors of their own environments and will permit detailed and, eventually, predictive models of community dynamics and responses.

Longer-term goals (10-20 years) relate to environmental monitoring on a large-scale. It will be important to be able to process a single environmental sample, such as one ml of seawater or one gram of soil, to identify the complement of microbes present, and to monitor the gene and protein content and expression state of this environment. Such advances will enable one to define a baseline activity for an entire population or habitat, natural or engineered, and study the effects on any perturbation on a range of parameters from the activity of a single gene to the post-translational modification of proteins to its effect on lateral gene transfer among species to the dynamics of the entire population of species. The ability to carry out these analyses on evolutionarily-related organisms will allow researchers to identify conserved regulatory networks and components essential for cellular processes such as aging, quiescence, and cell division.

Education and Work Force Issues

An overriding theme throughout the workshop was the shortage of investigators with training in bioinformatics, computational biology, and functional

All too often good science is not funded because the relevance of the research is not understood by the non-specialist.

genomics methodologies. Genomic data has been increasing at an exponential rate, and the instrumentation that has made this possible and will be required for functional genomics studies in the future is becoming ever more sophisticated. A commitment must be made to integrate work in microbiology and microbial ecology with advanced efforts in genomic sciences and to recruit individuals with training in bioinformatics (data storage and databases), computational biology (more complex analyses that may require development of new algorithms or software tools), physics, mathematics and statistics, molecular modeling, and engineering into the biological sciences to meet the challenges for the future. While the successful investigator in the 21st century may likely be an expert in a specific biological discipline, she/he will also need to be able to navigate the ever-expanding databases of genomic data and be familiar with the emerging new technologies for functional genomics studies.

It is essential that new interdisciplinary training opportunities become available as quickly as possible to address the needs of the scientific communities that will be the end-users of this information. Training must be available at all levels: for undergraduate and graduate students, for post-doctoral fellows, and for established investigators. The necessary training will best be accomplished by several mechanisms including development of new courses for undergraduates and graduate students, institutional training grants in genomics, full immersion summer courses, industry internships, and post-doctoral fellowships in established genome centers of excellence.

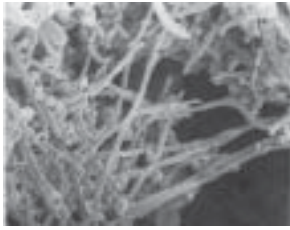
Outreach to Scientific Organizations and the Public

We humans as a species are intimately involved with the microbial species that are co-inhabitants of our planet. Although we often tend to think about only those organisms that cause disease in humans, we are absolutely dependent on the metabolic and biochemical diversity of microbial species for maintenance of the earth's atmosphere, for recycling organic and inorganic waste, for generating nitrogen for growth of plants, for providing a platform for evolution, for producing biomaterials, and for inspiring the development of biomaterials and technologies. In addition, we have deliberately made use of the metabolic capabilities of microbial species in the manufacture of food and antibiotics and in the bioremediation of industrial waste. As the world's population continues to grow this puts an ever-increasing burden on the planet's natural resources.

The application of genomic technology to the study of human pathogens was readily embraced. It was clear that there is tremendous potential to use the new genomic data from infectious species of microbes to accelerate the development of new diagnostics, therapeutics, and vaccines. The promise of genomic science in the area of infectious disease research has already begun to be realized. Novel therapeutic targets and vaccine candidates have been identified for a number of important infectious agents and are moving into clinical trials. The application of genomic technology to the study of microorganisms that exist in numerous environments from deep-sea vents to hot springs to the ocean, the soil, and our own bodies has an equal, if not greater, potential to accelerate the development of novel technologies for agriculture and the environment that will ultimately influence the health of our planet.

Genomics, especially microbial genomics, is a new frontier. One of the major lessons of genomics has been that we have always discovered more than we would have predicted at the outset. Today, we can easily imagine that in-

The Microbe Project: A BIO Advisory Committee Workshop



Methane producing microbes. Photo courtesy of Henry Aldrich.

Increased investment in microbial genomics will lead to a better understanding of evolution and genome stability or the ability to model transcriptional and translational regulation at the level of the cell. We can also predict that the ability to probe genetic responses in an environment or understand the interactions of organisms in a community will lead to novel insights into mechanisms involved in maintaining environmental and community homeostasis as well as those involved in intercellular and interspecies signaling. We can also easily predict that the ability to compare microbial physiology on a genomic level will allow us to understand the range of states of living organisms from cell division to stasis and death, which will give us insight into processes such as the evolution of aging and reproduction. What provides motivation to researchers in genomics is the sense of being on a frontier, of learning new things that could not have been predicted, of asking questions that could not even have been considered previously, and of anticipating that in the near future we will be able to ask questions that today are completely beyond our imaginations.

APPENDIX

Materials sent to participants in advance of the workshop.

The Microbe Project: A BIO Advisory Committee Workshop

Conveners: **Claire M. Fraser**
The Institute for Genomic Research
cmfraser@tigr.org

John Wooley
University of California, San Diego
jwooley@ucsd.edu

Objectives of the Workshop

Microbes were the first organisms on earth and predated animals and plants by more than 3 billion years. They are the foundation of the biosphere — both from an evolutionary and an environmental perspective (1). It has been estimated that microbial species make up about 60% of the Earth's biomass. The genetic, metabolic, and physiological diversity of microbial species is far greater than that found in plants and animals. And yet the diversity of the microbial world is largely unknown, with less than one-half of 1% of the estimated 2-3 billion microbial species identified. Of those species that have been described, their biological diversity is extraordinary, having adapted to grow under extremes of temperature, pH, salt concentration, and oxygen levels.

Perhaps no other area of research has been so energized by the application of genomic technology than the microbial field. It was just five years ago that TIGR published the first complete genome sequence for a free-living organism, *Haemophilus influenzae* (2), and since that first report another twenty-seven microbial genome sequences have been published, with more than 100 other projects underway (3). This progress has represented, on average, one completed genome sequence every two months and all indications point to this pace continuing to accelerate. Included in the first completed microbial projects are many important human pathogens; the simplest known free-living organism; “model” organisms *E. coli* and *B. subtilis*; thermophilic bacterial species that may represent some of the deepest branching members of the bacterial lineage; five representatives of the archaeal domain; and the first eukaryote, *Saccharomyces cerevisiae*.

All of the organisms that have been studied by whole genome analysis are species that can be grown either in the laboratory or in animal cells. It is important to remember that the vast majority of microbial species cannot be cultivated at all, and these organisms, which live in microbial communities, play essential roles in the overall ecology of the planet. Nevertheless, the study of “laboratory-adapted” microbes has had a profound impact on our under-

The Microbe Project: A BIO Advisory Committee Workshop

standing of the biology and the evolutionary relationships among microbial species. Microbiologists are now beginning to exploit this vast amount of new information in both basic and applied areas of research. The payoffs from these efforts will be significant and will promote advances in drug and vaccine design and in industrial and environmental processes.

Given the rapid progress in the field of microbial genomics, it is perhaps not surprising that a number of challenges have emerged and certain areas that deserve attention have been overlooked. These include, for example, the need to expand cooperation and coordination among the governmental agencies that are funding this work; address priorities for future sequencing projects; determine the role of small vs. large sequencing groups in the overall enterprise; define and adopt standards in gene annotation, develop consistent and fair policies governing data release by sequencing groups and its use by the scientific community, establish a long-term investment in databases and software for data mining and manipulation, meet challenges and opportunities in the area of functional genomics, and implement new programs to train the next generation of genome investigators.

Following on from the Interagency Report on the Federal Investment in Microbial Genomics, a new interagency working group will be convened to develop a coordinated interagency effort, now called “The Microbe Project.” NIH, USDA, DOE, and NASA in addition to NSF are involved thus far, and other agencies may join. Each agency’s mission will dictate its primary role in the Microbe Project. The NSF role clearly is basic science related to microbial diversity, including microbes in the environment. It seems likely that in future years some agencies will collaborate and develop joint programs.

Statement of Purpose and a List of Topics

The purpose of the proposed BIO AC workshop will be to provide advice that will help crystallize NSF’s role in The Microbe Project. The workshop will summarize the accomplishments, challenges, and opportunities in the microbial genomics field that are relevant to the NSF; provide some direct advice such as providing criteria for selection of microbes that should be sequenced with NSF support; and identify issues that should be addressed in greater detail including informatics, standards for annotation, and infrastructure needs in future workshops or other venues. The list of topics that could be discussed at this workshop include:

Where are we today in microbial genome sequencing — summary of funded microbial genome sequencing projects.

Where should we go from here – discussion of areas of research not yet funded that could increase the understanding of the microbial world including its biochemical and metabolic diversity; the evolution of microbial species; and the application of this information in basic biology, human and animal health, agriculture, the environment, and biotechnology.

How should priorities for future sequencing projects be determined? What criteria should be used to evaluate the choice of microorganisms for genome sequencing projects? How should genomics of larger organisms be addressed in the future?

What is the role of small vs. medium vs. large sequencing centers in meeting NSF objectives in microbial genomics? What is the role of academic/government/industrial partnerships? Does a virtual genomics facility made up of multiple participants make sense? How should future NSF-funded sequencing projects be coordinated to involve as many interested scientists as possible? How can sequencing infrastructure needs best be addressed?

What is required to further develop the genomic information infrastructure (sequences, databases, software) to enable the largest number of scientists to carry out genome-enabled science? What is the role of large and generalized vs. small and specialty databases? How is funding for these activities coordinated among multiple funding agencies?

What is the current state of the art in techniques related to functional genomics (DNA microarrays, proteomics, computational biology, and structural biology)? What are the challenges and opportunities in the coming years in the area of functional genomics? What is the role of large vs. small genomics facilities in elucidating the relationship between sequence and function on a large-scale? What are the most urgent tools and resources needed to drive genomic-level research? What is the role for NSF in establishing functional genomics centers?

What are the most urgent workforce issues that need to be addressed in order to maximally exploit breakthroughs that will come from genomics?

Participants

The invitees represent a cross-section from the fields of microbiology, environmental microbiology, evolutionary microbiology, microbial genomics, and bioinformatics. Many of the participants have had in-depth experience in the application of genomics to research questions.

BIO AC Members, Co-Conveners

Claire Fraser - The Institute for Genomic Research
John Wooley - University of California, San Diego

Invited Participants

Farooq Azam - Scripps Institute of Oceanography
Colleen Cavanaugh - Harvard University
Penny Chisholm - Massachusetts Institute of Technology
Ed DeLong - Monterey Bay Aquarium Research Institute
W. Ford Doolittle - Dalhousie University
Horst Feldbeck - Scripps Institute of Oceanography
Ken Halanych - Woods Hole Oceanographic Institute
Peter Karp - SRI International
Chad Nusbaum - Massachusetts Institute of Technology
Jim Oliver - University of North Carolina at Charlotte
Gary Olsen - University of Illinois
Mitch Sogin - MBL
Dieter Soll - Yale University
Jim Tiedje - Michigan State University
Maggie Werner-Washburne - University of New Mexico
Owen White - The Institute for Genomic Research

NSF Observers

Rita Colwell – NSF Director

Mary Clutter – Assistant Director for the Biological Sciences

Maryanna Henkart – Division Director, Division of Molecular and Cellular Biosciences

Matthew Kane – Program Director, Systematic Biology

Location of Meeting and Dates

The workshop will be held at the J. Erik Jonsson Center of the National Academy of Sciences in Woods Hole, MA from August 10 (all day) until early afternoon on August 11, 2000. We plan to start the first day with a series of short presentations by selected participants to give a feel for the current state of microbial genomics, technological capabilities, and projected needs. We hope that this will lead into the afternoon's discussions on how a NSF microbial genomics program should be structured and supported. Following lunch on the first day we will break into groups to discuss some of the specific questions outlined above and begin to compile a list of priorities and needs. Day 2 will start with short presentations by the break-out group leaders which should prompt further discussions. If necessary, the groups will reconvene for finalization of their documents. A final session of the second day will be for a wrap-up session. Our hope is to be able to produce a white paper that will summarize how this group envisions the present and future needs of the scientific communities in the broad field of microbial genomics and how these needs can best be met by NSF alone and in partnership with other agencies. The recommendations made in this document, we hope, will also serve as a platform in support of additional workshops and symposia related to specific issues.

References

Staley, J. J., Castenholz, R. W., Colwell, R. R., Holt, J. G., Kane, M. D.,

Pace, N. R. et al. *The Microbial World. The foundation of the biosphere.* American Society for Microbiology, 1997.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science* 269: 496-512 (1995).

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

Schedule

August 10, 2000

8:30 – 9:00	Informal breakfast at the Jonsson Center
9:00 – 12:30	Presentations by participants
9:00	Welcome – Rita Colwell, Director, NSF
9:10	Introduction and objectives of workshop Mary Clutter/Claire Fraser/John Wooley
	Microbial genome sequencing projects: state of the field and where do we go from here?
9:20	Ed DeLong (MBARI)
9:40	Ford Doolittle (University of Dalhousie)
10:00	Farooq Azam (SIO)
10:20	Mitch Sogin (MBL)
10:40	Break
	Genome databases: state of the field and where do we go from here?
11:00	Owen White (TIGR)
11:20	Peter Karp (SRI)
	Functional genomics: state of the field and where do we go from here?
11:40	Maggie Werner-Washburne (University of New Mexico)
12:00	TBD
	Summary of report of Marine Microbial Genomics Workshop held in April 2000
12:20	Penny Chisholm (MIT)
12:40 – 1:40	Lunch
1:40 – 2:00	Discussion of workshop objectives/Charge to participants
2:00 – 5:00	Break-out group discussions (with break at 3:30 p.m.)
5:00 – 6:00	Reception at the Jonsson Center
6:00 – 8:00	Dinner at the Jonsson Center

The Microbe Project: A BIO Advisory Committee Workshop

August 11, 2000

8:30 – 9:00	Informal breakfast at the Jonsson Center
9:00 – 10:30	Presentations by break-out groups (30 minutes each)
10:30 – 12:00	Break-out groups reconvene for final write-ups
12:00 – 1:00	Lunch
1:00 – 2:00	Final meeting of all participants and review of recommendations

The intention of the break-out groups is to address a limited number of specific topics in the areas of (1) genome sequencing, (2) bioinformatics and infrastructure, and (3) functional genomics. During the first afternoon, each break-out group will have the charge to discuss their area, address the questions, and produce a synopsis of recommendations for presentation to the entire group the next morning.

Following the presentations on day two, the break-out groups will reassemble and, using the revised list of recommendations, prepare a document that will be included into a white paper for presentation to BIO.

BI

**The National Science Foundation
Directorate for Biological Sciences
4201 Wilson Boulevard
Arlington, VA 22230**

