BIOINFORMATICS IN THE 21st CENTURY

A Report to the
Research Resources and Infrastructure Working Group
Subcommittee on Biotechnology
National Science and Technology Council
White House Office of Science and Technology Policy

Bioinformatics Workshop
February 3-4, 1998
Krasnow Institute for Advanced Study
George Mason University
Fairfax, Virginia

Prepared by:
Tracor Systems Technologies, Inc.
Rockville, MD
Under contract with
Krasnow Institute for Advanced Studies
George Mason University, Fairfax, VA

About the National Science and Technology Council

President Clinton established the National Science and Technology Council (NSTC) by Executive Order on November 23, 1993. This cabinet-level council is the principal means for the President to coordinate science, space and technology policies across the Federal Government. NSTC acts as a "virtual" agency for science and technology (S&T) to coordinate the diverse parts of the Federal research and development (R&D) enterprise. The NSTC is chaired by the President. Membership consists of the Vice President, Assistant to the President for Science and Technology, Cabinet Secretaries and Agency Heads with significant S&T

responsibilities, and other White House officials.

An important objective of the NSTC is the establishment of clear national goals for Federal S&T investments in areas ranging from information technologies and health research, to improving

transportation systems and strengthening fundamental research. The Council prepares R&D strategies that are coordinated across Federal agencies to form an investment package that is aimed at accomplishing multiple national goals.

To obtain additional information regarding the NSTC, contact the NSTC Executive Secretariat at 202-456-6102.

About the Office of Science and Technology Policy

The Office of Science and Technology Policy (OSTP) was established by the National Science and Technology Policy, Organization and Priorities Act of 1976. OSTP’s responsibilities include advising the President in policy formulation and budget development on all questions in which S&T are important elements; articulating the President's S&T policies and programs, and fostering strong partnerships among Federal, State and local governments, and the scientific communities in industry and academe.

To obtain additional information regarding the OSTP, contact the OSTP Administrative Office at 202-395-7347

DISCLAIMER: This document reflects the proceedings of a workshop organized by the Research Resources and Infrastructure Working Group of the National Science and Technology Council, Committee on Science, Subcommittee on Biotechnology. This workshop, which included a panel of more than 20 experts in bioinformatics and related fields, was organized to provide advice to the Subcommittee on Biotechnology. This document is not intended to reflect government policy.

Table of Contents

Introduction

Technological advances and ubiquity of the Internet offer unprecedented opportunities for scientists to gain access to, share, and analyze critical data and information stored in databases. These vast stores of information have a rich potential to expedite scientific

discovery and prevent costly duplication of experiments. Yet there is a price to pay for this wealth of information: the scientific community now faces the daunting challenge of storing, retrieving, analyzing, and rendering useful these rapidly growing data sets. For funding agencies, the challenge is to meet the ever-changing needs of the research community by making sound investments in research, infrastructure, and training related to bioinformatics.

At the behest of the Research Resources and Infrastructure Working Group, established by the National Science and Technology Council's Subcommittee on Biotechnology, a panel of more than 20 experts in bioinformatics and related fields convened to discuss critical issues surrounding bioinformatics, identify problems and challenges, and offer potential solutions. This report summarizes the key issues raised at the Workshop on Bioinformatics, held February 3-4, 1998, at the Krasnow Institute for Advanced Study, George Mason University, Fairfax, Virginia.

Executive Summary

Advances in laboratory tools and technologies now allow scientists to collect unprecedented amounts of data; fortunately, advances in computational sciences and communication technologies have kept pace, allowing biologists to share data across disciplines and address increasingly complex problems. Most scientists now have desktop computers that have more raw processing power than the first CRAY supercomputers.

Although new computational tools and information technologies are opening new vistas for biology and medicine, funding agencies now face the formidable task of identifying those projects and research areas that will most significantly benefit the scientific community and enable scientific advancement well into the next millennium. At the Workshop on Bioinformatics, held February 3-4, 1998, at the Krasnow Institute for Advanced Study, George Mason University, more than 20 experts in bioinformatics and related fields engaged in wide-ranging discussions related to the future of bioinformatics, obstacles that must be overcome, and possible actions that might be taken by the federal government or other entities.

While there was much disagreement about certain problems and solutions, the following opinions were echoed repeatedly during the two-day discussion and seem to merit particular attention:

- Achievements in biology and medicine in the 21st century will require a substantial investment in bioinformatics.
- Bioinformatics projects must be driven by user needs.
- Alternative funding and review mechanisms are needed for support of bioinformatics infrastructure and enabling technologies.
- Mechanisms are needed for interfacing funding agencies with professional societies to help set priorities for supporting bioinformatics research and infrastructure.

Recommendations

Workshop participants suggested that the federal government target three broad areas for support: Basic research into bioinformatics and its applications, bioinformatics infrastructure (e.g., databases) and other user resources, and education and training in bioinformatics.

In addition, the following observations and recommendations were made:

1. The federal government should invest research dollars for the biological sciences in bioinformatics. Bioinformatics will be indispensable for the advancement of science in the 21st century.

2. An interface must be created between funding agencies and professional societies to help set priorities for bioinformatics research and infrastructure. Professional societies

and other organizations that represent user interests can help the federal government set priorities and get feedback related to:

- The setting of standards (both technical and nontechnical) for database interconnectivity.
- Needs and priorities for databases, analytical tools, and other enabling technologies.
- Assessment of current databases and related infrastructure.

3. There is a national need for training and education in bioinformatics. In particular, educational efforts should target end-users (biologists and other domain scientists); bioinformaticists, who must create user-friendly tools; and senior faculty, who must take a leadership role and educate newcomers to the field.

4. Database interconnectivity and the setting of standards will be critical for examining complex, interdisciplinary biological problems. Having made substantial investments in the collection, storage, and analysis of biological data, it would behoove the federal government to help ensure that these data are put to the best possible use and are usefully interconnected to other critical datasets. The following possible actions were proposed:

- An interagency committee/task force should be created to address problems related to technical/computational standards and interconnectivity.
- Agency representatives should participate more fully in regional, national, and international standards-setting bodies and should demand that grantees adhere to standards where they exist.
- Professional societies should help set priorities for setting standards

5. Funding agencies should explore alternative granting mechanisms and review processes for databases and other bioinformatics infrastructure. The following comments and activities were proposed:

- To responsibly fund bioinformatics infrastructure, the federal government needs stronger links to the user community.
- The following two-step review process might be considered: 1. Identify priorities and needs by interfacing with the scientific community. 2. Fund only those proposals that address/meet those needs.
- Proposals for bioinformatics research and infrastructure are often declined because they are not fully appreciated by review committees. True "peer review," by individuals familiar with bioinformatics, is needed.
- Ongoing database/infrastructure projects should be regularly reviewed and assessed by the user

community. Funding mechanisms for infrastructure should go beyond the fixed-term, fixed-amount award.

6. New Intellectual Property Laws may destroy free flow of scientific data and information. The federal government should protect access to biological data for use in education, research, and other public interest purposes. The scientific community should remain vigilant and become aware of proposed legislation.

Role of Bioinformatics in the Biological/Biomedical Sciences

Bioinformatics will be at the core of biology in the 21st century. In fields ranging from structural biology to genomics to biomedical imaging, ready access to data and analytical tools are fundamentally changing the way investigators in the life sciences conduct research and approach problems. Complex, computationally intensive biological problems are now being addressed and promise to significantly advance our understanding of biology and medicine. No biological discipline will be unaffected by these technological breakthroughs.

Reliance on bioinformatics and related computational tools is perhaps most evident in the field of genomics, where sequencing data and related datasets are growing at an exponential rate, far outstripping efforts to manage and analyze these data. Every 10 weeks, more sequence data is deposited in GenBank than went into GenBank in the past 10 years, one workshop participant commented.

Declaring that the golden age of genomics has arrived, Dr. Anthony R. Kerlavage, director of bioinformatics at The Institute for Genomic Research (TIGR), noted that a dozen genomes, representing about 20,000 genes, have been completely sequenced to date, and 50 additional genomes are expected to be completed within the next three years. Lagging behind, however, are efforts to identify the role and function of these genes and their protein products. As genome researchers gradually shift their focus from gene structure to function, the challenge to bioinformaticists is to make such information accessible, understandable, and valuable to the scientific community.

Dr. Kerlavage described several lines of investigation related to microbial genomics now underway at TIGR. One involves classifying genes by role and function, which is proving useful in comparative genomic studies; another attempts to minimize the genome, or knock out genes until the organism can no longer survive. He identified an urgent need for managing and interpreting the growing amount of data generated by new chip-based technologies (microarrays). These powerful techniques, developed less than five years ago, allow high-speed, high-capacity analysis of gene expression. Dr. Kerlavage expects microarray techniques to generate a wealth of information that must be standardized, stored, and made available in the near future. With proper analysis, such data can help narrow the focus and prevent costly duplication of biological experiments. Millions of dollars and experimentation time might be saved if scientists have ready access to data that have already been compiled and archived, Dr. Kerlavage said.

Conclusions and Recommendations

Workshop participants agreed that bioinformatics will be critical not only to the future of genomics but to most areas of biological and biomedical research. Participants identified three general areas that

require support:

- Basic research into bioinformatics and its applications;
- Bioinformatics infrastructure and other user resources;
- Education and training in bioinformatics.

databases: Establishment, Maintenance, Scientific Review, and Support

As computational tools and communication technologies have steadily improved during the past few decades, biologists have become increasingly dependent on having ready access to shared data and analytical tools to enhance their research. Advances in bioinformatics allow biologists to rapidly collect and analyze enormous amounts of data, much of which are now stored in databanks or databases. By collecting critical information in user-friendly databases, made available to the scientific community, the federal government's investments in data acquisition and storage can be returned many times over. But this is much easier said than done. Dr. David Matthews, curator of USDA's GrainGenes database at Cornell University, described some of the choices and challenges faced by those who create, manage, and maintain large databases for the biological community.

Establishment. User needs should drive the creation and development of biological databases, workshop participants agreed. The scientific community should have input not only into setting priorities for new databases but also into performance reviews of existing databases and related projects.

Maintenance. Databases rarely allow for the ambiguities inherent in the life sciences. While physical data are specifically (phenomonologically) defined and can be readily codified for database input, biological data are often variable and open to interpretation. To help resolve ambiguities and ensure data quality, some biological databases have curators who oversee the selection and inclusion of data. However, curation can be costly and time-consuming. And because many of today's databases are growing exponentially, it may be impractical to have a central curator or peer review of data, many discussants noted.

When data content is relatively straightforward, as in GenBank, investigators can deposit their data directly without curation; other databases that are more complex might require contributors to complete a form that arranges data in a uniform format. However, at a certain point it becomes inefficient and impractical to train scientists to organize their own data to meet the needs of the database, commented Dr. David Matthews; he suggested that it might be useful to create mechanisms to support scientific experts in preparing data for contribution to some databases.

Although voluntary contribution of data is a critical component of many large databases, lack of incentive to contribute is a persistent problem, several participants agreed. Sequence databases rarely face this dilemma, since scientists must deposit sequence data in appropriate databanks as a condition of publication in scientific journals. Without voluntary contributions of data, central database-building facilities often must extract data from the literature or other sources, significantly driving up the cost and effort of database maintenance. As a result, critical data are often lost to the research community, or unavailable in a useful form, said Dr. Lois Blaine, director of bioinformatics at the American Type Culture Collection. Perhaps incentives should be developed for depositing data in crucial databases, she proposed.

Funding Mechanisms and Review. For a database to succeed, it must be assured of stable and continuous financial support, said Dr. Matthews. Financial stability boosts user confidence in the

database and encourages voluntary submission of data. In some respects databases are comparable to repositories of living materials: both may be absolutely essential for biological research but are a perpetual struggle to maintain and fund, commented Dr. Kenneth Paigen, senior staff scientist at The Jackson Laboratory. Although they may have little commercial value, some databases are so critical to the research community that they require a strong financial commitment from federal agencies.

Note: Issues related to funding mechanisms for databases and related infrastructure were discussed in greater detail in the session titled "Ensuring Access to Information: Federal Support for Infrastructure" led by Dr. Robert Robbins.

Conclusions and Recommendations

To improve the usefulness of databases and related resources, workshop participants proposed several plans and activities, including those listed below. Although no real consensus was reached on these suggestions, they seemed to generate sufficient interest among workshop participants for inclusion in this report.

Software Development. Many discussants argued that publicly funded databases have a strong software development component, which is absolutely critical to the success

of large database programs. Dr. Matthews proposed creation of a mechanism for funding development of software that might be generally useful to multiple database programs.

Analysis Nodes ("One-Stop Shopping"). Dr. Kerlavage identified a need for establishing analysis "nodes," where biologists could access a whole suite of analytical tools. Because there is currently no common interface for these critical resources, researchers must search the Internet to find critical tools at disparate sites, and many scientists are unaware of or unable to locate available resources. A few Web sites partially fill this niche. For example, ANGIS (Australian National Genomic Information Service) has valuable links to software, databases, and other bioinformatics resources, and the Lister Hill National Center for Biotechnology Information provides access to BLAST (Basic Local Alignment Search Tool), Entrez, and other tools. However, a larger, more comprehensive area for "one-stop shopping" is needed, argued Dr. Kerlavage, who suggested that creation of such a site might be a trans-agency function.

Focused Problem Solving. Dr. Kerlavage also suggested that the government fund the focused development of new tools, preferably under contract, to solve specific problems related to databases.

Standards and Interconnectivity

As biology becomes an increasingly collaborative undertaking, advances in computer technologies and bioinformatics are creating new possibilities for collaboration and discovery within and across scientific disciplines. However, most existing databases (and their associated biological disciplines) have grown up independently, with tremendous variability in nomenclature use, data content, and analytical tools. If biological data are to be effectively exchanged, integrated, and analyzed, the need for standardization must be addressed. Setting standards will be a formidable and expensive task, workshop participants noted. But with databases growing at an exponential rate, it might be prudent to begin to address these problems now, before they become even more unmanageable and costly to solve.

Dr. Lois Blaine, director of bioinformatics at the American Type Culture Collection, identified two types of standards that are critical for interoperability of databases: Technical/computational standards (e.g., hardware, software) and semantic/terminology standards (e.g., nomenclature, concepts). An additional consideration, which lies somewhere between the two, is data models.

Several international and interdisciplinary bodies have examined the importance of such standards for data exchange. For instance, the Committee on Data for Science and Technology (CODATA), founded in 1966 by the International Council of Scientific Unions, is an interdisciplinary committee that works to improve the compilation, evaluation, and dissemination of data on an international level. As part of its mission, CODATA also explores the need for standards and other options that might facilitate data exchange.

Technical/Computational Standards. To make the most of the federal government's substantial investments in the creation and maintenance of biological databases, the government should have a stake in ensuring interoperability and establishment of computational standards, suggested Dr. Paigen. Software and other tools can be designed to facilitate interconnectivity between databases; for example, CORBA is a new tool for facilitating information exchange between databases. But problems with technical/computational interconnectivity occur at many levels, and many obstacles remain.

Practically speaking, database overseers currently have little incentive to improve technical interconnectivity, many participants said. Such an endeavor would likely require extensive revision of data, software development, and the setting of standardsCall of which are costly, time-consuming, and not funded under current grants. Requirements to rework data to conform to certain inter-database standards would likely require three times as much funding, Dr. Matthews commented. He also noted that technical interconnectivity among databases could be difficult to maintain, in part because hardware and software are evolving rapidly, and making transitions to new technologies would require considerable cooperation among database owners.

Relating an opinion once expressed by Dr. Peter Karp, Dr. Robert Robbins, vice president for information technology at the Fred Hutchinson Cancer Research Center, said that database connectivity or referential integrity may not entirely depend on setting standards for nomenclature or software; rather, subtler and easier-to-solve problems might first be considered. For instance, some connectivity problems are caused when the key structure of a database is redesigned without notifying interconnected databases that point to those primary keys. This type of problem occurs frequently and is solvable.

What the government can usefully do is fund workshops to consider current barriers and practical options for technical interconnectivity, some participants suggested. Perhaps such workshops could identify which computational "layers" should be stabilized or standardized.

Semantic/Terminology Standards. Problems of semantics and terminology in the biological sciences are more intractable than technical/computational problems, workshop participants agreed. Achieving agreement on nomenclature within a discipline is extremely difficult, let alone across disciplines, Dr. Blaine commented.

Dr. Blaine identified four characteristics that are important for the development of nomenclature standards: They must be developed by experts; be accepted at an international level; have long-term funding/support; and be accessible and practical to use.

Within biological disciplines, names of organisms are often dictated by international code. For instance, in virology an international code is regulated by the International Committee on the Taxonomy of

Viruses, who meet periodically to discuss new data and adjust classifications as necessary. The group represents a public view that can be used by database producers, said Dr. Blaine, although there may be disagreement and the code is amenable to change. Other organizations work to standardize nomenclature for proteins, genes, and other biological entities. But even when standards exist they are not always used, Dr. Blaine noted, in part because standards are not effectively publicized and scientists may be unaware that standards exist.

Traditional semantic/conceptual barriers between scientific disciplines create an even greater hurdle for database interconnectivity. A common problem is that different vocabulary may be applied to similar or identical entities, and scientists themselves may not recognize that the objects are related.

Data Models. Standards may also need to be set at the level of database model. Relational models were once considered essential for interoperability, said Dr. Blaine, but now object-oriented data models are becoming more prevalent, and may be more suitable for biological data.

Reaching agreement about data models can be tremendously difficult, added Dr. Matthews, who described recent unsuccessful efforts for plant genome databases to define a common model. Data models are in a difficult position, commented another participant; the models should be part of the infrastructure, but they depend on the development of standards in semantics.

Conclusions and Recommendations

It would be a sound fiscal investment for the federal government to help ensure interoperability of databases, including the creation of standards. Although expensive, database interconnectivity will be critical to the future of biology and medicine. If standards are set later rather than sooner, establishing interconnectivity will be substantially more costly, commented Dr. Sylvia Spengler, principal investigator of the Human Genome Program at Lawrence Berkeley National Laboratory. Whatever mechanisms are chosen for achieving standardization, they must be flexible enough to adapt to unexpected needs or advances in science and technology, cautioned some discussants.

Most participants agreed that funding agencies should try to address the problems of technical/computational connectivity, perhaps via an interagency committee, since such standards may be more readily achieved than those of semantics/terminology. Although discussants acknowledged that creation of nomenclature and conceptual standards is essential, there was some disagreement as to whether the problem is beyond the scope of this workshop. Many participants recommended that the issue be referred to professional societies and database producers.

Workshop participants called for development of mechanisms that allow funding agencies to interface with professional societies, universities, and industry to identify problems and priorities for establishing and maintaining database interconnectivity.

Ensuring Access to Information: Federal Support for Infrastructure

For public-sector research to remain significant and vital into the 21st century, funding agencies must ensure that the research community has access to appropriate information resources, said Dr. Robert J. Robbins, vice president for information technology at the Fred Hutchinson Cancer Research Center. Without access to large collections of data, it will be impossible to conduct quality research and address complex biomedical issues. It is imperative that federal agencies recognize this and take up the

challenge of ensuring access, Dr. Robbins said.

Dr. Robbins proposed that federal support for infrastructure, especially information infrastructure, be supported by new funding mechanisms. The funding methods typically applied to investigator-initiated research are too slow for meeting information infrastructure needs, which are changing rapidly as technologies continue to improve. Federal agencies should consider developing faster, more efficient mechanisms for supporting large-scale public information resources, possibly even to the extent of shifting from supply-side to demand-side funding, he argued.

Perhaps even more important to consider, Dr. Robbins said, is that traditional proposal review processes designed primarily for investigator-initiated research can inadvertently lead to inferior infrastructure. Typical research proposals are judged on the merits of the proposal, and funding agencies have little input into the proposed research. But when funding infrastructure, Dr. Robbins suggested that agencies are buying access to central resources on behalf of the research community; therefore, agencies are obliged to obtain the best possible resource for the community. In such cases, project officers may need to guide applicants toward improving proposed programs and products.

Dr. Robbins also asserted that typical grant review processes prevent fulfillment of unique visions described in proposals for information infrastructure. When information resources must answer to a review committee, whose members may have limited knowledge of bioinformatics, the resource may be asked to broaden its efforts or eliminate novel components of the project, which may ultimately weaken the product. Dr. Robbins pointed to the success of National Center for Biological Information (NCBI), which need not answer to a typical review committee, to illustrate his point. Dr. Robbins proposed that NCBI succeeds because of its entrepreneurial vision and its successful relationship with consumers. In contrast, resource users carry little weight under traditional grant mechanisms, Dr. Robbins argued.

Many discussants noted that the government often expects databases to become commercially viable. But if the government is willing to support database projects but expects market forces to eventually take over, said Dr. Robbins, the project must be funded to nurture establishment of market forces. The government must also make adequate provisions for public interest access by the research, education, and library communities whenever public data are privatized.

Workshop participants commented that the World Wide Web, which seemed to appear overnight, was driven by market forces and was not funded by the federal government. The technological conductivity that the Web provides has forced information providers to comply with this standard and, in some cases, work together to interconnect their products.

Conclusions and Recommendations

Funding agencies should develop new mechanisms for funding and reviewing the usefulness of databases and other bioinformatics projects, workshop participants agreed. Participants also recognized a problem with the current system for proposal review, which is rarely performed by true peers who understand bioinformatics.

Some participants suggested giving greater power to project officers, who should become active in meeting and understanding the needs of the user community. However, other discussants cautioned, in these times of fiscal constraint and reduced staffing levels in the federal government, it may be impractical for project officers to play such a prominent role in the many grants they oversee.

The general consensus among participants was that the scientific communityCin particular, the

contingent that uses a particular resourceCchave input into setting priorities for funding of new bioinformatics projects and into the review of ongoing projects. This could be accomplished via workshops or meetings of professional societies.

Dr. Robbins proposed a two-step review process, in which funding agencies first establish priorities (with input from the scientific community) and then select "vendors," or proposals, that can best meet these prioritized needs.

Many participants agreed that user feedback during the grant period could help ensure performance. Perhaps continuation of funding might be contingent on positive user reviews. However, discussants also cautioned that is not effective to simply reduce the funding for a less-than-optimal database program; this leads to an even more inferior product.

Intellectual Property

Mr. Paul F. Uhlir, associate director for special projects at the National Research Council and director of the U.S. National Committee for CODATA (Committee on Data for Science and Technology), informed workshop participants of recent and proposed changes to international and domestic laws that affect copyright and protection of intellectual property as applied to digital information and databases. These laws may have negative consequences for the full and open exchange of scientific data, which is a hallmark of the research enterprise.

The rapid proliferation of digital data in recent years has raised concerns over protection of intellectual property, since digitized information can be readily copied and broadly distributed. In the United States, scientific and other works of authorship have long been protected primarily by copyright, which allows for "fair use" of protected information by scientists, educators, and others working for the public good. However, current U.S. copyright law does not extend to databases that are mere factual compilations and are not Aoriginal and creative works of authorship.@

Mr. Uhlir described a new law adopted by the European Union (E.U.) in March 1996, which creates unprecedented protection for database content and places severe restrictions on the concept of fair use and the conditions under which databases can be accessed in the networked environment. The new law, the European Directive on Databases, will have a chilling effect on the principle of open exchange of both public and private scientific data, Mr. Uhlir said. The effect will be most keenly felt in internationally oriented research on such topics as environmental change and biodiversity, or in data-intensive research that integrates data from multiple sources. The law will also increase the overall cost of conducting research, since commercial fees may be charged for access to data, and increased administrative costs will be needed to enforce legal restrictions on data use. Perhaps most disturbing, Mr. Uhlir continued, is the potential for large-scale, but difficult-to-measure, opportunity costs, which are likely to arise if simple exchanges of data and access to individual databases become legally threatening or prohibitively expensive.

In December 1996, at a diplomatic conference sponsored by the World Intellectual Property Organization, participants rejected a draft international Treaty on Intellectual Property in Respect of Databases that had been proposed by the E.U. and the United States. Based on the European Database Directive model, such a treaty would have protected the contents of databases and prohibited unauthorized uses of "substantial portions" of a database, as defined by the database owner. This would have created an entirely new international legal norm for database protection, requiring the United States

and other countries outside the E.U. to amend their own intellectual property laws. Although not designed to protect individual pieces of data (i.e., facts), in practice such a treaty would in essence restrict access to facts and most likely require scientists and educators to pay commercial prices for access to such bits of information. Some discussants noted that scientists regularly sign away copyright to their own data and other material when signing contracts for publication in scientific journals. This can severely limit reuse of this information in databases and other digital information products and services. The scientific community should be made aware of this problem and possibly form a united front to keep publishers from acquiring unwarranted and excessive rights to their intellectual property.

Conclusions and Recommendations

Mr. Uhlir suggested that workshop participants consider the legal aspects of federally funded bioinformatics projects. He recommended that participants reaffirm the "public good" aspects of all basic research data created under federal grants, and oppose restrictions on the open flow of scientific data. On an international level, he recommended encouraging scientists in Europe and elsewhere to continue with open exchange of data and cooperative research, and resist temptations to adopt the restrictive provisions that are now available there, even for public government institutions.

Finally, Mr. Uhlir suggested that the government exercise caution whenever privatizing certain data management and dissemination functions and protect access to such data for research, education, and other public interest uses. Such privatization should always be done on a nonexclusive basis.

Training

As reliance on databases and computational techniques continues to pervade the life sciences, the demand for well-trained professionals with expertise in both biology and information technologies will necessarily climb as well. However, the field of bioinformatics is trapped in a kind of netherworld, vitally important to the advancement of science yet unrecognized as a discrete discipline by many funding agencies and universities. As a result, proposals for bioinformatics-related research are often dismissed during the peer review process, and surprisingly few universities offer programs in bioinformatics.

Conclusions and Recommendations

The federal government should invest in bioinformatics training, discussants recommended, because such skills will be indispensable to the future of biological research. Educational programs should target three categories of individuals: end users, or biologists, who need training in using bioinformatics as a tool to enhance their research; master's level students, whose education will enable them to develop the tools and technologies needed for applied bioinformatics; and predoctoral students, who will receive formal training in both a computational science and a biological science and ultimately become leaders and educators in this emerging discipline. In addition, some discussants recommended creating summer bioinformatics courses for undergraduates enrolled in applied mathematics, computer science, or related programs.

Educational funding for bioinformatics should not lie solely in the hands of the federal government, workshop participants agreed. Industry also depends on having a qualified bioinformatics workforce and should be called upon to support training programs and fellowships. Some discussants expect individuals with Master's degrees in bioinformatics to be most marketable to industry, whereas doctoral

training will be required for academia. Dr. Harold Morowitz, director of the Krasnow Institute for Advanced Studies at George Mason University, commented that students in the university's bioinformatics program are rapidly employed by pharmaceutical companies, often before obtaining their degrees.

Training in bioinformatics will require a unique mode of cross-disciplinary education. Predoctoral instruction should address representational issues, fostering the ability to mathematically express a biological issue or topic. Workshop participants also identified a need for individuals trained as database or tool builders, who have a solid background in software engineering and some knowledge of biology.

Efforts should be made to legitimize bioinformatics as a profession and a field of study, discussants said. To stimulate "respect" and support for bioinformatics, Mr. Uhlir recommended creating annual awards that recognize "excellence in research" or "significant advances" in bioinformatics. Such awards might be sponsored by professional societies, foundations, or corporations, rather than the government, to honor both students and researchers.

Appendices

Bioinformatics Workshop

February 3-4, 1998

Krasnow Institute, George Mason University

Meeting Agenda

Tuesday, February 3, 1998

8:30 a.m. Opening Remarks
Harold Morowitz, Krasnow Institute Charge to the Group

Judith L. Vaitukaitis, National Center for Research Resources, NIH

Morning Sessions:

I. Role of Bioinformatics in the Biological/Biomedical Sciences

Anthony R. Kerlavage, The Institute for Genomic Research

II. Databases: Establishment, Maintenance, Scientific Review, and Support

David Matthews, Cornell University

III. Standards and Interconnectivity

Lois Blaine, American Type Culture Collection

Afternoon Sessions:

IV. Ensuring Access to Information: Federal Support for Infrastructure

Robert Robbins, Fred Hutchinson Cancer Research Center

V. Intellectual Property

Paul F. Uhlir, National Research Council

VI. Training

Harold Morowitz, Krasnow Institute

Wednesday, February 4, 1998

8:30 a.m. Preparing the Report and Options

Harold Morowitz, Krasnow Institute

Synopsis and Discussion of Plan

John Wooley, U.S. Department of Energy

Bioinformatics Workshop Participants

Harold Morowitz (Chair)

Director, Krasnow Institute

George Mason University

Fairfax, VA

Peter Arzberger

Associate Director

Center for Advanced Computational Science and Engineering

University of California, San Diego

David Benton

SmithKline Beecham Pharmaceuticals

King of Prussia, PA

Lois Blaine

Director, Bioinformatics Division

American Type Culture Collection

Rockville, MD

Douglas Brutlag

Professor of Biochemistry and Medicine

Stanford University School of Medicine

Stanford, CA

Daniel W. Drell

Biologist, Human Genome Program

Office of Biological and Environmental Research

U.S. Department of Energy

Germantown, VA

Paul Gilna

Program Manager

Division of Biological Infrastructure

National Science Foundation

Arlington, VA

Anthony R. Kerlavage

Director of Bioinformatics

The Institute for Genomic Research

Rockville, MD

David Matthews

Curator, GrainGenes Database

Department of Plant Breeding and Biometry

Cornell University

Ithaca, NY

Kenneth Paigen

Senior Staff Scientist

The Jackson Laboratory

Bar Harbor, ME

Robert J. Robbins

Vice President for Information Technology

Fred Hutchinson Cancer Research Center

Seattle, WA

Henry L. Shands

Acting Assistant Administrator

International Research Programs

Agricultural Research Service

U.S. Department of Agriculture

Beltsville, MD

Sylvia Spengler

Principal Investigator

Human Genome Program

Lawrence Berkeley National Laboratory

Berkeley, CA

Marvin Stodolsky

Molecular Biologist, Human Genome Program

Office of Biological and Environmental Research

U.S. Department of Energy

Germantown, MD

Paul F. Uhlir

Associate Director for Special Projects

Commission on Physical Sciences, Mathematics, and Applications

National Research Council

Washington, DC

Judith L. Vaitukaitis

Director

National Center for Research Resources

National Institutes of Health

Bethesda, MD

Keith B. Ward

Biomolecular and Biosystems Division

Office of Naval Research

Arlington, VA

John Wooley

Associate Director, Office of Energy Research

U.S. Department of Energy

Germantown, MD

Observers: (Due to the informal and interactive nature of the conference, observers were able to participate in the discussions and contribute to the conclusions).

Barbara T. Bauldock

Biological Resources Division

U.S. Geological Survey

Reston, VA

James H. Beach

National Biological Information Infrastructure

Biological Resources Division

U.S. Geological Survey

Reston, VA

Gladys Cotter

Assistant Chief Biologist, Informatics

Biological Resources Division

U.S. Geological Survey

Reston, VA

Richard DuBois

Health Scientist Administrator, Biomedical Technology Area

National Center for Research Resources

National Institutes of Health

Bethesda, MD

Maryanna Henkart

Director, Division of Molecular and Cellular Biosciences

National Science Foundation

Arlington, VA

Dov Jaron

Director, Biomedical Technology Area

National Center for Research Resources

National Institutes of Health

Bethesda, MD

George S. Michaels

Associate Professor of Computational Biology

Institute for Computational Sciences and Informatics

George Mason University

Fairfax, VA

Louise Ramm

Deputy Director

National Center for Research Resources

National Institutes of Health

Bethesda, MD

Deborah Sheely

Assistant Program Director, Plant Systems

National Research Iniative Competitive Grants Program

U.S. Department of Agriculture

Washington, DC

Anna Tsao

Defense Sciences Office

Defense Advanced Research Projects Agency

Arlington, VA

Research Resources and Infrastructure Working Group

of the

Subcommittee on Biotechnology

Dr. James H. Beach

National Biological Information

infrastructure Program

USGS Biological Resources Division

300 National Center

12201 Sunrise Valley Drive

Reston, VA 20192

(703) 319-1173 Phone

jbeach@nbii.gov (E-mail)

Dr. Jim Brown

Director

Division of Biological Infrastructure

Room 615

National Science Foundation

4201 Wilson Blvd

Arlington, Virginia 22230

(703) 306-1470 Phone

(7030 306-0356 FAX

jhbrown@nsf.gov (E-mail)

Dr. Perry B. Cregan

Soybean and Alfalfa Research Laboratory

USDA-ARS, Bldg. 011, HH-19

BARC-West

Beltsville, MD 20705-2350

(301) 504-5070 Phone

(301) 504-5728 FAX

pcregan@gig.usda.gov (E-mail)

Dr. Gladys Cotter

USGS/BRD

300 National Center

Reston, VA 20192

(703) 648-4090 Phone

(703) 648-4042 FAX

gladys-cotter@usgs.gov (E-mail)

Dr. Dan Drell

Biologist, Human Genome Project

Office of Health and Environmental Research

ER-72/GTN

US Department of Energy

19901 Germantown Road

Germantown, MD 20874-1290

(301) 903-4742 Phone

(301) 903-8521 FAX

daniel.drell@oer.doe.gov (E-mail)

Dr. Richard DuBois

Biomedical Technology

National Center for Research Resources

National Institutes of Health

One Rockledge Center, Room 6146

6705 Rockledge Drive

Bethesda, MD 20892-7956

(301) 435-0755 Phone

(301) 480-3659 FAX

richardd@ep.ncrr.nih.gov (E-mail)

Dr. Paul Gilna

Program Director,

Database and Computational Biology Activities

National Science Foundation

4201 Wilson Blvd, Room 615

Arlington, VA 22230

703.306.1470 x 6410

703.306.0356 (Fax)

pgilna@nsf.gov

1-800-509-2493 (Page)

Dr. Ed Kaleikau

USDA-CSREES-NRI

Room 330-F, Aerospace Center

901 "D" Street, SW

Washington DC 20024

(202)401-1901 Phone

(202) 401-6488 FAX

ekaleikau@reeusda.gov (E-mail)

Dr. David Lipman

National Center for Biotechnology Information

National Library of Medicine

National Institutes of Health

Building 38A, Room 8N805

9000 Rockville Pike

Bethesda, MD 20892

(301) 496-2475 Phone

(301) 480-9241 FAX

lipman@ncbi.nlm.nih.gov (E-mail)

Dr. Robert E. Menzer

USEPA (8701)

401 M Street, SW Washington, DC 20460

(202) 260-5779 Phone

(202) 260-0929 FAX

Menzer.Robert@EPAMAIL.EPA.gov (E-mail)

Dr. Henry L. Shands

Assistant Administrator

Genetic Resources

USDA/ARS/OA, Rm. 319-A

Jamie L. Whitten Federal Building

14th & Independence Ave., S.W.

Washington, D.C. 20250-0300

(202) 205-7835 Phone

(202) 690-1434 FAX

shands@sun.ars-grin.gov (E-mail)

Dr. Deborah L. Sheely

Assistant Program Director

NRI Competitive Grants Program

USDA, CSREES

901 D Street, SW

Washington, DC 20024

(202) 401-1924 Phone

(202) 401-6488 FAX

dsheely@reeusda.gov (E-mail)

Dr. Anna Tsao

DARPA/DSO

3701 North Fairfax Drive

Arlington, Virginia 22203-1714

(703) 696-2287 Phone

(703) 696-3999 FAX

(703) 696-0218 FAX

atsao@darpa.mil (E-mail)

Dr. Judith L. Vaitukaitis--CHAIR

Director

National Center for Research Resources

National Institutes of Health

9000 Rockville Pike, Bethesda, MD 20892-2128

(301) 496-5793 Phone

(301) 402-0006 FAX

vaitukaitis@nih.gov (E-mail)

Dr. Keith B. Ward

Program Officer

Biological Sciences and

Technology Program

Office of Naval Research, Code 335

800 North Quincy Street

Arlington, VA 22217-5660

(703) 696-0361 Phone

(703) 696-1212 FAX

wardk@onr.navy.mil (E-mail)

RECOMMENDATIONS FROM INDIVIDUAL WORKSHOP PARTICIPANTS

On the final day of the workshop, many participants distributed their own lists of recommendations relating to bioinformatics. Some of their suggestions were thoroughly discussed and considered during the workshop; others were not subjected to critical review during the two-day session. Although no consensus was reached on many of these proposals, the Research Resources and Infrastructure Working Group may wish to consider their recommendations.

Peter Arzberger, University of California, San Diego

1. Government agencies should consider **review mechanisms that reflect the fundamental differences between research and infrastructure** (e.g., in procuring and monitoring that resource). Specific suggestions: Decouple decisions about resources from specific proposals in the same scientific area; implement an STC review mechanism (e.g., three years of guaranteed funding, with annual reviews and options for extending the resource).

Note: Stable funding for a resource is CRITICAL, both for planning purposes and for the "security" of users.

2. **Infrastructure MUST contain a development component** (e.g., training of users). See example below describing NSF support for advanced computing. I think it is essential that biological information resources are charged and expected to continuously develop the resource.

3. Federal agencies should ensure some degree of **interconnectivity between databases.** Databases should not operate as resources for a single community (e.g., the depositors). Agencies should focus on integration (e.g., mindset of PIs, mindset of program officers); this might best be accomplished via even "higher" authorities, such as via interagency efforts.

4. **Training.** Emphasize Ph.D. level or postdocs over masters level. There are several models for encouraging universities to establish training programs. The NSF program in biology (Research Training Groups, RTG, now replaced by NSF-wide IGERT) is a mechanism that encourages an interdisciplinary approach and pushes universities to adopt the programs after funding.

5. Get the word out regarding proposed changes to U.S. **intellectual property laws** and the profound impact new European laws may have on information sharing.

6. Encourage continued discussions on **standards.**

Note about NSF support for advanced computing: In 1984-1985, NSF responded to community requests for supercomputer access by establishing five supercomputer centers focused on providing access to "cycles," a pure service model. In 1990, NSF encouraged development of enabling technologies, and the centers welcomed their new intellectual roles in the enterprise. (The pure service model employed by many university academic computing centers was not successful.)

In 1995, NSF announced a new competition to continue its support for the advanced computational infrastructure but to also include scientists from the academic communities. These "partnerships" were asked to provide the scientific community with access to the resources, to develop tools and environments to improve the resources, and to provide education and outreach to a variety of communities.

Douglas Brutlag, Stanford University

I. Research programs in genomics and bioinformatics: To organize genomic information in biologically meaningful ways and then apply this information to health care.

1. Genomics research: Classify sequences with known functions and in known families; identify unique sequences specific for human diseases or unique to disease-causing organisms; develop diagnostic methods for detecting disease while still treatable (e.g., DNA sequence diagnostics, gene expression diagnostics, tissue- and organ-specific diagnostics).

2. Bioinformatics research: Identify potential drug targets; identify rational drug and other therapies for disease.

3. Informatics research fundamental to genomics and bioinformation: classification algorithms, statistics, artificial intelligence, data models, hardware approaches, graph theory. II. Infrastructure: To ensure that the above information is readily available to the communities that need them, including researchers, educators, and industry.

1. Develop international standards for representations of biological and genomic entities, so that information can be represented and exchanged in an automated fashion. Objects might include gene sequences, gene maps, gene products, metabolic maps, annotations, etc.

2. Create permanent government-sponsored repositories for such information (e.g., with the U.S. Patent and Trademark Office, the National Library of Medicine, or the Library of Congress. Repositories could also be subcontracted to commercial firms that are qualified to maintain complex databases).

3. Develop international collaborations for exchange of information.

4. Support high-speed Internet and Internet II infrastructure to ensure the widest possible distribution of information.

III. Training and education:

1. Support predoctoral and postdoctoral degree candidates performing research in above interdisciplinary fields.

2. Ensure that practicing genomicists and bioinformaticians are formally trained in both the biological and informatics fields. This will help ensure the biological relevance of their work and ensure that informatics approaches are solid.

3. Support novel teaching methods that can repackage educational information for individuals in industry.

4. Support collaborative efforts to train students for industry. Such educational programs might be funded by industry.

David Benton, SmithKline Beecham Pharmaceuticals

I. Database and software interoperability standards. The federal government should

encourage database and software developers to participate in standards-adoption processes and then implement relevant standards that promote interoperability among databases and software components.

The Object Management Group (OMG) technology adoption process should be used to establish standard object-oriented interfaces for database services. The OMG has recently established a Life Sciences Research Domain Special Interest Group to coordinate its activities in this "vertical market domain."

II. Training. The federal government should stimulate and support doctoral-level training programs to educate the next generation of bioinformatics researchers (computational molecular biology and genomics) and theoreticians. Training masters-level bioinformatics practitioners is important (particularly to industry) but of lower priority.

Lois Blaine, American Type Culture Collection

1. Agencies should **set aside funding for programs in bioinformatics,** and fund both infrastructure projects (software, tool development, databases) and research projects involving bioinformatics. Specific programs may be in direct line with agency missions and goals, but primary evaluation criteria should include how the planned project interfaces or operates with other biological resources.

2. Agencies should **share the burden of supporting major cross-disciplinary community databases.** There are some existing examples of such resource-sharing, but the number of such projects should increase.

3. Agency representatives should **participate more fully in regional, national, and international standards-setting bodies** and should demand that grantees adhere to standards where they exist. Many rank-and-file bench scientists are not even aware of the work of organizations such as the International Union of Biological Societies and others.

Anthony Kerlavage, The Institute for Genomic Research

1. Fund development of a **standardized datamodel** for a subset of commonly used data types (e.g., genes, transcripts, proteins, features, etc.).

2. Encourage development of **specialized databases** *(not databanks)* (e.g., microbial, plant, human).

3. Fund establishment of analysis "nodes" (e.g., like BIONET, ANGIS).

4. Provide **documentation and tool sets for access and utilization of data** by the biological community.

5. Fund **focused development of new tools** (i.e., contracts for solving known problems).

6. **Training** specifically for bioinformatics (Ph.D., M.S., B.S.) *and* for end users.

David Matthews, Cornell University

Priorities (in descending order of importance):

1. A fellowship program for M.S. level graduate studies in bioinformatics.
2. A grant program for investigators to pay for bioinformatics services (e.g., privatized databases,

informatics staff, contract software development).

3. A mechanism for supporting scientific experts in preparing data for contribution to databases.

4. A grant program for development of software generally useful to multiple database programs.

Kenneth Paigen, The Jackson Laboratory

Proposed actions for the federal government:
1. **Establish an interoperability working group to set standards** (computational and nomenclature) required of all grantees.

2. **Connect with major scientific societies** (e.g., neurosciences, cell biology, microbiology) and ask for working groups to describe informatics needs with distributed priorities attached (100 points distributed among items). Proposed changes to federal policy:
1. **New database proposals must present a growth plan;** increases in annual funding will be contingent upon meeting the plan's goals and milestones.

2. All **databases must present the status of their progress and plans** to at least one, preferably two, major scientific meetings each year.

3. **All database grant renewals are on a rolling basis,** with annual reviews. At each review successful databases receive another three years of guaranteed support; questionable performance, only another two years of funding; and bad performance will receive notice that the project will be open for competition.
Some needed databases:

1. Cell Anatomy: Proteins, organelles, cell types and conditions.

2. Immunology: A model of the immune system.

a. Dynamics, showing development

b. Store information by associated function

c. Be able to manipulate parameters to predict outcomes

3. Gene regulation: Promoters, enhancers, transcription factors, spliceoforms.
4. Animal models of disease: Details about the models—the species, strain, mutant, and procedures that provide appropriate experimental materials.

Robert J. Robbins, Fred Hutchinson Cancer Research Center

I. We support the findings of the 1995 NSF workshop and urge federal agencies to recognize that **access to information and information technology will be essential for 21st century biology.** In other areas of human endeavor, support for large-scale information technologies consumes 5-10 percent of total gross revenues; successful 21st century biology will require similar levels of investment in bioinformatics.

II. Information technology moves at Internet speed, while federal proposal-review-funding cycles are slow. Agencies should consider developing **newer, faster, more efficient methods for supporting large-scale public information resources,** possibly even to the extent of shifting from supply-side to demand-side funding.

The need for some demand-side funding will become mute when advances in biotechnology reach the point where the majority of molecular biology data are produced in the private, not the public, sector.

III. Federal support for infrastructure, especially information infrastructure, requires that agencies recognize that they are **acting as procurement officers for the scientific community** and modify their actions accordingly. At a minimum, this must include active attempts by program officers to ensure and improve the quality and interoperability of goods and services procured and may even require adoption of a two-phase review process, with the first step being the establishment of priorities and the second selecting "vendors" to meet these prioritized needs.

Henry Shands, Agricultural Research Service, U.S. Department of Agriculture
Suggestions for the Office of Science and Technology Policy (OSTP) on Research Resources and Infrastructure:
1. To help guide federal agencies that manage databases or fund research that depends on databases, **OSTP should provide guidelines that clearly establish the administration's position on electronically stored mass data.** This could be done through the normal departmental budget process and the Office of Management and Budget review process. The position should enunciate:
That the information/data are important to the U.S. scientific (or other) community

That each agency should implement a plan to support databases at an appropriate level of funding that will make data available in a user-friendly format, in a timely manner, consistent with common practices. 2. OSTP should request that **agencies report the funding levels of their genomic databases relative to the amount of research data** through the budget process.

3. OSTP should **evaluate the genetic resources collections and their funding through the Office of Management and Budget** and departmental budget offices. Some collections have stakeholders in other agencies, and these should be identified. In some cases, it may be appropriate that funding and responsibility shift from one agency to another as missions change. *Sylvia Spengler, Lawrence Berkeley National Laboratory*
Realistic training budgets and stipends are needed across programs and agencies. Start stipends at $30,000, ranging up to $40,000.

Commitment to data availability and accessibility is needed.

Standards for interoperability are needed (e.g., via support for working groups). Suggested roles for the federal government:
1. Create **training grants** specifically for bioinformaticists: Begin with graduate students; extend up to PhD, down to MS; exceptional stipend.

2. Support **distance learning** (curriculum development and availability), not just through NSF but across agencies. (e.g., FIRST and R01s or R21s for new bioinformatics faculty/researchers, possibly computer scientists as well).

3. Develop (inter)agency mechanisms for **tracking priority achievement in long-term projects.**

4. Possible **measures for setting priorities:** Use by researchers, pay-off, value-added for users/role in research; data in the database are publicly available (i.e., not private or proprietary).

5. Review panels for **infrastructure vs. bioinformatics:** Create infrastructure panels with funding from the infrastructure budget; ensure bioinformatics competency (not just ad hoc) on panels, since

bioinformatics cuts across many other disciplines.

6. Think about **public vs. private efforts:** Questions of access, ownership etc.

Paul Uhlir, National Research Council

Policy recommendations for access to federal biological research data:
1. Endorse the principle of **"full and open" availability of basic research data** created, maintained, and disseminated with federal government funding. By "full and open," we mean that "data are made available with as few restrictions as possible, on a nondiscriminatory basis, for no more than the cost of reproduction and dissemination."

2. Support the OMB Circular A-130 **prohibition against the commercialization of federal government data dissemination functions.**

3. In those cases in which the **government chooses to privatize certain data management and dissemination functions,** it must protect access to those data for research, education, and other public interest uses.

4. In all **cooperative intergovernmental research activities,** the federal government should promote and adhere to exchange of data on a full and open basis.