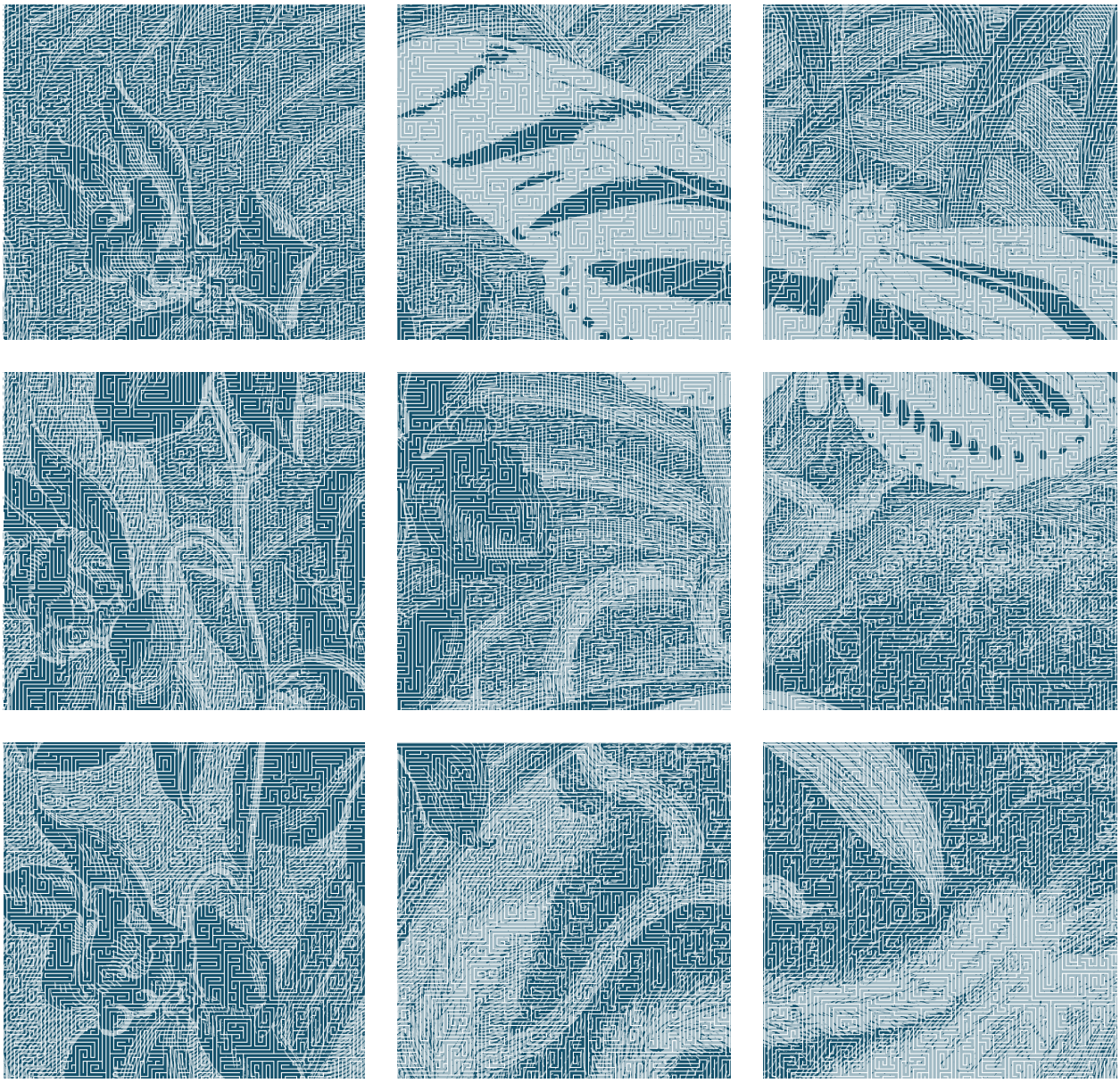
A report from the American Academy of Microbiology



An Experimental Approach to Genome Annotation

AMERICAN SOCIETY FOR MICROBIOLOGY

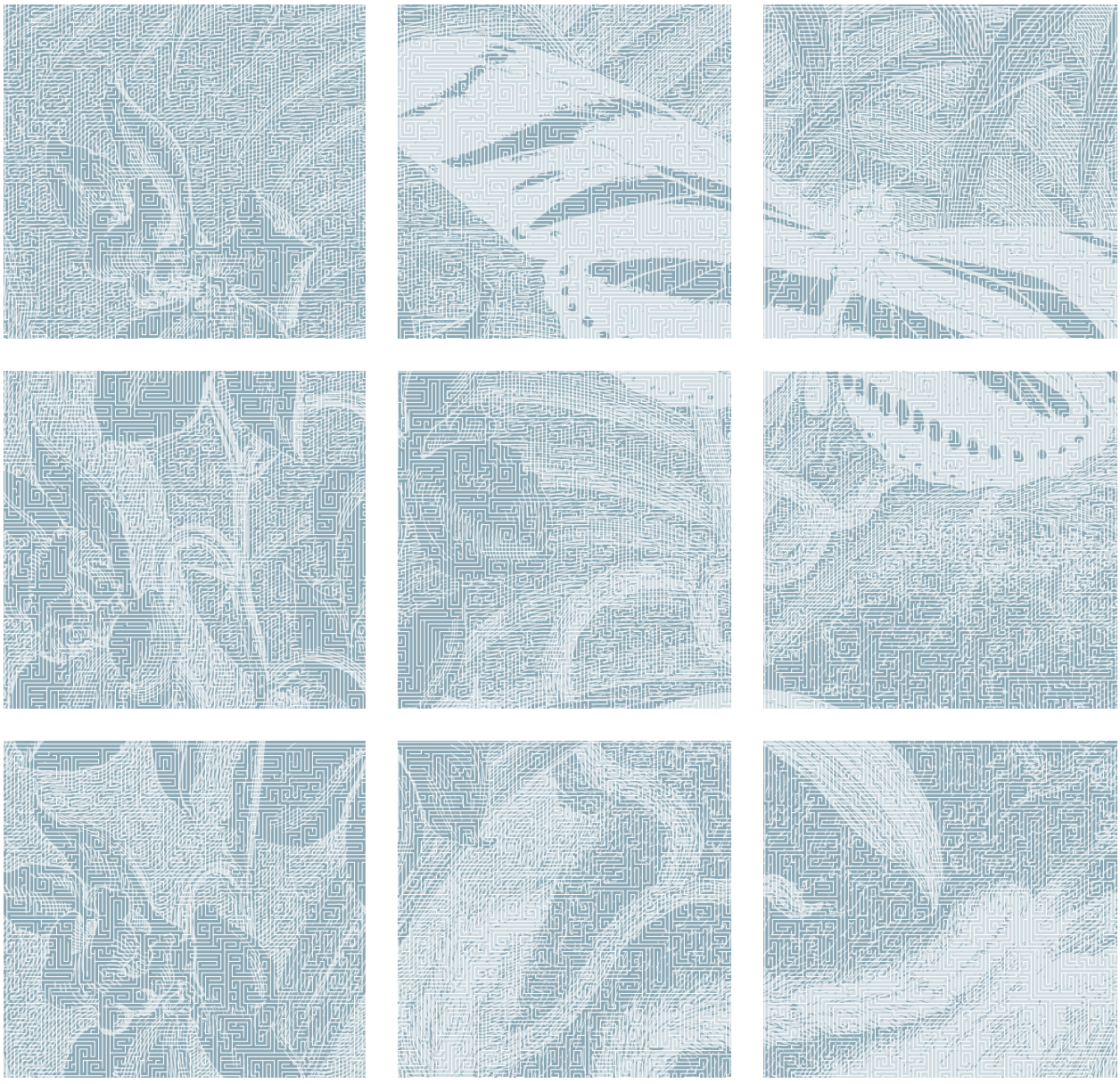
American Academy of Microbiology
1752 N Street, NW
Washington, DC 20052
http://www.asm.org

This report is based on a colloquium sponsored by the American Academy of Microbiology held July 19-20, 2004, in Washington, DC.

The American Academy of Microbiology is the honorific leadership group of the American Society for Microbiology. The mission of the American Academy of Microbiology is to recognize scientific excellence and foster knowledge and understanding in the microbiological sciences.

The American Academy of Microbiology is grateful for the generous support of the National Science Foundation. The American Academy of Microbiology strives to include women and underrepresented scientists in all activities.

An Experimental Approach to Genome Annotation



By Richard J. Roberts, Peter Karp, Simon Kasif, Stuart Linn, and Merry R. Buckley

Board of Governors, American Academy of Microbiology

Eugene W. Nester, Ph.D. (Chair)
University of Washington

Kenneth I. Berns, M.D., Ph.D.
University of Florida Genetics Institute

Arnold L. Demain, Ph.D.
Drew University

E. Peter Greenberg, Ph.D.
University of Iowa

J. Michael Miller, Ph.D.
Centers for Disease Control and Prevention

Stephen A. Morse, Ph.D.
Centers for Disease Control and Prevention

Harriet L. Robinson, Ph.D.
Emory University

Abraham L. Sonenshein, Ph.D.
Tufts University Medical School

George F. Sprague, Jr., Ph.D.
Institute for Molecular Biology, University of Oregon

David A. Stahl, Ph.D.
University of Washington

Judy D. Wall, Ph.D.
University of Missouri

Colloquium Steering Committee

Richard J. Roberts, Ph.D. (Chair)
New England Biolabs, Beverly, Massachusetts

Peter Karp, Ph.D.
SRI International, Menlo Park, California

Simon Kasif, Ph.D.
Boston University

Stuart Linn, Ph.D.
University of California, Berkeley

Carol A. Colgan
Director, American Academy of Microbiology

Colloquium Participants

Cheryl Arrowsmith, Ph.D.
University of Toronto, Ontario, Canada

Tadhg Begley, Ph.D.
Cornell University

Robert Bender, Ph.D.
University of Michigan

Barry R. Bochner, Ph.D.
Biolog, Hayward, California

Eric Brown, Ph.D.
McMaster University, Hamilton, Ontario, Canada

Frank Collart, Ph.D.
Argonne National Laboratory

Valerie de Crecy-Lagard, Ph.D.
The Scripps Research Institute

Andras Fiser, Ph.D.
Albert Einstein College of MedicineL

Michael Y. Galperin, Ph.D.
National Library of Medicine,
National Institutes of Health

Jon Goguen, Ph.D.
University of Massachusetts Medical School

Howard Goldfine, Ph.D.
University of Pennsylvania School of Medicine

Eugene Kolker, Ph.D.
Biatech, Bothell, Washington

Eugene Koonin, Ph.D.
National Library of Medicine,
National Institutes of Health

Frank W. Larimer, Ph.D.
Oak Ridge National Laboratory

Thomas Leyh, Ph.D.
Albert Einstein College of Medicine

Paul Ludden, Ph.D.
University of California, Berkeley

Edward Marcotte, Ph.D.
University of Texas at Austin

Kenneth Nealson, Ph.D.
University of Southern California

Eugene Nester, Ph.D.
University of Washington

Andrei Osterman, Ph.D.
The Burnham Institute, La Jolla, California

Julian Parkhill, Ph.D.
The Anger Centre, Hinxton, Cambridge, England

Dan Robertson, Ph.D.
Diversa Corporation, San Diego, California

Margaret Romine, Ph.D.
Battelle Pacific Northwest National Laboratory

Steven Salzberg, Ph.D.
The Institute for Genomic Research, Rockville, Maryland

Jeff Skolnick, Ph.D.
Buffalo Center of Excellence in Bioinformatics,
Buffalo, New York

Gary Stormo, Ph.D.
Washington University School of Medicine,
St. Louis, Missouri

Alfonso Valencia, Ph.D.
CNB-CSIC, Madrid, Spain

Eric Vimr, Ph.D.
University of Illinois

Jeremy Zucker, Ph.D.
Dana-Farber Cancer Institute, Boston, Massachusetts

Observers

Elaine Akst, Ph.D.
Fundamental Space Biology Division,
National Aeronautics and Space Administration

Patrick P. Dennis, Ph.D.
Division of Molecular and Cellular Biosciences,
National Science Foundation

Daniel W. Drell, Ph.D.
Health Effects and Life Sciences,
U.S. Department of Energy

Irene Anne Eckstrand, Ph.D.
National Institute of General Medical Sciences,
National Institutes of Health

Brad W. Fenwick, Ph.D.
U.S. Department of Agriculture, Kansas State University

Maria Y. Giovanni, Ph.D.
National Institute of Allergy and Infectious Diseases,
National Institutes of Health

Maryanna P. Henkart, Ph.D.
Division of Molecular and Cellular Biosciences,
National Science Foundation

John Houghton, Ph.D.
Biological and Environmental Research,
U.S. Department of Energy

Eric Jakobsson, Ph.D.
National Institute of General Medical Sciences,
National Institutes of Health

Matthew D. Kane, Ph.D.
Division of Environmental Biology,
National Science Foundation

Rachel E. Levinson
Office of Science and Technology Policy

Ann Lichens-Park, Ph.D.
U.S. Department of Agriculture

Anna Palmisano, Ph.D.
U.S. Department of Agriculture

Joanne Tornow, Ph.D.
Division of Molecular and Cellular Biosciences,
[o]National Science Foundation

Executive Summary

The American Academy for Microbiology convened a colloquium July 19-20, 2004, in Washington, DC, to address the critical challenge of prokaryotic genome annotation and to seek ways to accelerate progress in the field. Recent advances in DNA sequencing have produced a spectacular amount of new data; literally hundreds of thousands of sequenced prokaryotic genes now await annotation. These genes can be enumerated, compared, and grouped by sequence similarity into families, yet an understanding of their biochemical functions is lacking. Genomics provides that rare opportunity in science where the boundaries of current knowledge can be clearly defined. The annotation initiative proposed in this document will extend those boundaries and will likely lead to new applications and new progress in healthcare, biodefense, energy, the environment, and agriculture. This research could also impact many commercial enterprises, such as the chemical, food and dairy industries.

Colloquium participants included microbiologists, biochemists, and bioinformaticians. Observers from the National Institutes of Health, the National Science Foundation, the U.S. Department of Energy, the National Aeronautics and Space Administration, the Office of Science and Technology Policy, and the U.S. Department of Agriculture were also in attendance. Participants discussed the currently available sources of genome annotation information and the strengths and limitations of those sources. Four areas of concern in genomic annotation were identified:

(1) As many as 40% of all predicted genes in completed prokaryotic genomes have no functional annotation.

(2) Many genes have a predicted function, but that prediction has not been experimentally validated.

(3) As many as 5-10% of predicted gene functions may be incorrect.

(4) Many known enzymes have no corresponding genes identified in the sequence databases.

Much of the currently available annotation information is provided by computer programs that predict the functions of newly sequenced genes on the basis of their similarity to genes of known (or predicted) function. This technique is inherently limited in both breadth and accuracy by the small size of the core foundational set of genes with experimentally established functions. By expanding that foundational set through a systematic program of biochemical study of genes of unknown function, we can dramatically increase the quality of prokaryotic genome annotations, and enhance our understanding of current and future genome sequences.

The experimental elucidation of function for a hypothetical gene can be a significant challenge for the biochemist. However, in the past five years new bioinformatics techniques, mostly based on comparative genomics, have been developed that can provide clues about the function of a gene. Functional genomics methods, such as gene expression chips, can also provide hints about gene function. Such clues can greatly accelerate experimental studies by suggesting plausible hypotheses to be tested in the laboratory.

Colloquium participants agreed that accurate and complete annotation is vital to making full use of genomic data. However, there are great deficiencies in currently available annotation sources. Moreover, there are few sources of dedicated funding for experimental approaches to annotation. In light of these facts, it was recommended that a new initiative be undertaken that would synergistically combine computational methodologies for functional prediction with a systematic experimental approach to test those predictions. It would also broaden the foundational set of experimentally determined gene functions by finding missing genes for known enzymatic functions. Such a program would both increase experimental knowledge and spur further accuracy in bioinformatics prediction leading to repeated cycles of validation and prediction.

As part of the proposed initiative, a new resource focused on annotation should be developed. The central component is a database containing:

- Predictions regarding the functions of genes of unknown function, deposited by bioinformaticians, based on computationally inferred clues, which will serve as a starting point for experimental investigations.
- The results, positive or negative, of those experimental investigations, which in many cases will establish new gene annotations backed by rigorous experimental work.
- A prioritized list of sequenced genes for which no functional information is currently available.
- A list of biochemically-characterized functions for which no gene has yet been assigned (referred to as orphan functions).
- Data on previously characterized proteins currently in the public databases.

The basic design of the database was discussed, and recommendations for hosting, administration, and management of the database were put forth.

Achieving an accurate and detailed annotation of a newly sequenced genome is a critical, but often difficult, step in the process of analyzing the sequence data. This is especially difficult for organisms where genetic tools have yet to be developed. Unfortunately, the pace of experimental elucidation of gene function is very slow compared to the pace of sequencing and computational prediction of function. Thus, rather than attempt to experimentally explore the functions of every unknown gene in every sequenced genome, it is preferable to focus experimental investigations on the most informative targets. For instance, the scope and accuracy of existing bioinformatics techniques would be greatly enhanced by obtaining one or a few experimental functions for members of gene families that are found in many organisms. This experimental annotation initiative will encourage and enable experimental biochemists to participate in the annotation of prokaryotic genomes.

The initial focus of this particular initiative would be prokaryotes – bacteria and archaea – because (a) they possess relatively small genomes comprising genes that are usually easily defined, (b)a great deal of prokaryotic genomic sequence data is available in the public domain, and (c) because they are experimentally tractable. Schemes need to be developed to determine which among the prokaryotic gene products and orphan proteins should receive attention first. In one possible plan, priority would be given to families of similar genes that are found in many different genomes, because determining a biochemical function for one member will likely implicate all family members as possessing the same or a similar function.

The details of the bioinformatics part of the initiative, such as database design and operation, should be open to the discretion of those researchers who apply for funding to construct it, but certain broad recommendations for the content and administrative aspects of the resource were formulated by the colloquium participants. For example, the database should include not only protein gene products but also functional RNA products. It was stressed that the input of both bioinformaticians and experimentalists would be vital to the success of the initiative, and their collaboration should be encouraged. The creation of an external database advisory board was also recommended. Funding would be required to support the bioinformaticians who will make and evaluate the bioinformatics predictions and generate and maintain the computational resource. However, the largest requirement for funding would be to support the experimental biochemical work testing bioinformatics predictions. It was proposed that one or more pilot projects be undertaken to assess the feasibility of the approach before embarking on a large scale initiative.

The potential impact of the proposed initiative is difficult to overstate since it would affect all aspects of biology. The participants feel that this project is essential to enable the next step in moving genomic science forward from accumulating a large depository of sequences towards achieving a true understanding of the basic elements of prokaryotic biology. Without a forward-looking initiative like the one proposed here, the functional data needed to propel systems biology forward will not be available, and those trying to understand the complex interactions of genes and their products in living cells will continue to work with many components of unknown function. In addition, elucidating the enzymatic functions essential for prokaryotic life will impact our understanding of eukaryotic organisms, which possess many of these same genes. This initiative will also foster closer collaborations between experimental and computational scientists and help to reinstate the importance of biochemical research. Finally, although much of the project will focus on traditional biochemistry, the initiative can be expected to stimulate new advances in functional screening, new functional genomic technologies such as phenotype arrays, and significant industrial and commercial opportunities in the form of new targets for both medical and industrial applications of prokaryotic biology.

Introduction

Since the emergence of large-scale sequencing technologies in the 1990s, the complete genomes of hundreds of organisms have been sequenced and archived. The success of these sequencing programs has been breathtaking. Today, the genomes of organisms as diverse as bacteria and humans and as extraordinary as puffer fish and loblolly pine have been catalogued. Sequences abound, and we are in the midst of a genomics revolution.

However, the more difficult work of interpreting genomic sequences has hardly begun. Roughly 40% of predicted genes have not been assigned even tentative functions. It is rare in science to be able to clearly delineate the boundaries of current knowledge, but that is exactly where genomics stands today. The genome sequences available at this time are a great resource, and to fully realize the potential they represent they must be annotated accurately through a combined approach of bioinformatics and experimentation. Since many genes are found in some form in more than one species,

assigning function to any individual gene can impact our understanding of many different organisms. Hence, a focused effort in functional annotation of individual genes can have an extensive impact on our understanding of many species and systems. Within these sequences lie novel drug targets, new enzymes for biotechnology, and likely an abundance of novel regulatory elements.

Simple annotation is only a first, but essential, step in understanding the complexity of the whole organism. Functional annotation should be more than a cataloguing of protein functions. It should include information about the interactions of the gene products – these interactions result in a hierarchy of function about which we know very little. The consequence of these interactions is a system that is far greater than the sum of its parts and results in a self-replicating organism. One goal in determining the functions of the genes, therefore, is to provide a basis from which the workings of the whole organism can be explored and ultimately understood. This is the realm of systems biology.

Systems biologists seek to achieve an understanding of the whole organism by studying the products of the genome, how they interrelate, and how they come together in synergistic networks that accomplish the complex functions of life. Because systems biology deals directly with the products of the genome, the field relies heavily on careful functional annotation to describe the individual roles of these products. Unfortunately, efforts in systems biology are presently hampered by the slow pace of gene annotation. It is therefore essential that we uncover the full biochemical potential of prokaryotic genes if we are ever to understand the internal machinery of the simplest forms of life.

The slow pace of functional annotation

Despite the necessity of experimental, functional annotation, progress in this area has been sluggish, resulting in calls for action from members of the genomics community (Roberts, RJ, PLoS Biology 2: E42, 2004; Karp, PD, Genome Biol. 5: 401, 2004). The slow pace of annotation has led to the current situation, in which a large number of putative gene annotations are based on a much smaller foundational set of functionally characterized gene products. This inverse pyramid is not a satisfactory basis from which to form hypotheses. In this situation, with limited breadth of knowledge about the diversity of gene functions, functions can be easily misconstrued, and too much confidence might be placed on sequence correlations between distantly-related genes. To make full use of the resource of genomic sequences, therefore, experimental annotation of gene products must be given high priority.

One reason for the current gap between sequence determination and the functional characterization of genes is a dearth of funding opportunities for annotation. Up to now, funding has been scattered and fairly limited, posing a significant barrier to progress in annotation. The current funding paradigm in the United States does not allow for short, intensive investigations of the functions of individual genes or gene families; funding sources are more focused on high throughput research strategies or directed towards particular biochemical problems of interest to an investigator.

Current sources of functional annotation and their limitations

A number of different sources of functional annotation, some well-known, others more obscure or under development, are presently available to researchers, but they have a number of limitations and do not meet all of the requirements of current research. The scientific literature is the best source of annotation, but the relevant information is often scattered among many papers and may not be widely known. Moreover, a great deal of relevant information was published prior to availability of genome sequences, but this information was not consistently incorporated into current lists of annotations.

The advent of computational tools has also caused some problems in acquiring and compiling annotation data. As computational methods have come to dominate genome annotation, early annotation errors have been propagated. Also, functions discovered since the first annotation was made have been missed in some cases and have not been incorporated into the databases. To address these problems, the field requires a single, central source of annotation information that is regularly updated and is distinct from the current archival databases. This resource could be used as a point of reference for future annotators and other biologists.

The annotation sources available today frequently apply thresholds of evidence that must be exceeded before a given annotation can be included in a database. This approach has the advantage that it avoids misleading users by omitting wildly speculative functions. However, clues that might be indicative of a gene's function, but do not rise to the standards required by the genome annotators, are unavailable to researchers. The reliability of annotations made available through these resources is also a concern, as databases rarely provide quantitative estimates of the trustworthiness of the evidence used in annotation. Moreover, the experimental methods employed vary among research groups. As a result, incorrect annota-

tions made using faulty techniques can go undetected and may be propagated from one genome to another.

Updating annotation information can also be difficult. New information and gene product characterizations that appear in the literature often fail to be transmitted to electronic annotation sources because updating procedures have not been put in place. Conversely, making annotation errors known to the scientific community also poses a problem and is often neglected. (Model organism databases such as EcoCyc, Flybase, and the Saccharomyces Genome Databases are exceptions in that these databases perform intensive literature-based curation efforts that extract new experimentally determined functions from the literature and update their respective databases. However, these efforts are limited to less than 10% of current genomes and the annotations are not automatically passed to related genes in other organisms. Furthermore, these databases do not currently store predictions or clues about function from either computational or experimental methodologies.)

Finally, the names that these annotation sources employ for genes and gene products elicit a great deal of confusion among researchers. Currently very few abstracts that describe protein functions use standard functional descriptions, or systematic gene identifiers.

Recent breakthroughs in bioinformatics

The majority of bioinformatics function prediction algorithms infer the function of a gene based on similar-

❉ *The Enzyme Genomics Initiative* ❉

One group has initiated systematic work to find the genes associated with known enzymatic functions. SRI International has begun a project called the Enzyme Genomics Initiative, the goal of which is to find at least one genetic sequence for each known enzymatic function. SRI's estimates, taken from cross-referencing the EC numbers available in the ENZYME database with information on enzyme genes from a number of databases, including Swiss-Prot and TrEMBL, conclude that there are at least 1,400 enzymes for which the genetic sequence is not known. SRI refers to these enzymes as "orphans", a term suggestive of their indefinite genetic heritage.

Researchers working on the SRI initiative have devised a ranking system for predicting the ease with which the gene for a given orphan could be revealed. Highest marks are given to those enzymes that come from an organism whose genome has been fully sequenced or for which at least part of the protein sequence is known. In these cases, identifying the gene simply requires computationally matching the biochemical properties of the enzyme to the open reading frames in the genome sequence and then testing the prediction experimentally. A second priority is given to those enzymes that have been experimentally characterized recently. Lesser marks are granted to enzymes with available information regarding certain physical characteristics, like molecular weight, isoelectric point or the results of protease digestion.

See http://bioinformatics.ai.sri.com/enzyme-genomics/ for more information on SRI's Enzyme Genomics Initiative.

Taken from a lecture presented by Peter Karp, SRI International

ity of its sequence to that of previously characterized genes. But what if none of the homologs of a gene have assigned functions? What if a gene has no homologs in the public sequence databases? These situations present serious roadblocks for applying sequenced-based methods for prediction of gene function.

A lack of hints about the possible function of a gene also poses a problem in applying experimental approaches to functional characterization. Hints about gene function are helpful because they can indicate which biochemical assays should be applied to confirm or refute the hypothetical function. Knock-out mutants for the gene of interest could provide clues to function, unless a knock-out has no observable phenotype. Gene expression studies can also provide hints in cases where a gene is co-expressed or co-regulated with a relatively small number of other genes with common functions.

Computational approaches can also provide indications of gene function. Recent breakthroughs in bioinformatics have produced three classes of computational techniques that can increase the chance that experimentalists will have a starting point for their research. The first class of methods is based on comparative genomics techniques that infer functional associations between genes from gene patterns observed across many genomes. For example, imagine that a gene of interest, A, is adjacent in an organism of interest to gene B, whose function is known. By itself this observation is of little consequence since the neighbors of a gene frequently have completely unrelated functions. But imagine further that we observe five or more diverse organisms in which the homolog of gene A in each organism is adjacent to the homolog of gene B in that organism. Observation of conserved chromosomal proximity across many genomes supports an inference that genes A and B have related functions, such as being in the same pathway. Similar techniques based on phylogenetic co-occurrence, and the existence of rare gene fusions, are also used.

The second class of methods is based on applying machine learning techniques to very large datasets of information under the assumption that proteins with certain functions tend to have certain physico-chemical properties. This type of method has been used to infer whether or not a protein is an enzyme and to infer which of the six top-level Enzyme Commission classes an enzyme belongs to, based on information such as the amino-acid composition, molecular weight, and pI of the protein.

The third class of methods is based on integrative technologies that bring together different sources of evidence, such as protein-protein interaction screens, functional genomics screens, pathway information, RNAi screens, and computational predictions, to piece together a global picture of the proteome of a given organism.

An annotation initiative

In light of the pivotal importance of functional annotation to the progress of biological science, and because of the lack of clearly targeted support for experimental annotation and limitations of the current annotation initiatives, it is recommended that a centralized annotation initiative be established in the United States. (A related initiative, the Biosapiens project, is already underway in Europe.) The new initiative proposed here would (a) require bioinformaticians, in collaboration with experimentalists and the database managers, to establish a prioritized list of genes for experimental study and (b) engage experimentalists to test those predictions in the laboratory. In addition, enzymes for which good biochemical assays exist, but for which no genes have been reported should also be targeted for study under the initiative [see Box 1]. The results would lead to a database of reliable, experimentally-derived annotations that would be used by the entire biological community. The establishment of a central database to support experimental annotation would alert biologists to the need for careful interpretation of genomic sequences, and would provide immediate practical assistance to genome annotators. By providing funds for small, distinct annotation projects, the initiative could support and highlight an important niche within the biological community.

The experimental assignment of function for a given gene product is often a discrete, achievable project that can be accomplished within a reasonable amount of time by new Ph.D. students, rotation students, or even good Honors undergraduate students. The key is for the student to work in an established laboratory where reagents, substrates for biochemical assays, and technical know-how are available. A focused effort to assign functions to unknown genes would provide many excellent opportunities for training of graduate-level students.

Under the initiative, a functional prediction and validation database would tie together catalogs of uncharacterized gene products and orphan proteins with functional information that has been acquired through experimental means. The database would allow bioinformaticians and experimentalists to post new predictions and experimental results along with detailed information about the methods and evidence used to derive them. The database would require the contributions of a reliable staff of professionals who are able to work with the community of annotators and maintain the quality required by such a valuable resource.

The initiative would also enable further improvements to and validation of bioinformatics algorithms. By facilitating comparisons of functional predictions generated using bioinformatics with the results of experiments designed to test those predictions, the predictive power of the computational tools available through the database could be continually evaluated and improved. Thus, as the database expands so will the power of the predictive methods.

An annotation initiative of this type would benefit all of the fields that rely upon genomics and informatics by providing detailed knowledge about gene products and generating hypotheses about their biological role. Biodefense, biotechnology, synthetic biological chemistry and drug development would all be helped and the functional data generated would be of key significance for systems biology and many of the post-genomic initiatives that study multi-gene systems.

Outlining the annotation problem

An annotation initiative to support and compile accurate functional annotations should be designed to focus on the most worthy and central annotation problems at hand. With countless genes of unknown function and many scores of functions with unknown genes, where should the work of functional annotation begin? Also, how should erroneous functional assignments in publicly available databases be brought to light and corrected?

The initial thrust of the initiative should be the annotation of prokaryotic genomes. Bacteria and archaea possess relatively small genomes that could be more easily interpreted in a reasonable amount of time than those of eukaryotes. Also, scores of such genomes have been sequenced and serve as a great resource for experimentalists. Finally, prokaryotic organisms are currently the main platform for synthetic biology, molecular engineering, and systems biology. Hence, there is an urgent need to compile a carefully curated parts-list for future research and biomedical and commercial applications.

Setting priorities for experimentation

A prioritization scheme is needed to determine which genes among the available prokaryotic genomes should first be targeted for experimentation. A number of distinct schemes can be put forth, but the most effective and realistic approach would combine many considerations. Ideally, the cost and length of the experimental study would be balanced against the probability of success and the overall utility of the studied gene product.

One obvious criterion could be the size and promiscuity of a family of proteins. For example, if experimentally determining the function of one member of a cluster of orthologous groups would provide an understanding about many of the other members, then that protein should be a priority for investigation.

Targets for experimental work could also be scored according to the reliability of the available functional evidence. In this scheme, those genes for which the data are missing or extremely vague would take highest priority, those genes with slightly more reliable data would come next, etc. Another priority is those gene products for which the information is likely to be reliable, but imprecise. These would include broad functional predictions, such as "a glycosyl transferase." It would require experimental work to determine the specific substrates and pathways that are relevant to such a protein.

The gene products of well-studied model organisms like Escherichia coli could be targeted for experimentation in an effort to achieve a complete understanding of one organism that could then be applied in other contexts. Another approach might be to focus on genes that participate in common functions, such as sporulation or DNA repair, or are members of a common metabolic pathway. A very successful example of such a study that focused on tRNA modification is outlined in Box 2. Along these same lines, genes for which a great deal is known (aside from function) could also be appropriate targets for experimentation.

Biological significance or impact on society can also be guides to establishing priorities. It may be desirable to target those genes that are associated with human pathogenesis, like the genes involved in host interactions, antibiotic resistance, infectivity, and virulence, in an effort to make an impact on public health and medicine. It may also be useful to target gene products that are of economic significance, such as enzymes for industrially or environmentally important processes.

Whether a formal prioritization scheme should be used to guide funding decisions for the initiative remains to be decided. On one hand, a guide to priorities could serve to focus the efforts of a funding program on the most promising targets. On the other hand, it may be advisable to leave the decisions in the hands of experimentalists, who would have to find

appropriate assays and who would be required to make a convincing case for their work to a review panel or a goal-oriented funding initiative.

Finding genes for enzymes with known functions

Prioritization is also needed in cases where a cellular function has been assigned to a given enzyme, but the genetic sequence of the protein is not known (see Box 1). It is advisable to place the highest priority on those enzymes which have already been isolated and purified. N-terminal sequencing of these proteins should enable easy identification for an organism with a completely sequenced genome. Outside of these instances, priority should be placed on identifying the genes associated with important gaps in metabolic pathways.

Another criterion for prioritizing enzymes could be the degree to which the function is distributed phylogenetically. Finally, emphasis could be placed on enzymes that act on important metabolites or upon drugs.

❉ *Finding Missing Genes by Comparative Genomics* ❉

To provide accurate and complete annotations of the available genomes, it is necessary to identify a gene or gene set for every cellular function. In one approach to this task, Valerie de Crecy-Lagard is tracking down the genes t hat are associated with tRNA modification using bioinformatics and common sense biology.

De Crecy-Lagard tracks down functions that haven't been attributed to a gene through a combination of pathway reconstruction, subsystem analysis, and consultation with experts in different areas of biochemistry. Once a missing function has been identified, she then applies several bioinformatics tools (available through public databases, including COG, String, PhydBac, KEGG, MetaCyc, and SGD Model) to track down genes that may produce that function. The phylogenetic occurrence of a given function can be a powerful device for finding the appropriate gene; if the function does not occur in organisms X, Y, and Z, for example, then candidate genes for that function might be absent in those species as well. Several databases enable phylogenetic queries that facilitate finding missing genes in this way.

Gene clustering is another important indicator. If other genes in the pathway have been found to cluster at a certain point on the genome, the missing gene might be found within or next to the cluster. Co-expression patterns can provide leads as can the detection of fusion proteins, shared regulatory sites, and protein-protein interactions. Mining databases for these types of biological clues enables the researcher to compile enough information to formulate hypotheses that can be tested in the laboratory.

Taken from a lecture presented by Valerie de Crecy-Lagard, University of Florida. Other lectures exemplifying how informatics can guide experimentation were presented by Eugene Kolker (Biatech, Bothell, Washington) and Andrei Osterman (The Burnham Institute, La Jolla, California)

It is important to note that genes identified by this phylogenetic approach will have homologs in the sequenced genomes and consequently gene-function relationships established by this method will complement the genome-based approach.

Incorrect functional assignments

The problem of incorrect functional information in publicly available databases poses a serious problem. Charles Darwin wrote, "False facts are highly injurious to the progress of science, for they often endure; long, but false views, if supported by some evidence, do little harm, for every one takes a salutary pleasure in proving their falseness." Incorrect annotations provided to the public are often seen as "false facts," results that can mislead researchers and lead to wasted time and resources. However, it is important that researchers realize that annotations are interpretation, and subject to error, and should therefore be seen as "false views," and open to challenge and/or verification. Incorrect functional assignments should be corrected and disagreements between predictions reconciled, but this is difficult given the size and archival nature of current databases. Creating a new, independent database, and initially populating it with only functionally characterized genes without transitive annotation will be of immediate benefit to researchers and genome annotators, providing a solid foundation for subsequent annotation. This database will form the core, to which will be added the functional assignments that will be the outcome of this initiative.

To avoid and correct the dissemination of inaccurate data, the annotation database should be designed to allow dynamic access and should contain means for reassessing previous assignments and incorporating new data and evidence as they become available. On the experimental side, more value should be placed on the validation or the disproof of function. Disproving a functional assignment should be publishable in the relevant literature if the correct assignment has been made or in the central database if no definitive assignment is provided.

A publicly available database of functional assignments that contains the results of rigorous experimental work and includes bioinformatic predictions with levels of certainty indicated would be extremely useful. It is essential that experimental assignments be clearly distinguished from computational predictions. Furthermore, functions inferred from high-throughput experimentation should be distinguished from more reliable low-throughput biochemical studies. This information would solve many of the problems derived from the use of incorrect assignments and would be invaluable to the genome centers seeking to annotate new genomes.

Database requirements

Many different kinds of data would be acquired, integrated, and maintained by this annotation initiative. These include (a) a current dataset of experimental and predicted annotations found in the sequence databases, (b) a new set of predicted annotations, which would drive the experimental aspect of the initiative, and (c) the annotations that emerge from this experimental effort. These annotations would need to be incorporated into the definitive set. These data could all be handled within a single comprehensive database from which any given set of data might be retrieved. Alternatively, two or more integrated databases might be considered. The design specifics of data handling should be left open to those researchers who will construct it. However, some important features to address in designing data handling include:

- The scope of the database(s),
- How genes are defined in the context of the database,
- The specific information required for an entry (which may include information from informatics studies and/or experimental data) and the organization of that information, and
- How the information will be curated and maintained.

Although most of the particulars of the data handling remain to be determined, certain questions of content and standards were discussed at the colloquium. These issues include the criteria for a designation of a "definitive annotation," ways to tackle verification problems when mistakes arise, the inclusion of functional RNA in the database, the hosting, administration and management of the database, and the predicted costs of the project.

Setting criteria for definitive annotations

Not all annotation evidence is equivalent, and the concept of a "correct" annotation is subjective to some extent. In assembling a set of reliable gene annotations, it will be necessary to set criteria that define the type and degree of evidence required for a "definitive" annotation and to decide what kinds of less definitive data should be included. One way to set the bar is to ask whether the annotation explains the known biological properties of the gene. Under some conditions, a general assessment of the biochemical activity of the gene may be sufficient for moving forward. While it would be ideal to meet the high standards of publications like the Enzyme Handbook, this is not always possible, even in cases where the gene product has been thoroughly

characterized. Any incremental information about a gene's product, even a negative result, is probably worth including in the data set.

High throughput methods available today are often suggestive of function, but on their own they do not provide validation. In cases where a gene product has been sufficiently characterized to merit inclusion in the database, clues about function from high throughput methods should be added.

It should be noted that among computational biologists seemingly congruent functional predictions can be made separately by more than one group. However, if the tools these groups used to arrive at the predictions are related to one another or if the methods depend on the same underlying calculations, then the predictions would be correlated and should not be represented as separate and independent.

The verification of functional predictions

A central feature of the proposed initiative is to engage experimentalists in testing functional predictions made by bioinformaticians. This is not to suggest that computational prediction and experimental validation are independent activities. On the contrary, they form a closed loop initiative where computational predictions are validated or refined through experimentation and fed back into the computational pipeline to generate new cycles of hypothesis generation and validation. To facilitate this type of collaboration, the data set of computational predictions needs to be presented to the experimentalist in a user-friendly manner that allows the experimentalist to assess potentially contradictory predictions and use his or her biochemical judgment to select targets for study.

In verifying the functions of known genes or the genes of known functions, the preferred methods will be strongly dependent upon the problem at hand. When strong predictions are available, then cloning of the gene and direct biochemical assay in the hands of an expert could quickly lead to an answer. In the case of genes of unknown functions, a number of methods may come into play in an attempt to refine predictions. High throughput assays of the activity of expressed unknown proteins against large arrays of single potential substrates could prove helpful. However, these high throughput methods may not reveal sufficient detailed functional information to produce a definitive annotation on their own. Similarly, microarray assays to identify the gene products that are expressed in concert with the unknown gene or under stress conditions that alter expression may provide clues for more detailed biochemical exploration, but would not provide definitive evidence of function.

In the case of known enzymatic functions for which no corresponding gene has been described, protein purification and sequencing are likely to be the best, most straightforward approach to identifying the gene. Once a potential relationship has been established, then a simple biochemical assay could prove the assignment. If a candidate gene were present in strains such as E. coli or Bacillus subtilis, for which complete knockout libraries exist, then the suspected function could be assayed directly in both the appropriate knockout strain and its wild type counterpart. However, cloning the suspected gene and assaying its product directly would still be advisable.

It may not always be possible to verify conclusively the substrate of a particular enzyme, since many enzymes show specificities that are either broad or seemingly imprecise. Moreover, many unknown gene products may have complex enzymatic functions or they may lack an enzymatic function altogether, complicating efforts to uncover their functions by direct experimentation. Consequently, the proposed database should include functional clues as well as the results of experimental characterizations.

Many proteins do not exhibit an enzymatic activity, but rather play a role that depends on protein-protein interactions. The chaperonins, which help other proteins fold, adaptor proteins, and some of the integral membrane proteins that play a structural role, are good examples of proteins that are not easily tested by direct biochemical assays. In these cases, more indirect assays will be needed, and genetic tests may be helpful. Hence, the annotation initiative must be open to a wide variety of data from biochemical and genetic experiments that will help to define gene function.

Functional RNAs

Since the initiative should include all genes, it must include genes that encode functioning RNAs, such as tRNAs, inhibitory RNAs, specialized RNAs such as tmRNA, and perhaps others that remain to be discovered. Because some of these RNAs lack broad conservation across species, there are unique difficulties in prediction for RNA encoding genes which would need to be addressed.

Management of the Database

Hosting and administration

The question of where the annotation database should be housed is of significant practical concern. The administrators of smaller, individual prediction databases may be reluctant to share predictions with a centralized database like that proposed in this document if they perceive that the central database is biased in some way. Thus, there may be merit in seeing the database hosted at the National Center for Biotechnology Information (NCBI), which has already expressed interest in contributing to this initiative.

Design and management

It is critical to involve not only bioinformaticians in the design and curation of the proposed database, but also the experimentalists who will use the database to set their research agenda. Proposals to design and realize the database should require the collaboration of both types of professionals. It would also be helpful if, as part of this initiative, bioinformatics tools were made more accessible to experimentalists to enable these professionals to make functional predictions and address those predictions in their own laboratories.

The quality of the user interface of the database will be important to both bioinformaticians and experimentalists; the database must be intuitive and easy to navigate. Moreover, the process by which predictions, evidence, and methods may be submitted must be simple for prospective contributors to use and must permit bulk submissions of thousands of predictions programmatically, as well as bulk removal and replacement of old predictions with improved predictions made by the same investigators.

It will be critical for experimental researchers to deposit the results of their efforts into the database as quickly as possible. Promptness is a difficult objective to enforce, however, and there may be problems in convincing researchers to treat their data as community property. We recommend that the initiative adopt a strict requirement that all experimental results must be deposited into the database at the time of publication, and that there must be database accession numbers associated with each publication. These publications must be submitted by researchers as a criterion for continued funding.

The establishment of an external advisory board for the database project is recommended. Curation of the database should be a community effort in which advice is solicited from experts in specific subjects. Detailed descriptions of the organization and operation of the database are not necessary at this stage; proposals should be encouraged to include innovative solutions to meet the general goals of the initiative. It is important that the database be a public resource that provides facilities for both internet access and full downloads (to enable analysis and data mining) freely to the community.

Coordination with Model Organism Databases

The database project must have tight interactions with the model organism databases that have been established for many important experimental organisms. The EcoCyc project serves as a model of this type of revision; the project constantly combs the E. coli experimental literature for newly identified gene functions and incorporates those results into EcoCyc. (Annotation data should also be transferred from EcoCyc to the annotation database of the initiative. Likewise, experimentally determined gene functions should be transferred from the annotation database to EcoCyc.)

Peer review of data submitted to the database

Peer review of data and other supporting materials submitted to the database is critical to ensuring the integrity of this resource. However, lengthy peer review processes can stymie progress by increasing the amount of time between the discovery and the wide availability of the findings. It is recommended that submitted data be made immediately available through the database and that these data be accompanied by a clear warning statement that they have not yet been peer reviewed. The eventual outcome of the peer review process could then be posted in the database as it becomes available.

Communication and collaboration among annotation researchers

Communication between the two types of researchers involved in gene annotation, namely bioinformaticians who make function predictions and the experimentalists who take those hypotheses into the lab and test them, is not always effective. The annotation initiative described in this document could serve to bring these groups closer together in the interest of producing authoritative gene annotations in a timely, cost-efficient

manner – particularly if joint funding opportunities exist. It could also serve to motivate the biochemists by giving them a sense of community with the genomics initiatives currently underway.

Linking predictions available in the database with the names and contact information of the bioinformaticians who make the predictions would bring the bioinformaticians and experimentalists closer together, would assist experimentalists in acquiring the information they need to carry out their work and would help to provide scientific credibility to the bioinformaticians. Courses and tutorials on bioinformatics for experimental scientists could be offered under the auspices of the initiative to help these researchers better understand prediction methods and how to interpret and use them effectively. Finally, the submission of collaborative proposals should be encouraged in order to bring the different types of expertise of these professionals together on the same project.

For the sake of experimentalists, and to make the best use of function predictions, it is important to provide as much information as possible on the sources of the function predictions presented in the database. This information should include, but not be limited to the sources of the predictions, the degree of confidence or quality values of the prediction, links to the raw data used to make predictions, the results of expression analyses, and any phylogenetic data that have been generated.

The expertise of bioinformaticians could also be made available to experimentalists through a "help desk" system sponsored by the database initiative.

Predicted Costs and Funding of the Project

It is important that funding be available for improvement of bioinformatics function-prediction methods, whether through this initiative, or through a separate initiative. Although significant progress has been made in the newly developed bioinformatics methods, this field is still an extremely active one, and we expect significant improvements will be made to these methods for a number of years. Moreover, the database of experimentally-verified annotations will provide a higher quality "gold-standard" set of functional annotations that will aid in further refinement of these methods.

The costs of experimental verification will vary quite widely, depending on the gene of interest and the functions being investigated. For instance, a student project in an established laboratory testing a strong prediction might take only a few weeks with a low incremental cost. Since this could easily be a major component of the initiative in its early stages, it will be crucial to develop funding mechanisms able to support such activities. However, development of higher throughput methods, for example to check ligand binding, could be more expensive. Projects like this could be funded through more conventional routes.

The typical funding approach, in which proposals are solicited from interested investigators, would be most appropriate for funding the work behind this project. Some modifications to this scheme may be appropriate, however, including encouraging the use of a wide range of funding, from large down to very small awards. Also, the normal process of reviewing applications could be modified by conducting the review process in a manner similar to the review of scholarly publications, instead of the usual review method, in which study sections are employed.

Another possibility would be to encourage scientists to submit proposals for supplements to existing grants that might permit funded researchers to assign students to work on specific genes within the laboratory's area of expertise. This might help to elicit the involvement of the experimentalists who will provide the data at the heart of the project by providing specifically targeted funding.

Funding for some of the work described in this initiative might be provided from sources that are already available. For example, a request for proposals has been issued under the auspices of the PSI that provides an opportunity for bioinformaticians and experimentalists to work with a set of proteins that are already earmarked for funding because their structures are known or in the process of being determined.

Finally, it seems desirable to launch a small pilot project to test the efficacy of the proposed initiative. If early success can be achieved at modest cost, then there would be some impetus to raise the stakes and launch a major attack on the problem.

Summary of Recommendations

- Considering the importance of genome annotation to full exploitation of sequence information, progress in experimental annotation has been slow, largely due to a lack of available funding for experimental annotation approaches. It is recommended that an annotation initiative be undertaken to cat-

alyze and coordinate funding for experimental approaches to genome annotation.

- Given the current lack of a reliable source of functional annotation data, a central data resource should be established. It should incorporate:
  - a database of peer-reviewed, experimentally verified gene annotations,
  - a catalog of the genes that have yet to be annotated, which users can sort by gene family, species, priority, etc.,
  - a catalog of the functions for which a gene remains to be found, and
  - all available experimental information relevant to function.
- Priority in designating funding for annotation through the initiative should be placed first on prokaryotic genomes. In selecting among prokaryotic genes, emphasis should be placed on those gene products that belong to large protein families, since knowledge of these genes is most likely to impart an understanding of the biology of many diverse systems. In this way, a small investment of experimental work and funding can lead to big rewards in understanding many species.
- Progress in functional annotation could be enhanced to some extent by developing mechanisms and information systems that encourage collaboration between bioinformaticians and experimentalists. This would allow experimental scientists to quickly access and test the predictions made using informatics tools, while providing bioinformaticians access to experts on the function of particular genes and enzyme systems.
- In efforts to identify the gene encoding a given product, funding priority should be given to those proteins or RNAs that have been purified and characterized, as simple sequencing would then lead to identification of the gene.
- The design of the database should remain open to the input of researchers who choose to submit proposals for constructing and maintaining it. However, it must support the maintenance and update of working hypotheses about gene function.
- To avoid conflicts with the interests of potential contributors, the database should be hosted by an unbiased organization without a vested interest in the content of the data.
- The collaborative contributions of experimentalists and bioinformaticians should be encouraged through the requests for applications announced by the database coordinators.
- A variety of funding types will be needed, including small awards to support students who might work in an experienced investigator's laboratory for a short period of time.
- The resources produced by the initiative should be made public and be freely available to the global scientific community. ◆

This report was designed by
Pensaré Design Group
www.pensaredesign.com