Background and Challenges

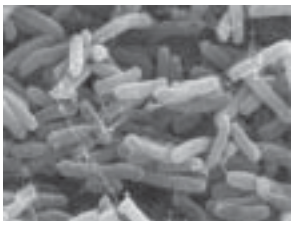
Workshop Charge:

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.

Sequencing and Biodiversity

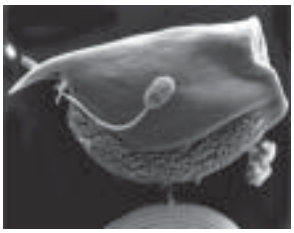
Genomics is the latest and most exciting tool aimed at unraveling the incredible complexity of microbial life on earth. The field of microbial genomics has made enormous progress during the past five years since the first report on the completion of the genome sequence from a free-living organism in 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.



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

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



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 microstructure of microbial communities in the sea and in the soil, and genomics is the key to helping us interrogate the systems.

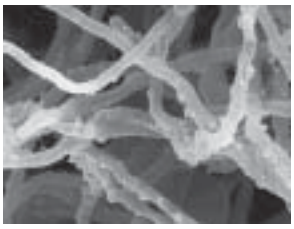
Analysis of complete genome sequences is beginning to provide a great deal of insight into many questions concerning the evolution of microbes. One area of insight has surrounded the occurrence of genetic exchanges between different evolutionary lineages – a phenomenon known as horizontal, or lateral, 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 Entamoebe. 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.

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.

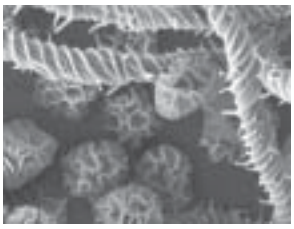
Current challenges include the lack of standards among groups for genome annotation; the lack of uniformity in data presentation, accessibility, output capabilities, and archiving of software versions; and relatively poor software documentation.

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.



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.

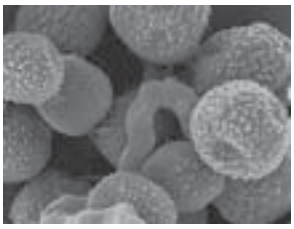
Recommendations

"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.

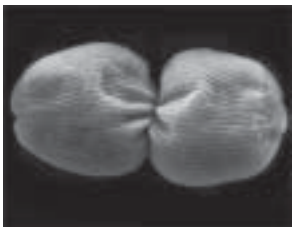
microorganisms. CSDBs may be geographically distributed, and the workload may be distributed among various groups that have specific expertise in unique areas of research. The advantage of this kind of approach is that it can greatly minimize duplication of effort and duplicate funding of research activities, 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



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.

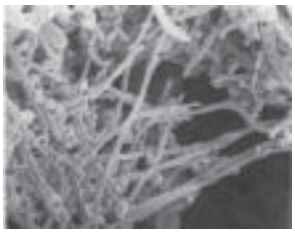
genomics methodologies. Genomic data has been increasing at an exponential rate, and the instrumentation that has made this possible and will be required for functional genomics studies in the future is becoming ever more sophisticated. A commitment must be made to integrate work in microbiology and 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-



Methane producing microbes. *Photo courtesy of Henry Aldrich.*

creased 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.