AMERICAN ACADEMY OF MICROBIOLOGY

MICROBIAL GENOMES:

BLUEPRINTS FOR LIFE

This paper was written with support of the U.S. Department of Energy under Contract No. DE-FG02-99ER62738.

DOE Patent Clearance Granted
MPDvorscak May 24, 2001
Mark P. Dvorscak Date
(630) 252-2393
E-mail: mark.dvorscak@ch.doe.gov
Office of Intellectual Property Law
DOE Chicago Operations Office

1752 N STREET, N.W.

WASHINGTON, DC 20036

WWW.ASMUSA.ORG/ACASRC/ACA1.HTM

This report is based on an American Academy of Microbiology colloquium held March 19-21, 1999, in New Orleans, Louisiana.

The colloquium was supported by the following sponsors:

U.S. Department of Agriculture

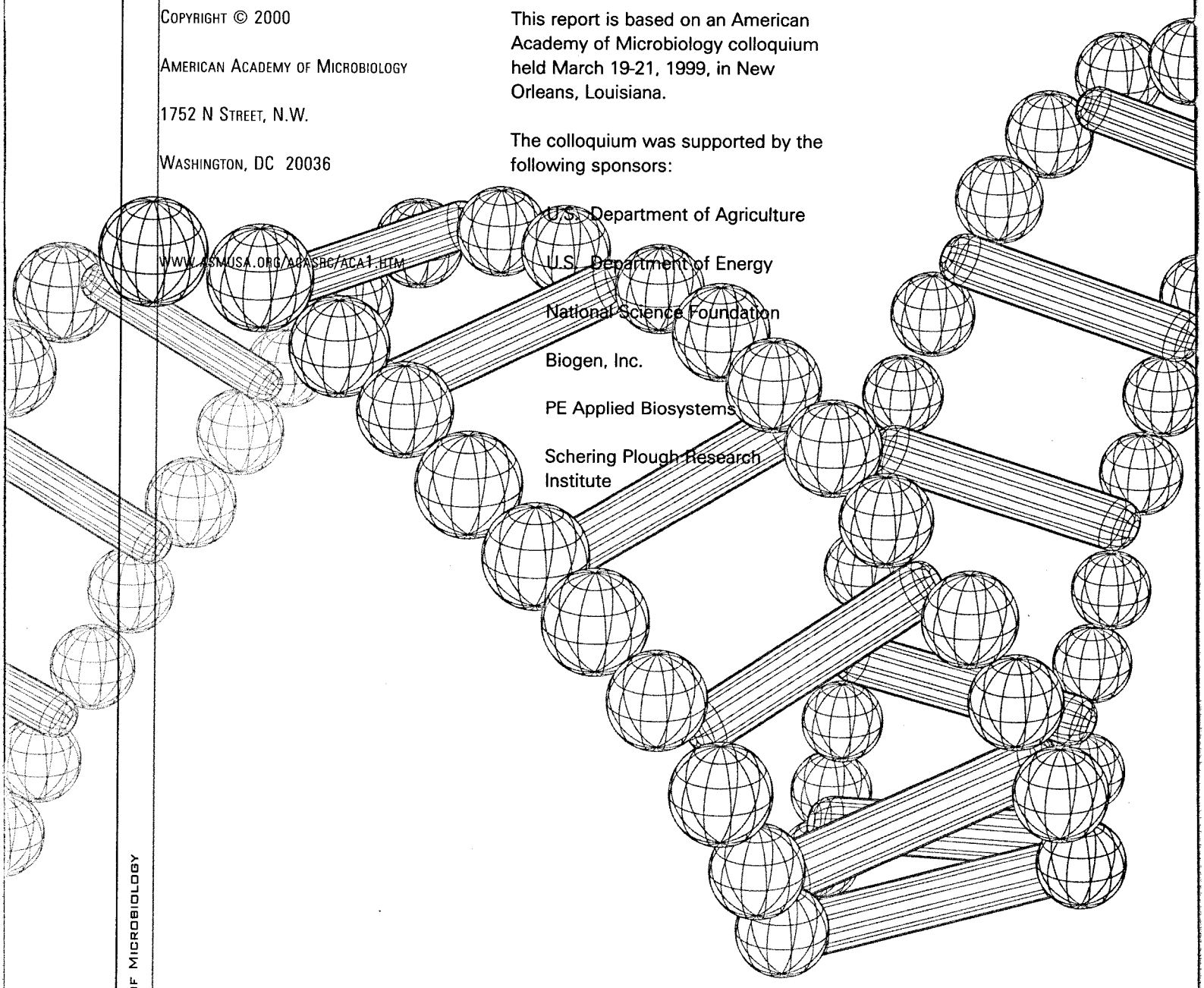
U.S. Department of Energy

National Science Foundation

Biogen, Inc.

PE Applied Biosystems

Schering Plough Research Institute

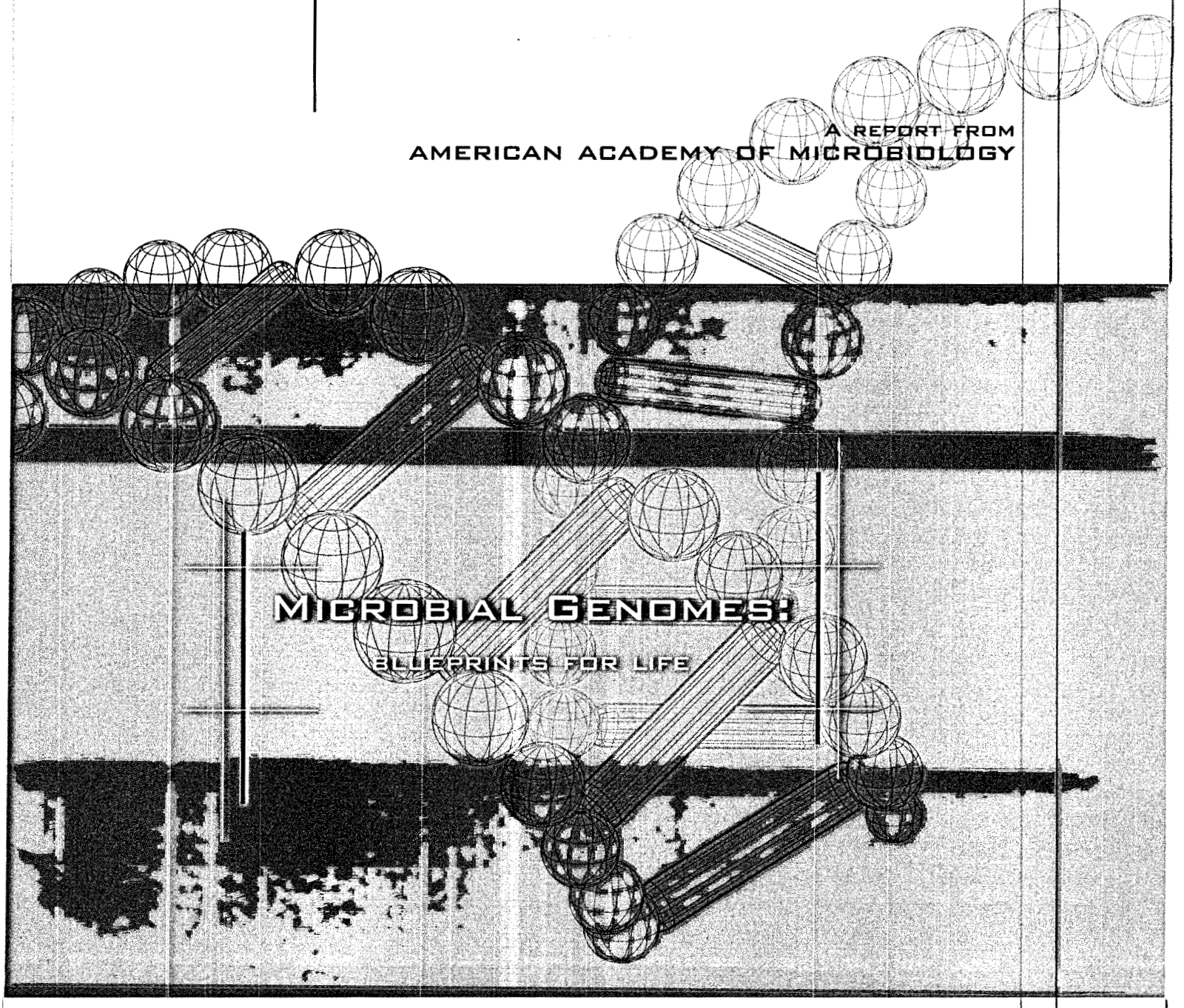


DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

A REPORT FROM
AMERICAN ACADEMY OF MICROBIOLOGY



MICROBIAL GENOMES:
BLUEPRINTS FOR LIFE

DAVID A. RELMAN, M.D. AND EVELYN STRAUSS, PH.D.

AMERICAN ACADEMY OF MICROBIOLOGY

MICROBIAL GENOMES: BLUEPRINTS FOR LIFE

Board of Governors

Colloquium Steering Committee

David A. Relman, M.D. (Chair)
Stanford University

David Schlessinger, Ph.D.
National Institute of Aging, Baltimore, MD

Hamilton Smith, Ph.D.
Celera Genomics, Rockville, MD

Mitchell Sogin, Ph.D.
Marine Biological Laboratories, Woods Hole, MA

J. Craig Venter, Ph.D.
Celera Genomics, Rockville, MD

Colloquium Participants

Joan W. Bennett, Ph.D.
Tulane University

Alison M. Berry, Ph.D.
National Science Foundation, Arlington, VA

Donald A. Bryant, Ph.D.
Pennsylvania State University

Allan M. Campbell, Ph.D.
Stanford University

Rita R. Colwell, Ph.D., Sc.D.
University of Maryland

Shiladitya DasSarma, Ph.D.
University of Massachusetts, Amherst

Julian E. Davies, Ph.D.
TerraGen Diversity, Inc., Vancouver, BC, Canada

Robert E. Davis, Ph.D.
U.S. Department of Agriculture, Beltsville, MD

W. Ford Doolittle, Ph.D.
Dalhousie University, Halifax, NS, Canada

Stanley Falkow, Ph.D.
Stanford University

Michael Fonstein, Ph.D.
University of Chicago

Claire M. Fraser, Ph.D.
The Institute for Genomic Research, Rockville, MD

Marvin E. Frazier, Ph.D.
U.S. Department of Energy, Germantown, MD

Thomas R. Gingeras, Ph.D.
Affymetrix, Inc., Santa Clara, CA

Harold S. Ginsberg, M.D.
National Institutes of Health, Bethesda, MD

Michael Gottlieb, Ph.D.
National Institutes of Health, Bethesda, MD

D. Jay Grimes, Ph.D.
The University of Southern Mississippi

Radhey S. Gupta, Ph.D.
McMaster University, Hamilton, Ont., Canada

Maryanna P. Henkart, Ph.D.
National Science Foundation, Arlington, VA

Richard E. Isaacson, Ph.D.
University of Illinois

H. Mark Johnston, Ph.D.
Washington University School of Medicine, St. Louis, MO

A. Dale Kaiser, Ph.D.
Stanford University

Noel T. Keen, Ph.D.
University of California, Riverside

Jeffrey H. Miller, Ph.D.
University of California, Los Angeles

Frank Christopher Minion, Ph.D.
Iowa State University

Richard Moxon, Ph.D.
John Radcliffe Hospital, Oxford, England

Kenneth H. Nealson, Ph.D.
California Institute of Technology

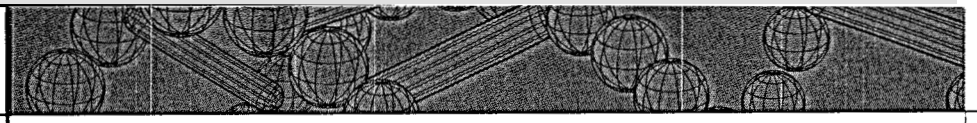
Eugene W. Nester, Ph.D.
University of Washington

David A. Relman, M.D.
Stanford University

Monica M. Riley, Ph.D.
Marine Biological Laboratories, Woods Hole, MA

R. Michael Roberts, Ph.D.
U.S. Department of Agriculture, Washington, DC

Richard J. Roberts, Ph.D.
New England Biolabs, Beverly, MA



David Schlessinger, Ph.D.
National Institute of Aging, Baltimore, MD

Karen Shaw, Ph.D.
Schering Plough Research Institute, Kenilworth, NJ

Melvin I. Simon, Ph.D.
California Institute of Technology

Mitchell L. Sogin, Ph.D.
Marine Biological Laboratories, Woods Hole, MA

James T. Staley, Ph.D.
University of Washington

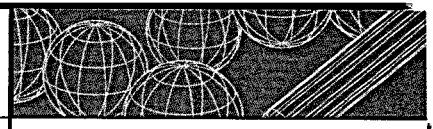
Karl O. Stetter, Ph.D.
University of Regensburg, Regensburg, Germany

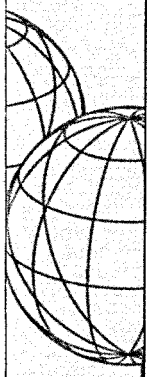
STAFF

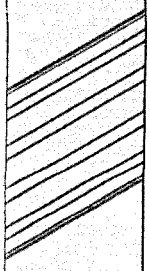
Carol A. Colgan, Director
American Academy of Microbiology

Peggy McNult, Manager
American College of Microbiology

Report Editor
Evelyn Strauss, Ph.D.
Santa Cruz, CA







Executive Summary

A colloquium was convened by the American Academy of Microbiology to consider issues relating to the blossoming field of microbial genomic science. The colloquium was held in New Orleans, Louisiana, on March 19-21, 1999. The principal findings of the colloquium are summarized below.

Microbes carry enormous genetic wealth and biological aptitude. Humans have already exploited these resources to enhance the quality of our lives in many ways. We have adopted the biochemical competence and versatility of these tiny creatures in medicine, agriculture, ecology, and studies of evolution.

The advent of DNA sequencing on a large scale brings with it the increasing ability to unravel and compile the genetic secrets of any living thing. Microbiologists are currently in a position to delve into these small organisms and to emerge with valuable information that can be put to both practical and theoretical use. This information will increase our understanding of how microbes contribute to health and disease of our bodies and of natural environments. It will also promote advances in food production, bioremediation, and drug design. It will extend our insight into how organisms weave the complex web that sustains them—and indirectly ourselves—and will allow us to reconstruct the origin of life and understand how it continues to evolve.

Investigators have already begun to tap into this tremendous store of knowledge. As the field of microbial genomic science has begun to grow, however, various problems have become apparent. Many of these are typical of a rapidly expanding enterprise; others are specific to this new scientific arena. The unique benefits of genomics can point to the importance of these problems and the need for coordinated attention. The colloquium participants discussed how to address these challenges so as to foster most effectively the expansion and vigor of this novel and invaluable endeavor.

Genomic science has the power to revolutionize microbiology, which in turn will transform and improve life on this planet. In order to realize this potential, strategic interventions are required now.

THE ADVENT OF DNA SEQUENCING ON A LARGE SCALE BRINGS WITH IT THE INCREASING ABILITY TO UNRAVEL AND COMPILE THE GENETIC SECRETS OF ANY LIVING THING. MICROBIOLOGISTS ARE CURRENTLY IN A POSITION TO DELVE INTO THESE SMALL ORGANISMS AND TO EMERGE WITH VALUABLE INFORMATION THAT CAN BE PUT TO BOTH PRACTICAL AND THEORETICAL USE.

INTRODUCTION

MICROBES LIVE EVERYWHERE, INCLUDING IN DEEP-SEA VENTS MILES UNDER THE OCEAN, BOILING HOT SPRINGS, GLACIERS, AND EVEN ROCK, AS WELL AS MORE FAMILIAR SETTINGS SUCH AS SOIL, PONDS, AND ANIMAL INTESTINES. THEIR ABILITY TO SURVIVE IN DIVERSE ENVIRONMENTS REFLECTS AN ASTOUNDING ARRAY OF BIOCHEMICAL APTITUDE.

The small size of microbes belies their powerful influence. They surround and inhabit us, and every creature on earth depends on them for life. In order to understand and exploit microbes, human beings must inventory and understand the vast repertoire of microbial activities. Because a creature's capabilities are reflected in and dictated by the DNA it carries, a microorganism's complete—or genomic—DNA sequence provides a blueprint for its biochemical and behavioral endowment.

Microbes live everywhere, including in deep-sea vents miles under the ocean, boiling hot springs, glaciers, and even rock, as well as more familiar settings such as soil, ponds, and animal intestines. Their ability to survive in diverse environments reflects an astounding array of biochemical aptitude. They churn out the nitrogen that plants need for growth and emit gasses that create and maintain the critical composition of the earth's atmosphere. In our bodies, they help digest food, ward off attempts by harmful organisms to invade, provoke misdirected immune responses against our own tissues, and even shape the development of our immune systems.

Humans have co-opted microbial talents to enhance the quality of our lives. For example, many large-scale industrial processes depend on microbes. In addition, we exploit them for making food and medicines and put them to work cleaning up our sewage and industrial waste. Microbes can dramatically influence ecosystems because many possess unique abilities to recycle both organic and inorganic substances. As the world's human population increases and natural resources diminish, environmental challenges grow. Humans are trying to harness the vast metabolic potential of microorganisms to apply microbial solutions to some of these ecological problems.

Microbes are not only vital to our bodies, forests, factories, and all of life. They also represent a record of evolution. Because all life descended from tiny, single-celled creatures, microbes open a window into antiquity. By peering inside these simple organisms, we learn about our origins and those of every living thing. A comparison of new genome sequences with those previously known shows us aspects of biology unique to an organism, as well as features that are shared with others. Elucidating the patterns of particular genes (and the capabilities they confer) among microorganisms improves our understanding of how different creatures are related. This information is critical for attempts to study, capitalize on, and interfere with the activities of microorganisms. Discovering which microbes contain which genes reveals the capabilities demanded by particular types of environments or lifestyles, for example, and points to the deviations that allow organisms to adapt to unusual habitats or behaviors. It also provides the basis for predictions about what capacities a new microorganism might possess, based on the activities of its genetic neighbors.

Despite their importance, we know strikingly little about microbes. They compose greater than 50% of the

INTRODUCTION

living matter on this planet, yet we have identified far less than one percent of all the predicted microbial groups. The microbial world, therefore, represents a vast reservoir of untapped knowledge and potential utility. Microbes offer a view into the capabilities of living things that far exceeds the capabilities that will be revealed by, for example, the human genome sequence alone. The amount of DNA carried by a member of the human species is equivalent to the DNA cache of approximately 1,000 microbes, and there are far more than 10 times this number of different microbes in the world around us. Microbial genome sequencing, therefore, offers an extremely efficient way to generate a picture of the tremendous biological diversity that living creatures possess, particularly because microbes inhabit such an array of bizarre and extreme environments. To profit from this natural biological archive, we need to embark upon a coordinated, large-scale effort to uncover and interpret a comprehensive set of microbial genome sequences.

Already governmental, academic, and private agencies are deciphering the genetic code of many microbes. Since the debut of whole genome sequencing in the mid-1990s, the genomes of 29 microbes and three chromosomes from two lower eukaryotic parasites have been sequenced; more than 100 other sequencing projects are under way. Yet, inevitably for an enterprise that is advancing rapidly and generating an explosion of valuable information, problems have arisen. Organization and dissemination of technology and data have been haphazard. The absence of a broad discussion regarding priorities and goals has contributed to the lack of coordination and absence of generally accepted standards. Sequencing technology has matured at an unexpectedly rapid pace, making quite feasible goals that would not have been imaginable just a few years ago. But, funding is inadequate for the kind of large-scale, integrated effort that should be undertaken.

This colloquium was organized with the aim of directing attention toward the challenges that face this new and promising field in the hope of productively interceding while the enterprise is still young and malleable. Now is the time to conduct discussions that include people with a broad range of interests and expertise about current and foreseeable challenges. Already, federal agencies have begun to ask for guidance on these issues. It is essential that the sequencing efforts accelerate as rapidly and as smoothly as possible. With timely, well-aimed interventions, it should be possible to transform the vision of a large-scale microbial genome sequencing effort into reality, and to maximize the effort's efficiency and economy. The time is ripe for fruitful intervention.

NOW IS THE TIME TO CONDUCT DISCUSSIONS THAT INCLUDE PEOPLE WITH A BROAD RANGE OF INTERESTS AND EXPERTISE ABOUT CURRENT AND FORESEEABLE CHALLENGES.

Potential value of complete microbial genome sequences: vision for the future

Because the average length of a microbial gene is 1000 nucleotides, an amount of sequence equivalent to the size of the human genome (about 3.5 gigabases) carries about 3.5 million gene-sized pieces and should therefore contain at least 350,000 unique genes.

Microbial genome sequences wield enormous power in their potential to touch and shape a tremendous range of human activities. Information and products gleaned from the sequences and associated studies will profoundly impact our health, environment, agriculture, and industry in terms of enhancing both quality and economics in these far-reaching areas.

Microbes contain an impressive repertoire of biochemical activities. This diversity far exceeds that of higher organisms, such as insects, plants, and animals. At least 10 percent (and usually closer to 30-40 percent) of the genes in every complete microbial genome sequence analyzed so far are "new"; they have not been previously discovered in another organism. Because the average length of a microbial gene is 1000 nucleotides, an amount of sequence equivalent to the size of the human genome (about 3.5 gigabases) carries about 3.5 million gene-sized pieces and should therefore contain at least 350,000 unique genes. Furthermore, each microbial genome is small—about one thousandth the size of a mammalian genome. Because microbes occupy myriad environmental niches and a single human organism occupies one, the same amount of sequence from these tiny creatures represents a much wider array of function. We are surrounded by a deep pool of biochemically unique abilities. We have only begun to dip in.

Many molecules produced by living creatures can perform sophisticated feats. Natural selection over evolutionary time scales has honed the ability of biological molecules to carry out chemical deeds with extraordinary precision and specificity. Enzymatic reactions excel in their efficient use of raw materials, and in their production rates. In contrast to many human-made catalysts, enzymes can distinguish subtle differences in their substrates. This allows them to, for example, pick one of several identical chemical groups in a single molecule and add something to it or transform it into something else. Biological molecules can even discern mirror image molecules–a feat that is critical for designing and creating, for example, drugs that need to fit properly into the "glove" of some natural enzyme in the body. Furthermore, reactions performed by biological molecules far outnumber those performed by synthetic ones. Scientists have uncovered many natural reactions that have not been mimicked by laboratory methods.

Enzymes and the microbes that carry them exhibit useful behaviors. They naturally make methane and other sources of energy, fabricate useful chemicals such as antibiotics, and break down toxic substances to harmless products. Food engineers have added microbes (or, in some cases, their enzymes) to production vats in order to manufacture the high fructose corn syrup that sweetens our drinks. Cheese-makers use microbial rennin as a less expensive substitute for the calf version of the same enzyme. Microbes and their products permeate the baking, dairy, and alcoholic beverage industries. Furthermore, certain microbes thrive under harsh conditions, such as extremely high temperatures and pressures. These organisms contain

POTENTIAL VALUE OF COMPLETE MICROBIAL GENOME SEQUENCES: VISION FOR THE FUTURE

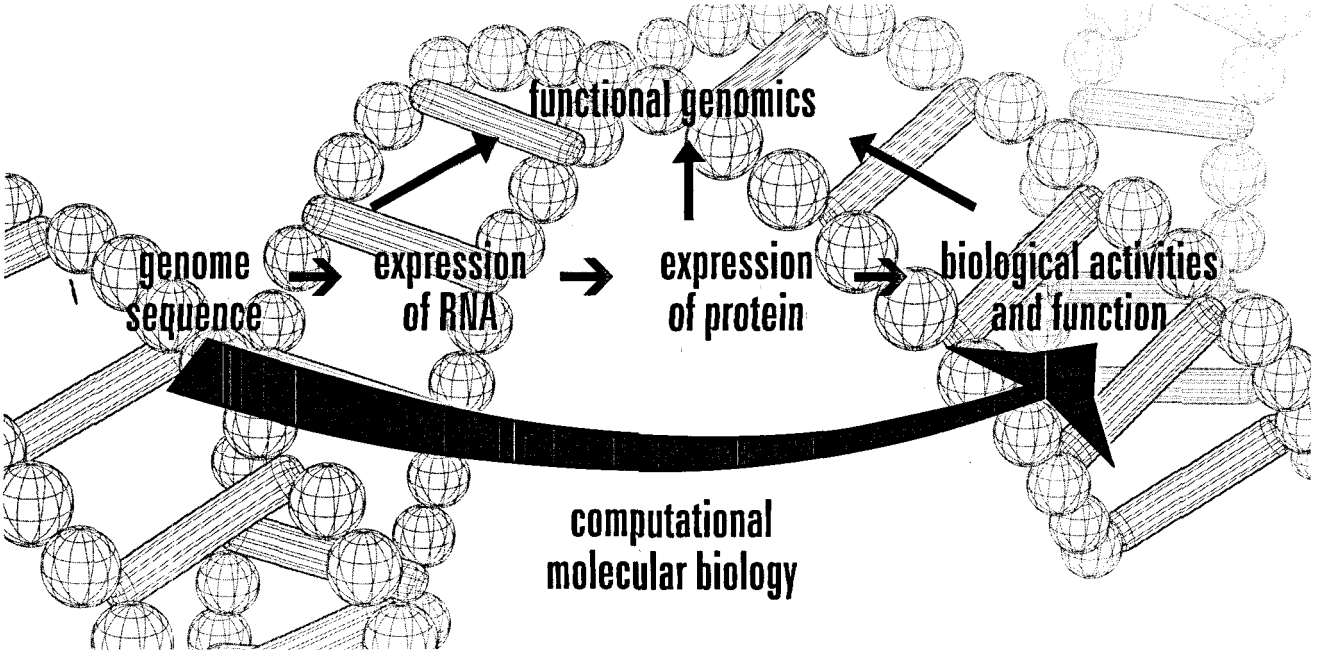
hardy enzymes that can be used to carry out reactions that are currently beyond the reach of synthetic organic chemistry; some of these robust enzymes have already been applied in industry. Through genetic engineering, it is possible to adapt and fine-tune the already exquisitely sensitive and powerful microbial reactions to better serve human purposes.

Microbes also provide valuable information about evolution. Studying them has already revealed previously unknown mechanisms by which genes are "shared" among different branches of the tree of life, which are beginning to redefine the "topology" of this tree. Further investigations will undoubtedly reveal more secrets about how all creatures arose and developed.

Microbes, therefore, represent a veritable treasure chest of biochemical tricks and historical facts, most of which currently remain buried. Sequencing provides a tool with which to begin uncovering this mass of information (see Figure). A very small number of microbes have been discovered and even fewer of those have been, or are being, sequenced. Furthermore, because the function of up to 40 percent of microbial genes on average is unknown, scientists have not yet unraveled all the biochemical secrets of even one microbe.

Sequencing the genome of an organism entails spelling out the letters—the nucleotides, or bases—that compose the entire chromosome (or, in some cases, chromosomes). The sequence of these nucleotides allows scientists to predict where genes start and end, and the order in which amino acids are strung together to make the proteins that these genes encode. In many instances, the resulting pattern of amino acids resembles that of a protein with a known function. In such cases, scientists have a clue as to the new gene's—and its protein product's—role in the cell or its biochemical potential. Even if a gene does not resemble one with a known physiological job, investigators can still determine evolutionary relationships between organisms by discerning families of related genes.

In addition to these conventional uses of genetic information, sequencing can provide a doorway to creatures that are difficult to grow in the laboratory. Because DNA can be obtained from microbes whether or not they're



Potential value of complete microbial genome sequences: vision for the future

The field of microbial genomics will give rise to large amounts of information about microbes' biochemical talents. Our understanding of complex microbial communities in diverse environments will advance.

viable, and whether or not we know how to propagate them in the laboratory, this method provides a powerful way to determine an organism's presence and genetic relatedness to other organisms. There is no demand for investigators to first figure out how to keep it healthy in a test tube instead of in its usual habitat miles under the ocean or some other odd and difficult-to-mimic environment. In some particularly productive circumstances, results from a sequence analysis might suggest ways to grow a previously uncultured microbe in the lab, by revealing relationships to organisms for which scientists have already solved this problem.

The past five years have seen the complete sequencing of about 30 organisms, most of them microbes. Within the next decade, the genetic blueprints of hundreds more organisms—including humans—likely will be fully spelled out. As the sequencing enterprise picks up steam, however, greater coordination for projects will be essential and challenges of appropriate storage, analysis, and dissemination of the data they produce will need to be met.

In the next five years, the rate of sequencing will accelerate; a simple bacterial genome sequence may be completed in a day! Colloquium participants hope that scientists will be able to enter and manipulate data easily within widely accessible databases that include up-to-date information about sequence and associated biology from all over the world. Computer tools will help investigators analyze their data rapidly; researchers will be able to ask sophisticated questions and obtain reasonable answers. This will allow them to proceed quickly from new sequences to testable hypotheses that will reveal biological function. Numerous affordable technologies will exist for analyzing gene expression and protein action on a genome-sized scale. As a consequence, the amounts and types of data about gene function will explode.

The field of microbial genomics will give rise to large amounts of information about microbes' biochemical talents. Our understanding of complex microbial communities in diverse environments will advance. We will understand much more about the role of microbes in keeping our bodies and surroundings healthy as well as in making us sick. This amassing knowledge will fuel a significant increase in the number and types of commercial products in the fields of agriculture, medicine, industry, and environmental cleanup.

CURRENT STATUS

As of July 2000, scientists have published genome sequences of 29 microorganisms; over 100 more are in progress. So far, the bulk of sequencing has been performed by a few, dominant centers. These include governmental, as well as private, agencies. Sequencing technology is advancing rapidly. Methods for generating sequences have become faster and more efficient. As a result, costs per nucleotide of sequence have decreased. They still, however, prohibit most smaller groups, such as individual academic laboratories, from tackling entire genome sequences. Eventually, individual investigators will likely enjoy the option of sequencing an entire microbial genome by themselves, in a reasonable period of time; this would greatly expand the breadth and depth of the gene and gene product discovery effort. However, at the present time, most investigators cannot consider this a realistic possibility.

Coverage of the microbial world has been spotty. In particular, genome projects have focused on organisms that cause human disease. Microbes of agricultural importance or those that associate harmlessly with animals (commensals) have not received nearly as much attention. Similarly, organisms that inhabit the nether regions of our planet, and which therefore represent the great diversity in the microbial world, have largely been ignored.

Much of the information about the function of genes has not been demonstrated directly; rather, it has been inferred from the sequence. Genes that share a large amount of sequence identity or similarity encode proteins that commonly perform the same or related functions. As a result, comparison of a sequence to a gene with a known function can provide clues as to the new gene's activity. However, this method has its limitations. Nature can use similar amino acid sequences (represented in genes by similar nucleotide sequences) for different biochemical activities, and relying on sequence information alone to reveal a gene's properties can lead researchers astray. Furthermore, cells employ similar biochemical reactions in different ways. For example, an investigator might infer from a particular sequence that the gene product in question adds a phosphate group to another protein. However, the target protein and the biological effect of that phosphate will remain a mystery until experiments are performed that directly probe the function of the unknown gene.

Scientists have a long tradition of unearthing what a gene product does by direct experimental inquiry. Decades-old techniques, as well as methods that exploit tools that have been developed in the last few years, can reveal a gene product's function. Conventional procedures and modern, whole genome-level approaches are converging in some new technological schemes. For example, several powerful new strategies represent a convergence of conventional genetic techniques and rationale with a modern, high-throughput twist.

Despite the fact that genomic investigations hold the potential to transform microbiology, the sequenced genomes have so far impacted the field less than expected. In part, this is because often the sequencing community has failed to plan for how genomic data would be integrated into the experimental biology; as a result, the necessary links between gene sequences and function are lagging. The microbial sequencing endeavor currently lacks a mechanism for establishing a broad discussion of priorities about all aspects of the effort. Which microbes should be sequenced? How should scientists ensure that there is good coverage of important microbes from a wide variety of fields? Which technologies and approaches should be put in place to use the genomic information being generated, in an optimal fashion? What kinds of innovations are most needed to generate and interpret the data? How should quality control standards be developed? When during a project should sequence be released? How will funds be generated to support this ambitious, yet tractable and valuable, endeavor? All of these questions and many more are critical to the success and productivity of microbial sequencing projects, yet they have not been considered in a well-orchestrated way.

Moreover, current sequencing efforts are not coordinated. Groups are duplicating each other's efforts by sequencing the same organisms. Once completed, the raw data from publicly funded projects are available, but not always from commercial projects. However, additional information—for example, results investigators might have about the physiological role of particular sequences—is not entered in any standardized format. For this and other reasons, it

CURRENT STATUS

often remains hidden to the scientific community at large. Furthermore, dissemination of information about new technology has not been optimal, in part because no centralized entity is charged with the responsibility for this task.

Better database management tools would greatly enhance efforts to make optimal use of the data that are pouring in from the various sequencing projects, but there has been little organized discussion of the specifications and development of such tools. The field of bioinformatics—where biology and computer science intersect—is booming. Funding organizations are beginning to support this new enterprise and those conducting it, but it will require more concerted support to thrive. Financial support for improved informatics tools and for sequencing *per se* is not currently sufficient for the kind of large-scale efforts that should be undertaken.

Multiple funding sources exist for various microbial sequencing projects, and each has its own perspectives and priorities. These include the Department of Defense (DOD), Department of Energy (DOE), National Aeronautics and Space Administration (NASA), National Institutes of Health (NIH), National Institutes of Standards and Technology (NIST), National Science Foundation (NSF), and the Department of Agriculture (USDA). Together, the different agencies cover a great deal of intellectual and practical territory. However, their combined potential could be strengthened by coordination between them in terms of long-term management or thinking. The enterprise could become more efficient and powerful than it is now.

Challenges, Opportunities, and Potential Solutions

Coordination of microbial sequencing projects

In order to realize the enormous potential of microbial genomic science, all aspects of the endeavor must be better coordinated. Sequencing is currently being performed through a variety of avenues—federal, private, academic, business—and by different countries. Currently, no agency or mechanism is overseeing and organizing these sequencing efforts, or ensuring that they receive the funding and other support necessary for them to continue. This situation has led to duplication of efforts.

The absence of an orchestrated and inclusive discussion among interested parties has led to disorganization in other ways as well. There is no universally agreed upon set of priorities or objectives to guide sequencing activities as they gain ever-greater momentum. Members of the community have only informally, if at all, conferred about the growing need for more powerful tools for analyzing sequence data and for uncovering the biological function of the DNA being sequenced. A discussion of approaches to further drive down sequencing costs would also benefit the community. There is no standard method to disseminate information about such new techniques and devices as they are developed. In some cases, sequencing projects are divorced from the biologists who will use the information, and there has been little formal discussion of criteria to use, or the most appropriate research groups to be involved, when choosing an organism to study.

Standards are not in place for appropriate use and attribution of data posted on the Internet, nor have members of the community agreed about basic issues of quality control with regard to the primary sequence data. There is not even a widely accepted notion of what constitutes a "complete" genome sequence or of the point at which investigators should make the data publicly available. No central repository currently exists for organisms (or DNA) that have been sequenced.

Numerous tactics could combat some of these problems. A federally funded web site, for example, could allow investigators to inform the community about their sequencing projects. This would presumably reduce duplication of efforts. The most widely used website where sequencing projects are now posted (www.tigr.org) lists only sequencing projects that are well under way. By that time, it is too late to gather input from scientists to make sure, for example, that the most appropriate strain is the one being sequenced. Early involvement of eventual data users will strengthen the integrity and efficiency of the entire project. Genomic data will be of most use when it leads to experiments.

Colloquium participants acknowledged that private firms might not list their projects on a public site, and even if they did, their usefulness would remain dubious unless they shared the data. Even with such inadequacies, however, such a list would increase efficiency. As sequencing activities gain momentum, the potential for waste increases tremendously. Coordination will decrease duplication of efforts. While some coordination efforts are underway, much more is needed.

Agencies involved in sequencing projects would benefit from communicating with each other in other ways as well. A discussion about priorities and objectives to guide sequencing activities would serve everyone well, even if specific goals differ between segments of the community. Agencies could profit from each other's experience, build collaborations, and construct an even more powerful and efficient sequencing enterprise together. Furthermore, an alliance of interested parties could encourage other organizations to develop genomics programs; involvement of a variety of associations will ensure the balanced mix of projects required to fulfill the potential of microbial genome projects. Integration of multiple agencies at the international, as well as the national, level is critical. Coordination is needed for all aspects of the microbial genome enterprise to safeguard its ability to thrive and produce.

Sequencing: selection of *microorganisms and strategies*

Strategies: What to sequence: Scientists face several challenges in choosing which microorganisms to sequence and in organizing the sequencing efforts. Every step toward accomplishing these tasks presents multiple options. For example, choices abound about which microbes to sequence, given that they play critical roles in many different fields, including health, agriculture, industry, environmental clean-

Challenges, Opportunities, and Potential Solutions

up, ecology, and evolution. Furthermore, many microbes benefit animals and humans while living inside them; these and others may yield new information about the interrelationships of life. Large-scale microbial sequence information will enrich all of these areas. So far, scientists have weighted their efforts toward disease-causing microbes. This area merits focus, but so do the others. Real and potential roles of microbes in agriculture and bioremediation, for example, need to receive as much emphasis as roles in causing disease, and efforts to sequence nonpathogenic as well as pathogenic organisms should be bolstered, since only a small percentage of microorganisms cause disease.

Colloquium participants formulated general criteria for the choice of genomics projects, but did not endorse any specific list of microorganisms to be targeted. Genomics as a science has many legitimate objectives, which different genome projects and subject areas can fulfill differently. Choosing which sets of organisms to sequence, too, raises issues that will be resolved differently depending on a project's goal. Comparing microbes that are almost identical offers the possibility of pinning down particular genetic sequences that determine specific traits and behaviors. Analysis of close relatives can reveal genes that, for example, allow one organism to disseminate throughout an animal host while its cousin remains confined to a particular site. Studying a wide range of organisms also offers benefits, such as expanding our knowledge of the tremendous breadth of environmental niches and associated biochemical abilities represented by microbes and their biochemical repertoire. For these reasons, sequencing closely, as well as distantly, related organisms provides value.

In the near future, scientists should consider numerous criteria when selecting organisms to sequence. Depending on the project, different questions will be more or less relevant. Some examples are:

- Is DNA available? Can it be easily sequenced?
- Is the organism genetically tractable? Can it be experimentally manipulated?
- What is the phylogenetic significance of the organism? (Is it located within a poorly–studied branch of the tree of life?)
- Does it have environmental and/or ecological significance? (Could it be employed in bioremediation? Does it play a role in cycles of critical chemical elements?)
- Does it cause disease?
- Does it display unique physiology or other biological features? (Does it carry unusual structures? Does it live in a novel environment? Does it go through a complex developmental cycle? Does it produce unusual products? Does it interact in a remarkable way with other organisms? Does it inhabit a unique environmental niche or demonstrate a peculiar lifestyle?)
- Will its study impact other disciplines?
- What is its economic importance? (What impact, if any, will it have on agricultural, pharmaceutical, and/or chemical industries?)
- Does it have other practical significance? (Does it cause disease in plants or animals? Might it be useful for engineering traits in other organisms? Does it provide benefit to another organism? Can an antibiotic be produced?)
- Is there a community of scientists to study the organism so that genomic data can be linked to functional data?
- Will the sequencing be carried out in a cost-effective manner?

The participants stressed that there is no such thing as a poor choice of microorganism to sequence, especially as the costs decrease. Each sequence adds significant value to the pool of available sequences. For this reason, the participants did not prioritize these criteria; their importance depends on the scientific question being asked. Individual investigators should retain a substantial role in the process of choosing microorganisms for sequencing, as these are the researchers who will use the resultant data.

Publicizing lists of organisms in a timely manner whose sequences are being deciphered will increase the efficiency of the global sequencing effort. Sequences held by proprietary commercial interests, however, should probably be ignored; private firms usually elect to keep their sequencing data out of the public sphere and this information is, therefore, of no utility to the scientific

community at large. The participants, however, would like to devise ways to encourage companies to share their sequencing data, at least at some level. For example, the public might fund some private genome projects in exchange for release of proprietary data. Alternatively, companies might be compensated for making their sequence available by bringing in the expertise of outside scientists who might find the biological function and practical use of the sequences quickly, thereby increasing the value of the investment.

Strategies: Who should sequence? Numerous options also exist for how to accomplish the sequencing. At one extreme, central establishments that can produce large quantities of high quality data cheaply and efficiently would perform the sequencing. At the other extreme, individual labs with specific interests and expertise in the biology of particular organisms would compile the relevant genome sequences. All types of hybrid arrangements can be envisioned, and some systems may better serve the goals of particular projects. In general, however, the participants see a need to integrate the strengths of both approaches. Sequencing projects and experimental programs aimed at elucidating the physiological features of microorganisms need to be linked much more firmly than they have been. This connection will ensure that the sequencing data are exploited to their fullest potential. Conversely, a close relationship between sequence generation and inquiry into the function of sequences is likely to guide sequencing programs and tilt them toward the organisms with the most potential for downstream biological investigation. Connecting data generation with application should enhance the quality, efficiency, and productivity of the overall microbial sequencing effort.

The hope is that individual investigators will play a major role in sequencing projects by collaborating with the large-scale sequencing groups. Large centers can exploit economies of scale, speed, access to expensive technologies, and in-house development of specialized expertise and new technologies. Centralized facilities, however, tend to marginalize users of the sequence data—the biologists who will perform experiments to interpret the biological relevance of the genomic information. Depending on the particular project, this problem can be overcome in various ways. Consortia of microbiologists interested in a particular organism or set of organisms might hire out the sequencing or collaborate in some other way with a large sequencing center. Perhaps public funds could support centers at which groups of investigators could have genomes sequenced with the aid of permanent center staff and technology. Access to funding and thus to the use of such facilities could be obtained by the usual scientific peer review process.

A university, in partnership with a federal agency or other universities, might host a genomics center that focuses on a particular subset of organisms. The particular subset would be chosen to best capitalize on the specialties of the local scientists, and different centers would develop different areas of concentration (e.g., marine microbes, soil microbes, skin pathogens, etc.). Multiple possibilities exist for strengthening ties between sequencers and experimental biologists. Colloquium participants encourage an exploration of these and other models.

Data management and analysis

Analytical tools: The bulk of effort and resources invested thus far have been directed toward generating sequences. As a result, sequencing technologies have made major advances, the cost has dropped significantly, and the quantity of genomic data has risen rapidly. The bottleneck in the microbial genomic sequencing endeavor has shifted from data generation to data management and interpretation. This increasing gap between having the data and knowing what it means must be bridged if the associated breakthroughs in medicine, agriculture, industry, evolution, and environmental sciences are to be realized. Part of the current problem is due to a lack of standardization. No one yet knows what the best database system will be. The potential of the entire enterprise, however, will be slowed until everyone adopts the same system.

To exploit in an optimal fashion the data that will continue to pour in from sequencing projects, new management tools are required. These will be used to manipulate sequences in a comprehensive database. In addition to the nucleotide sequence *per se*, researchers need to be able to retrieve large amounts of other information. For example, the database should include the inferred function

Challenges, Opportunities, and Potential Solutions

of a gene (based on similarities to genes with known activities) or, when available, information about the real function of the gene (based on experimental evidence). Other biological observations should also accompany the sequence: how much of the gene product is produced under what circumstances; how its activity or quantity changes in response to particular physical or chemical signals; whether it gets chemically modified and under what circumstances? Furthermore, this type of functional biological data needs to be codified in such a way that it is possible to compare all features associated with a particular gene or section of a genome with a stretch of DNA from another organism. Investigators must be able to probe similarities and differences in features that extend well beyond just the sequence of nucleotides in the DNA, and potentially detect previously unrecognizable patterns. Furthermore, the system should be set up so that it's easy to annotate files after their initial release. There must be a simple way to integrate new data seamlessly.

The type of computer tool to integrate the data in this manner does not yet exist. The field of bioinformatics, which applies the power of computer science to the problems of biology, is still a young science. It requires many forms of support, including money, direction, and personnel. It is very important that significant efforts are invested into planning and providing the most useful computer tools. Moreover, the scientific community needs to know about these new tools as they are developed. Finally, the tools must be accessible and easy to handle. Colloquium participants would like to encourage the eventual development of "Genomics Operating Systems"—self-contained, comprehensive, user-friendly packages to help new groups plan and conduct genomics projects. These large-scale databases and projects to develop informatics tools will require significantly more resources than are currently allocated to such activities. Probably a national effort, in the form of a multi-agency consortium or congressionally funded agency, is required to archive, curate, and distribute the data and the tools with which to use it.

Access and standardization: Sequence information must be made publically available to all communities of scientists, and it must exist in a form that is easy to access. Furthermore, sequences should be available to the public. Financial and other resources will be needed to support such a central repository. Rapid sharing of data about microbial genomes is critical to catalyze experimentation that will lead to potential advances in the many applied and theoretical fields of agriculture, medicine and environmental technology. But timely dissemination of information must be balanced with the need to ensure appropriate completeness and accuracy of the data. Furthermore investigators must have a reasonable opportunity to make use of their data or else they will reap no competitive benefit from generating the sequence.

Currently, genomic data are treated differently from other types of scientific data. Traditionally, researchers are taught that they shouldn't publish results until they confirm the accuracy. However, the Bermuda Statement (http://www.nhgri.nih.gov/Grant_info/Funding/Statements/RFA/data_release.html), recently confirmed through the Clinton-Blair Statement, is a policy statement developed in the context of human, not microbial, genome research. It mandates the immediate electronic release of DNA sequence data funded by public agencies. The rationale for early release of unfinished data is to optimize coordination among the various parties working on the same organism and to facilitate prompt, productive use of the data. For enormous, multi-facility projects such as sequencing the three billion base pairs of the human genome, there is value in quick release since there is such a lag before ultimate completion of the project. This is not true for shorter term, microbial genome projects, and they should be treated differently. Current sequencing technologies can generate one-fold coverage of entire bacterial genomes within a matter of days. Extremely quick data-release policies penalize small groups by forcing investigators to share data before they have had a chance to confirm its accuracy, begin its interpretation, and test hypo-theses. Thus, these policies stifle the effort of sequencers to explore the biology of their target organisms.

Furthermore, there are many anecdotal stories about the unattributed use of preliminary data after it has been posted on the World Wide Web. A more complete discussion of the ethical issues raised by this situation is included below in Ethical Issues section.

Challenges, Opportunities, and Potential Solutions

Many colloquium participants thought that sequence data should be treated the same as other scientific data. Prior to publication, however, groups performing the sequencing should be encouraged to make the data available to interested parties on request, providing the requesting party agrees to use it in a collegial fashion.

In any case, standardized guidelines should be adopted to address when, in the course of a project, data should be released. The National Institute of Allergy and Infectious Diseases (NIAID) has set forth a policy for microbial genome activities calling for release of data one month after three-fold coverage of the genome (or chromosome, in the case of a eukaryotic organism) has been achieved. A discussion should take place as to whether it is appropriate for the general community to adopt this rule. At a minimum, data should be released either at the time of publication or when the grant that funded the sequencing expires.

Timeliness of data release relates to accuracy as well as to access. Sequencing errors are inherent to all current methods and it is necessary to sequence any stretch of DNA multiple times before it can be considered correct. Furthermore, current methods of "shotgun" sequencing, in which random pieces of DNA are sequenced and then reassembled by a computer, inevitably leave holes in the compiled data.

Unfinished sequence is often of suitable enough accuracy and completeness to provide some immediate help to researchers; it can be argued that waiting until the difficult process of closing gaps has been achieved before releasing the sequence is too cumbersome. Furthermore, for some purposes, partial sequencing may be appropriate. "Leveraged sequencing," in which two genomic sequences are compared after one-fold coverage, is a way to uncover large differences between closely related organisms. This method has been proposed as a way to compromise between limited funds and the need for more sequence data. While it is useful for uncovering large differences, it does not reveal smaller, yet often important, ones. As technology improves and sequencing costs plummet, discussion of the utility of partial sequence or leveraged techniques should become moot. In the meantime, it should be kept in mind that, while a 90% completed sequence can be a useful tool, the unfinished 10% may contain some of the most valuable genetic data of all. Although partial information serves some purposes, complete genome sequences should be the ultimate goal for most projects.

Different groups and researchers disagree on when to call a genome sequence "complete." The definition should be standardized and applied more uniformly. Furthermore, at some point, the DNA itself must be made available. This could pose an unacceptable burden on an investigator. The American Type Culture Collection (ATCC) or other companies should be encouraged to undertake this role. Finally, the community needs to develop quality standards for acceptable error rates for sequence data.

Getting at function

The DNA sequences themselves represent just the beginning of genome-level study. Computer analysis can predict which portions of these strings of nucleotide letters are likely to represent genes or signals that flip genes on or off. It can even discern similarity to genes of known function, and thus suggest biological roles for the new piece of DNA. But similarity is not function. Proteins that share certain features may perform similar biochemical tasks, but in different pathways in the cell. They also might possess additional abilities not readily recognizable from the sequence. Furthermore, approximately 30-40% of microbial genes thus far sequenced are "orphans," having no close relatives at the level of DNA sequence. Sequence points the way toward function, but it is the microbiologist's challenge to translate these inanimate pieces of information into cellular activities. Currently, investigators are shaping new processes and tools with that goal and approach in mind, as well as adapting older methods for use with large amounts of sequence data.

Already, information in genome sequences is saving investigators significant amounts of time. For example, researchers have devised ways to determine which random pieces of a genome are activated under particular conditions. Dozens of genes can be identified simultaneously that behave in this fashion. By employing these techniques, investigators have discovered genes that are turned on only when inside an animal host, for example; these methods can be used for probing

Challenges, Opportunities, and Potential Solutions

other environments as well. This genre of experimental approach, as well as conventional genetic screens, is greatly aided by complete genome sequences. As investigators isolate pieces of DNA that get triggered in response to a particular environmental condition, they can compare them to the known sequences of the entire genome simply by logging on to a computer and probing a sequence database. In this way, they can jump in mere minutes from a small span of sequence that they know is important (from their experiment) to the entire gene to which it corresponds. Previously, an investigator would spend days or weeks generating the sequence of that same gene.

Generating the sequence of entire organisms opens up the possibility of analyzing function on a whole genome level. This new field of genetics—commonly called functional genomics—has begun to blossom as investigators develop new techniques for this purpose. The general idea is to analyze many genes in parallel.

With completed genome sequences, it is now possible to represent every gene in an organism's genetic store on a glass surface within an area the size of a postage stamp. Scientists have developed DNA chips (also called microarrays) that permit the analysis of tens of thousands of genes simultaneously from each of hundreds of samples. Typically, using such devices, investigators can generate information about the activity of a particular gene in multiple organisms or the activity of all genes in a single organism. In this way, a snapshot of biological activity within an organism can be produced, and compared, for example, with a picture of the same organism under a different environmental condition. This approach saves extraordinary amounts of time compared to conventional molecular biology approaches that require independent manipulations of each sample or small groups of samples. It is a powerful way of asking the same question of many different samples or genes.

Scientists are devising other approaches to screen entire genomes for biologically relevant activities. Some of the methods analyze the protein products of genes as opposed to measuring whether genes are on or off. One current scheme identifies protein pairs that bind to or interact with each other. This method permits comprehensive studies on sequences that encode a complete set of proteins from any organism.

Scientists are just beginning to devise efficient ways to probe the complex network of interactions that constitute a living organism. These techniques hold tremendous power and promise in analyzing biologically relevant features of organisms at the genomic level. However, producing these chips and performing the accompanying genome-level analyses is expensive. While some innovative labs are making their own arrays, the high overhead and need for a critical mass of local expertise proves a hindrance. Capital equipment and consumable materials on this large scale are expensive. Furthermore, large numbers of people (and their associated salaries) and/or costly robots are required for some types of experiments. Currently, these methods are not accessible to the average academic researcher.

While headway is being made to tackle biological function on a genomic level, we have seen only the first examples of such techniques. Similarly stiff equipment demands and costs are likely to impede most biologists from using procedures developed in the near future that can analyze large numbers of samples quickly and in parallel. In order for the costs to drop, standard approaches will presumably need to gain acceptance so companies have an incentive to invest their resources into developing less expensive methods. Because different research groups are currently taking different tacks, any one method is an expensive and risky proposition because there are a limited number of customers.

The development of innovative, high-throughput technologies that address the function of newly sequenced genes should be encouraged. Furthermore, an organized discussion about needs and possible approaches to new strategies would serve the community well in a number of ways, including pointing out the requirements for common tools and approaches. It is especially critical to maintain close ties between advances in functional analysis and new computer tools; different technologies aimed at probing biological function will require particular types of informatics support.

Challenges, Opportunities, and Potential Solutions

Funding

A large increase in funds is required to finance the sequencing and study of microbial genomes on a scale that would take advantage of existing technological and scientific opportunities and bring the most benefit to society. Microbes offer tremendous diversity that can be exploited for a multitude of purposes. In order to capitalize on this potential, thousands of these organisms should be studied.

In addition to more funding, grant procedures need to be adjusted to bolster this new field. Some aspects of genomics projects, such as large-scale sequencing efforts, don't fit well into common National Institutes of Health (NIH) funding mechanisms. The grants that support most academic labs (R01 grants) are not geared toward projects of such large scope or cost. As the field of microbial genomic science moves forward, many types of worthwhile projects will emerge that current funding programs are not geared to accommodate. Funding mechanisms and formats should keep up with this changing set of demands. Some aspects of genomic inquiry, such as conventional approaches toward elucidating biological function, might continue to fit into existing funding pathways.

It is critical that funding of sequencing efforts be coordinated with funding of informatics tools with which to analyze the data, and with subsequent investigations into biological function. All components must garner sufficient support in order to procure the potential value from microbial sequences. The total amount of funds allocated for microbial genomics should be increased dramatically, without redistribution of funds from other areas of biology.

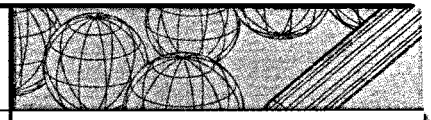
As indicated elsewhere in this report, large amounts of money are required to build and sustain a productive microbial genome enterprise. Funds are required to generate and analyze data; support innovation in all aspects of getting from sequence to function; coordinate activities; ensure that the data produced are easily manipulable and available to the scientific community; and foster an environment in which scientists from different disciplines are collaborating. Governments, foundations, and private companies are currently financing genomic sequencing efforts, but an even larger commitment is needed. Colloquium participants suggested an interagency collaborative effort at the federal, and eventually international, level. Federal research agencies could participate. Such a coordinated program would help combat the problems associated with the lack of organization demonstrated thus far.

Training and Education

This exciting juncture, where the power of genomic information meets the field of microbiology, is creating a need for a new breed of microbiologist and interdisciplinary scientist. In the genomic era, investigators will manipulate large quantities of data. The associated reliance on computer databases and informatics techniques is defining a new field of computational molecular biology. The community needs people who understand the demands and perspective of the biologists who will use the sequence data to infer and uncover function of the corresponding genes, and who can envision and write the most efficient and powerful programs for allowing biologists to plumb those data. This represents a new type of expertise—one that encompasses both a solid foundation in computer science and a sophisticated understanding of molecular biology.

In addition to this new brand of scientist, traditionally trained physiologists, geneticists, biochemists, and molecular biologists will still be needed. As the sequence database expands, there will be a concomitant need for people who can employ nucleic acid sequence information to design experiments to reveal the function of the sequences and test the novel hypotheses generated. These experts of diverse background must find common language and joint work settings for fostering and promoting these new relationships. As tools are developed that analyze all of the genes or proteins in an organism in parallel, it is possible that additional specialties will evolve. Such a development, and anything else that advances the pace of discovery associated with microbial genome projects, should be encouraged. We must ensure, therefore, that this genomic enterprise nurtures scientists who can bridge the biology-computer science gap, while maintaining a pool of investigators who can perform the irreplaceable job of direct experimental inquiry on microbes.

The genomic era will make new demands on the public as well as on specialists. People from all walks of



Challenges, Opportunities, and Potential Solutions

life will need to discuss various ethical and practical issues that arise from this new endeavor. (See below for a discussion of some of the ethical issues.) It is critical, therefore, that lay people understand the promise of the effort and can converse about the relevant issues. To that end, education about the power of microbiology and genome sequences should be boosted at all levels.

Ethical issues

Microbial genome projects raise a new type of ethical issue for scientists, associated with the reporting and handling of sequence data. Currently, genomic data are treated differently than other types of scientific data. Guidelines devised for the human genome project dictate the immediate electronic release of DNA sequence data funded by public agencies. This early release means that other investigators have access to the information prior to publication. Numerous issues pertaining to the timing of data release have been discussed above (Data management and analysis). The current situation raises some ethical challenges.

There have been some reports of unattributed use of preliminary data that has been gleaned from the Internet. Whether such use technically qualifies as egregious misconduct hinges on whether the posting of data at a website constitutes publication of those data. Nevertheless, the use of Internet data without proper attribution would seem to qualify as a new form of electronic plagiarism. Many individuals, however, may not realize that they are violating the traditional values of the research enterprise when they co-opt these data, especially given the Internet's reputation and function as a medium for free and open sharing. A clear statement on the ethics of electronic data usage and attribution should be developed for the sequencing community. The community needs explicit mechanisms for establishing and teaching the norms of proper behavior.

These are just some of the issues confronting scientists that are arising from the microbial genome sequencing projects. Currently, no committee is charged with tackling these issues. Representatives from different segments of the community that study microbial genomes could help organize a discussion about such challenges, devise solutions, and implement them.

As with many aspects of contemporary scientific inquiry, new ethical issues for the public as well as the scientist are emerging. In addition to a need for a mechanism to transmit professional ethics within the community of microbiologists, there is a need to develop public outreach programs that address societal concerns associated with microbial genomic investigations.

Religious leaders, environmental activists, and many other concerned citizens have raised thoughtful questions about the long-term implications of deciphering "blueprints for life." In effect, genomic research extends the debate that has raged for 25 years about the safety and morality of genetic engineering.

The ethical and societal implications of genomic research often extend beyond the professional competence of microbiologists. What does it mean to know the complete DNA sequence of the Ebola virus, the anthrax bacterium, and other agents of germ warfare? Does microbial genomic inquiry increase or decrease the risk of biological terrorism? Will genomic research expand the use of genetically modified organisms in human food? What are the implications of microbial genomic research for the release of genetically modified organisms into the environment?

The 20th century has taught us many lessons about the unintended, negative consequences of supposedly benign and life-enhancing scientific research. These questions need to be considered carefully. Public forums should be held to air and debate these and other concerns. The input of ethicists, sociologists, theologians, and other experts outside the scientific community should be enlisted to provide insights from a variety of perspectives.

Roles for Professional Societies

Currently no organization or group is providing leadership for the burgeoning field of microbial genomics. Professional societies, such as the American Society for Microbiology, have the expertise to guide the scientific community and the public at large on a wide variety of issues associated with this area. Because of its large size (over 42,000 members) and century-long commitment to the biological sciences, the American Society for Microbiology is in a good position to ensure the

continuity of efforts to collect and maintain sequence and associated data, and, in general, to play a proactive role in keeping the microbial genome sequencing endeavor vigorous and productive. ASM and other appropriate professional societies should consider the following:

- Establish a council-level committee on microbial genomics. This committee could help ensure that the society is able to fulfill the needs and goals outlined in this report.
- Advocate for an interagency effort specifically devoted to advancing the microbial genome enterprise.
- Work for broader societal support and understanding of genomics research. Such efforts might include devising school curricula about microbial genomics, relaying information about developments to the media, and advising public figures about relevant issues.
- Provide members with up-to-date information so they are prepared to champion the value of microbiology and genomic research to public, private, and governmental agencies.
- Ensure that educational institutions and programs can support and enhance the growing enterprise of microbial genomics. Such efforts might include encouraging universities to rebuild microbiology infrastructures; encouraging graduate programs to apply for pre-doctoral training grants in the field of microbial genomic science; and in setting up short courses (at Woods Hole or Cold Spring Harbor, for example) in basic bioinformatics to help investigators established in traditional microbiological disciplines to re-tool and re-train for genomics studies.
- Provide a clearinghouse for microbial sequencing (and associated) data. The organization might help organize and disseminate the data itself. It also might play a lead role in affording scientists easy access to information about sequencing projects that are under way.
- Develop a clear statement on the ethics of using electronically available sequence information
- Develop and promote uniform standards to ensure the quality of sequence data.
- Stimulate improvements and cost-reduction efforts in laboratory and computer techniques.
- Devise strategies that will encourage private firms to share their sequencing data.
- Encourage research on relevant ethical issues and organize forums in which ethical issues relating to non-human genomics studies would be publicly discussed.
- Make itself (or its members) available to advise funding agencies and other organizations or individuals who need guidance.
- Organize meetings with agency and institute directors to address the needs of microbial sequencing projects.
- Solicit contributions for a microbial genome trust fund.
- Join with its counterparts in other nations to encourage worldwide genomics research activities to facilitate international cooperation and funding.

Specific Recommendations

Colloquium participants developed a number of recommendations for the future. Those that were unanimously endorsed are the following:

- Develop a large-scale, coordinated, multi-agency, microbial genome sequencing effort. Stimulate collaborative projects between a sequencing group and a group that understands the biology of the organism and is well equipped (in terms of training and tools) to probe hypotheses that arise from the new sequence information with experimental approaches. Create a mechanism for open discussion within the scientific community about evolving needs, priorities, and planning.

- Coordinate all aspects of microbial genome projects among interested parties. This should aim to organize efforts to gather, disseminate, and use microbial genome sequence information. It should also cultivate efforts to promote all aspects of necessary continued support: funding, development of data management tools, creation of ethical and practical guidelines and standards, smooth input and access to data, constant examination of all aspects of planning appropriately for the next phase, and efficient use of all types of resources (human, financial, technological) to hone the genome sequencing endeavor and ensure its success.

- Provide a central and easily accessible location for completed genomic sequences and associated information about the biology associated with it. Create standardized and easily accessible sequence annotation and computational tools so users can efficiently and productively use the data.

- Devise new guidelines for data release, other issues of standardization, and for ethical issues related to public electronic databases. Establish a multi-agency committee that will organize a discussion among interested parties.

- Organize a mechanism with which to address societal concerns and inform the public about the benefits and implications of a large-scale microbial genome sequencing effort. Ethical issues that might be discussed include whether and to what degree sequences and their products could become private property, how to curb attempts to exploit microbial genome information for malevolent purposes, and whether and how to preserve microbial diversity.

- Modernize the K-12 educational curriculum to cultivate a society whose members understand the importance and ramifications of microbial genomics.

- Enhance high-level training and education in genomics and associated technologies. Encourage academic programs and departments to submit training grants that focus on some aspect of microbial genomics. Encourage funding of postdoctoral work in this area by stimulating the creation of fellowships.

- Nurture the development of methodology that reveals the biological role of microbial DNA sequences. To help speed movement from sequence to function, create funding opportunities for new experimental methods as well as for development of data management tools. These technologies and approaches should include novel ways to infer gene and protein function at the level of the individual gene/protein as well as at the level of all the genes/proteins within an organism. It is vital to fuel enthusiasm for new approaches without discouraging the use and development of the old.