REPORT OF THE INVITATIONAL DOE WORKSHOP ON GENOME INFORMATICS, 26–27 APRIL 1993 BALTIMORE, MARYLAND

DAVID T. KINGSBURY, CHAIRMAN
JAY SNODDY, EXECUTIVE SECRETARY

GENOME INFORMATICS I: COMMUNITY DATABASES

ROBERT J. ROBBINS, EDITOR

Authors:

Michael Waterman
Ed Uberbacher
Sylvia Spengler
F. Randall Smith
Thomas Slezak
Robert J. Robbins
Thomas Marr
David T. Kingsbury
Paul Gilna
Chris Fields
Kenneth Fasman
Dan Davison
Michael Cinkosky
Peter Cartwright
Elbert Branscomb
Helen Berman

ESP

ELECTRONIC SCHOLARLY PUBLISHING

HTTP://WWW.ESP.ORG

Robbins, R. J. [Ed.] 1994. Genome informatics I: Community databases. *Journal of Computational Biology,* 3: 173–190.

REPORT OF THE INVITATIONAL DOE WORKSHOP ON GENOME INFORMATICS, 26–27 APRIL 1993 BALTIMORE, MARYLAND

DAVID T. KINGSBURY, CHAIRMAN
JAY SNODDY, EXECUTIVE SECRETARY

GENOME INFORMATICS I: COMMUNITY DATABASES

ROBERT J. ROBBINS, EDITOR

Authors:

Michael Waterman
Ed Uberbacher
Sylvia Spengler
F. Randall Smith
Thomas Slezak
Robert J. Robbins
Thomas Marr
David T. Kingsbury
Paul Gilna
Chris Fields
Kenneth Fasman
Dan Davison
Michael Cinkosky
Peter Cartwright
Elbert Branscomb
Helen Berman

Note: The ideas in this report were authored by all of the meeting participants. The writing of the final report was done by Robert Robbins, based on notes taken at the meeting by Jay Snoddy, and on extensive input from meeting participants. The report drafts were reviewed and revised several times by all meeting participants before being approved.

PREFACE

On 26–27 April 1993, many workers, all actively involved in developing and deploying information resources for the Human Genome Project (HGP), attended a workshop in Baltimore, Maryland. The purpose of the workshop was to begin a systematic assessment of the state of information resources relevant to the HGP, especially community databases, and to provide recommendations for future improvements, both in terms of needed activities and improved policies.

This document reflects discussions that occurred at the workshop. As a meeting report, not a learned essay, it has no bibliography and no detailed summary of related efforts by others. The absence of formal citations and bibliographies is a natural consequence of the format and does not indicate a lack of respect for earlier work.

Although the text of the document has been edited and reorganized to facilitate reading, the general content and tenor of the meeting has been retained. The report is not an exhaustive treatment; many other relevant issues might also have been addressed. The omission of a topic reflects only the reality of the discussion, not an assessment of its importance.

This meeting and its report will be the first of a series to assess the state of genome informatics and to offer general advice and guidelines for future development. Additional meetings will address database interoperability, general requirements for genome informatics, the migration toward object–oriented technology, common data models, and other topics. Because genome informatics shares many concerns with the general information–infrastructure requirements of all biology, it is hoped that these studies may also prove useful to other scientific communities.

Comments and suggests regarding this report or related activities are welcome and may be sent to bioinfo@er.doe.gov.*

* ESP NOTE: This email address is no longer operational.

CONTENTS

First published in: *Journal of Computational Biology,* 3: 173–190.

INTRODUCTION

THE HUMAN GENOME PROGRAM (HGP) is producing large quantities of complex map and DNA sequence data. Informatics projects in algorithms, software, and databases are crucial in accumulating and interpreting these data in a robust and automated fashion at genome and sequencing centers. Furthermore, the data will need to be captured into robust community databases and accessed with equally robust analysis tools; biologists will need to ask questions of the data accumulated by the genome program and other research.

The future success of the genome project will depend on the ease with which accurate and timely answers to interesting questions about genomic data can be obtained. The Department of Energy (DOE), the National Institutes of Health (NIH), and other agencies must exercise leadership and control to ensure that needed informatics systems are developed and operated appropriately.

Recognizing the importance of informatics to the success of the genome project, DOE supports a portfolio of independent research projects in genome informatics, as well as core informatics activities at genome centers. To ensure the continuing high quality of these programs, last year OHER/DOE asked an independent panel to review the entire DOE program of informatics projects. The meeting reported here is a continuation of the long–term planning process initiated with that review. This meeting was also designed to feed into DOE development of new 5–year plans for the HGP.

In addition, this planning will be useful in other DOE or OHER/DOE research programs that will require an infrastructure to collect, interpret, and integrate diverse biological data. Some other programs or interests include health effects, mutation research, structural biology, biotechnology research (including applications in environmental biotechnology and the biological production of fuels, biomass, or other materials), and environmental research (including research and modeling that allow a better understanding of the effects of environmental perturbations on organisms, their ecology, and the environment).

TYPES OF INFORMATICS PROJECTS

Computer systems play essential roles in all aspects of genome research, from data acquisition and analysis to data management. Without powerful computers and appropriately designed data–management systems, high–volume genome research cannot proceed. DOE and other agencies support a variety of genome–related informatics projects, which can be broadly classified as data–acquisition, data–analysis, or data–management systems.

Data–acquisition systems

Data acquisition will be required at all research labs that are generating large amounts of data. Genome centers and similar laboratories will require very strong local informatics core support for acquiring data efficiently. Some specific examples of data–acquisition systems include:

Inventory Control Software: A major genome center may require several hundreds of thousands of reagents, gels, and other materials. Because manual tracking would be impractical and inefficient, automated inventory control systems are needed. Such systems are essential, albeit not glamorous, components of a major genome center.

Reagent Manipulation Software: Robotic systems are now required to carry out high–volume, high–precision laboratory manipulations in genome research. As the HGP switches into production mode, computer support for robotics will become increasingly important.

Sequence Production Software: At present, computer systems are required in almost all aspects of sequence generation and assembly. If the genome project is to attain its stated goals of obtaining whole–genome sequences by 2005, improvements in sequencing technology of several orders of magnitude will be required. Improved software systems will play a crucial role in these advances.

Visualization Software: Much genomic data begins as images. Interpreting gels, reading filters, and many other steps on the critical path of genome analysis require computer tools for automation and optimization.

Data–analysis systems

Genome data cannot be analyzed efficiently without computer systems. Studying sequences, predicting protein structures, and comparing genomes on an extensive scale all require additional informatics tools, such as:

First published in: *Journal of Computational Biology,* 3: 173–190.

Sequence Analysis Software: Sequence analysis is perhaps the best–known, best–established area of genome informatics. Performing alignments, detecting homologies, identifying coding regions, extracting features; and other computerized analyses of sequences are now so commonly performed as to be routine. At the same time, sequence analysis is a multifaceted and biologically profound area of research, demanding much continued work.

Protein Folding Software: Genetic information is transformed into function *via* proteins, whose functional specificities are determined by their three–dimensional shapes. Genome and related research will yield very important results when the prediction of protein structure and function from amino–acid sequences is possible.

Map Assembly and Integration Software: With larger maps (composed of different fundamental kinds and combinations of data) being generated, computation plays an increasingly central role in their assembly and integration.

Physical Mapping and Contig Assembly Software: Assembling n clones into an ordered physical map usually involves generating a large, $n \times n$ matrix of comparisons, then deducing a possible order from the matrix. With any reasonable n, performing these analyses manually is impossible.

Genetic Mapping Software: Software systems play a key role in the analysis of genetic mapping data. Many programs and algorithms are now in use, and improvements are proposed regularly. This will continue to be an active informatics challenge for HGP.

Comparative Genomics Tools: As the genome project matures and large amounts of genomic information are available for a number of species, comparative genomics will emerge as an active area of study.

Classification Software: Extracting features from DNA sequences, placing proteins into gene families, and tracking protein motifs are all key activities in genome research and all need computer tools. Much work has been done in these areas, and more will be required as HGP moves into production mode and sheer data volume precludes extensive manual classification.

Data–management systems

The genome project is generating information that cannot be accommodated by traditional publishing. Ready access to these findings is essential for interpreting current experiments and planning future work. Now, electronic data management and publishing systems are increasingly crucial components of genome research. These systems range from highly specialized databases supporting local research projects to general databases that support the entire community.

Local Databases: Local databases are usually developed at one site and designed to handle specific, local needs. They are usually a closed resource, available only to local researchers, and containing both raw data and refined information. Often they are also involved in local inventory control. Tightly integrated into local bench research, they must be flexible and capable of rapidly tracking changes in local experimental protocols. The requirement for flexibility and responsiveness often exceeds a need for robustness and general applicability. Formal user–support systems (in the form of extensive manuals, regularly staffed help desks, *etc.)* are rare, but informal support is crucial — local users must be able to work with the system. Most importantly, local systems must quickly meet the specific, idiosyncratic requirements imposed by specific protocols used by local researchers.

Collaborative Databases: Collaborative databases are conceptually in the middle ground between local and community systems. Information is collected from a larger set of collaborating researchers, perhaps those working on a particular chromosome. Again, both raw data and refined information are included, with an increasing emphasis on integrating findings from several laboratories. Requirements for robustness are greater than with local systems, since the system may be used by some who do not have access to local computer expertise. More adequate user support is also needed.

Collaborative databases are frequently discussed and proposed, but at present it is difficult to identify any extant examples. Many of the problems associated with building such databases seem to be sociological, rather than technical.

Community Databases: Community databases are shared resources, open to the entire research community. Although they have traditionally emphasized refined information over raw data, the trend now is for community databases to play an increasing role in making raw data *(e.g.,* the underlying data on which a published genetic map is based) available as a research resource. Community databases must be robust, because the majority of their users will be located off site. Formal user support, in the form of manuals, help lines, regularly staffed help desks, and training are required. Because these systems must meet the *consensus* needs of the entire community, care must be taken in their design to ensure that the databases are sufficiently flexible and robust to address the needs of different user communities. Because users need to integrate findings from several different community databases, each community database should be designed as a component of a larger information infrastructure for the biological sciences. Specifically, community databases should recognize the

First published in: *Journal of Computational Biology,* 3: 173–190.

biological interdependence of information in multiple databases and should provide support for integrated queries involving multiple databases. To this end, they should be built from standard components and must be well documented.

OVERVIEW OF COMMUNITY DATABASES AND PROJECTS

The meeting began with general discussion of OHER–supported research and the role of community databases in facilitating that research. Two main OHER–supported research programs with significant informatics components are Structural Biology and the HGP. Structural–biology research is, in part, an effort to relate biological function to molecular shape. This requires physically separate, but logically joint access to databases of protein sequences (PIR–International and SwissProt) and to crystallographic structures (PDB).

The HGP is a multinational, multimillion dollar research program with several deliverables, including nucleic acid sequences and genome maps (and the technology to produce and interpret these data). HGP cannot succeed without appropriate community databases to collect, manage, and distribute data. GenBank, EMBL, and DDBJ presently handle sequence data for all species. Genetic data (genes, maps, probes, *etc.)* are managed in species–specific databases such as GDB for humans, FlyBASE for *Drosophila,* and the Mouse Genome Database (MGD) for the mouse. Not merely archives where findings are stored for historical purposes, genomic databases are needed as dynamic resources to serve as analysis tools that will greatly influence the ongoing HGP bench work. As HGP data volume grows exponentially, bench work will become increasingly dependent upon access to appropriate information for both the design and the interpretation of experimentation.

The overview presentations concluded with a discussion of the role of community databases in OHER research programs. Community databases are essential for the success of OHER research. At the same time, all extant community databases have serious deficiencies and fall short of meeting community needs. Deficiencies derive both from rapidly changing needs and conceptual and technical flaws in the design of present systems.

This report summarizes the discussions of OHER informatics needs (especially for community databases), identifies new and unmet needs, and recommends actions to ensure that these needs will be met.

TYPES OF COMMUNITY DATABASE USERS

Databases and other informatics infrastructure can be justified only in terms of the scientific superstructure they support. Because different users of infrastructure have different requirements, a first step in considering the utility of informatics projects is to identify their users.

With regard to community databases of genome data, the group felt that users and producers of genome information fall into one of four broad categories: (i) genome centers, (ii) independent laboratories at major research organizations, (iii) individual researchers and small laboratories, and (iv) other users.

Genome centers

NIH/NCHGR and DOE genome centers are large, multi–investigator facilities with the specific mission of genome research. These centers usually employ an informatics staff to operate a robust local computing environment, depend on computers to manage their own data and to plan and interpret their bench research, and are connected to the Internet. With their local computer expertise and their ability to generate and use large amounts of genome data, these sites must have direct access *(i.e.,* network–mediated, program–to–program access) to the community databases, both for the extraction of data into their local computing environment and for the submission of new findings into the databases.

Such sites depend upon ready access to community database information for their operations. Local informatics staff develop custom software, both for maintaining and manipulating local data and for integrating access into multiple community databases. Because of varying research needs, these centers must have local capabilities and software. However, community databases can impede local efforts by making unexpected changes in data structures or interfaces or otherwise altering the behavior of their systems. Therefore, if HGP is to maximize the value of its collective investment in informatics, community databases must be operated to facilitate, not impede, the activities of informatics personnel at genome centers.

To this end, community databases should adhere to certain technical standards, such as making their systems interactively available over the Internet through stable and well–documented application programming interfaces (APIs). Community databases should also follow some minimal, HGP–wide semantic standards. One critical example is the use of stable, arbitrary external identifiers (accession numbers) for individual objects. Human retrieval of information is most intuitive when biological nomenclature can be used

First published in: *Journal of Computational Biology,* 3: 173–190.

as identifiers. However, periodic revision of biological concepts causes automated retrieval systems to fail if biological nomenclature is the only way to identify data objects. Automated systems cannot work reliably without accession numbers.

In addition, individual community databases must be charged with collecting, maintaining, and distributing information that shows how data objects in their system relate to data objects in other systems. An embarrassment to the HGP is our inability to answer simple questions such as, "How many genes on the long arm of chromosome 21 have been sequenced?"

Independent laboratories at major genome research organizations

These laboratories are smaller than centers and may not have a dedicated local informatics staff. They may have access to shared computational resources provided by their department or host institution. The host institution probably has Internet access. The laboratory may have direct Internet access and, if not, probably has indirect access through some shared institutional computers.

These laboratories need access to community databases and will also be providing data to them. The need for electronic, query–level access that integrates across community databases is especially pressing and will be increasingly important. Because these labs do not have dedicated computer staff, they usually work through interfaces provided by the community databases or perhaps through shared systems developed at their host institution, but not under their control. They may produce enough new data to require some kind of electronic data submission, but less than the bulk transfers required for genome centers. These laboratories will require computer tools to help them build and annotate their submissions, and to provide updates, corrections, and other improvements to data already submitted.

Individual researchers

Individual researchers who are not part of a major research center may have only limited access to powerful computing facilities, even on a shared basis. Their organization may have an Internet connection, but their own facilities are frequently not directly attached to the local organizational network, resulting in network connection only by dial–up to a networked computer.

Although individually these laboratories may not be involved in large data transfers, collectively they produce much important information. The workers need access to computerized tools that will operate on PCs or Macintoshes to assist in preparing data for database

submission. They will also need tools for integrated access to multiple community databases. Because their access will be only through formal interfaces developed by others, often without access to any on–site technical assistance, these workers need very robust, easy–to–use systems to assist them in interacting with the community databases.

Some of these laboratories may be housed in institutions that do not have Internet connectivity or even local area networking. Such workers need help in convincing their institutions of the need for network access.

Other users

Increasingly, educators and students in higher education (and even in high schools) are making use of genome information. These individuals have widely varying access to computer facilities and to networks. Because they use the information for pedagogical rather than research purposes, their needs for on–line help, browsing tools, and explanatory documentation are acute. Although the ultimate justification for the development of community databases must be in terms of the science, research, and practical developments that they support, the needs of educational users should not be forgotten. All future researchers will pass through the educational system, and training in the use of these systems will become an increasingly necessary and routine part of biological education.

Furthermore, genome data will increasingly be needed by clinical centers and genetic counselors. Other individuals may wish to access genomic information for a variety of reasons.

Although less sophisticated users must be supported, community databases and access methods must not be designed to meet their needs only.

USER REQUIREMENTS

Responsive databases

Community databases must meet user requirements and respond to changing needs. For this to happen, every community database should be subject to independent peer review and regularly evaluated by an independent panel of experts. Mechanisms must be in place to allow advice to flow readily from the community to public database operators.

First published in: *Journal of Computational Biology,* 3: 173–190.

Internet access

Routine data submission, data retrieval, and data analysis will work best through the Internet. Serious producers and consumers of genomic data will require Internet access. The genome community should work with other scientific and medical disciplines to increase awareness of Internet importance. Access to the Internet should be assumed for computational genome work. DOE and other agencies should strive to gain access for those who do not have it, rather than develop special technologies to accommodate them.

Complex queries

Complex queries are becoming more important as user sophistication increases and databases are used for data discovery. Initial use of sequence databases was often for the simple query of, "Has my new sequence been submitted before?" As more information is obtained from the genome project and other research efforts, complicated queries will become increasingly useful and needed. Answering many queries will require simultaneous, coordinated access to multiple databases (*e.g.*, queries that cross GDB and GenBank).

More support for complex, multi–database queries will require major efforts toward improving the integration and interoperability of community databases. Community database providers should work to meet these needs, and agencies should support these efforts. Incentives must be provided to community database operators for their local efforts to ensure database interoperability. All community databases should be available online through some directly accessible API. An API allows users to develop their own custom programs that can directly extract and manipulate data from community databases. Without an API, researchers must spend excessive time manually identifying, extracting, and formatting data from community databases before further analyses can begin.

Ad hoc queries

Databases must be designed with standard interfaces that permit the easy retrieval of answers to common questions. However, software developers cannot predict all present and future needs, so community databases should also permit ad hoc queries not accommodated by the standard interface. Such query capability should be unlimited, in that users should be able to obtain answers to any reasonable query that is, in principle, answerable from the contents of the database.

Unlimited ad hoc queries are possible only if databases provide technical support for them. This can be done, without a major

development effort, by taking advantage of commercial database systems that provide a query–language interface. In addition to making ad hoc queries technically possible, community databases must also make them practically feasible by providing enough documentation and training to allow users to formulate meaningful queries.

At present, all relational databases provide a standard query language, SQL, whereas object–oriented databases offer only product–specific query languages or none at all. At the same time, the richness of object–oriented data models makes these databases attractive for genome work. Thus, the goals of supporting ad hoc queries and of achieving a better data model are in conflict.

Integrated analytical. tools and databases

To use the information in community databases, users require software for analysis and synthesis. These should be incorporated into suites of software tools and servers that can run across the network, both to extract the necessary data from the databases and to perform the actual analyses. Developing such integrated systems cannot be done unless community databases and software tools provide external arbitrary identifiers, documented semantics and schemas, shared data concepts, and APIs.

Integrated community databases

As multi–database queries become more important, users will require access to integrated views of data from multiple databases. Choices among alternate interfaces and software that link databases will also be needed. Database links (*i.e.*, connections between objects in different databases) should be able to scale up to the large amount of biological information that will be incorporated in them. These links should also be as inherently unrestricted as possible in permitted queries. Ultimately, biologists should not have to know where data are located.

DATA SUBMISSION AND CURATION ISSUES

Even if great success attends genome bench research, the overall HGP will be viewed as less than a success if users are unable to share their findings with others by submitting their data into robust community databases in an easy and timely fashion. If repeating experiments becomes easier than locating previous results, genome informatics will have failed.

First published in: *Journal of Computational Biology,* 3: 173–190.

In addition, scientists must be able to trust information quality in the community databases and to obtain useful data from them easily. Systems for improved data submission and curation will be greatly needed.

Data submission and accession

Sequence databases were developed before HGP became a formal program. Initially these databases functioned in a manner similar to the production of review papers — experts read the literature to extract, then interpret, previously published findings. This worked well enough when data volume was very low, but by the mid 1980s the nucleotide sequence databases had a year–long backlog of identified but not–yet–accessioned data. At the time, two proposals for remedying the backlog were made: (i) speed journal scanning by reducing sequence annotation and tying the scanning to other on–going activities such as the preparation of MedLine and (ii) develop tools to allow researchers to submit their data directly to the databases. By 1991, direct data submission had solved the backlog problem. For example, the GenBank backlog had disappeared and the time necessary to accession an entry had dropped from months to days.

Also, for the first time direct data submission allowed editorial review and other quality control over sequence data. Directly submitted data could be subjected to a suite of analytical software to identify potential problems (ORFs that didn't translate as claimed, the presence of vector sequence, *etc*.). Results of these analyses could be returned to the researcher, who could then respond with an improved submission, as appropriate. This effectively replaces three error–generating steps (preparing data to send to the journal, typesetting the journal, and scanning the journal) with one error–correcting step. Compared with journal scanning, direct data submission gives better data faster. The continuous improvement of methods for direct data submission and curation must be a key goal for genome informatics.

With genome research, data are generated with components that need to be submitted to multiple databases. A coordinated research program at a genome center might well generate nucleotide sequences, new markers, raw and integrated maps, cDNA partial sequences, clone information, and other data that are now published through a number of different databases such as GenBank, dBEST, GDB, ATCC, and others. Genome centers would be unnecessarily burdened if they were obliged to prepare their data for multiple databases and to take responsibility for the validity of links among related elements in all the databases. Community databases should never require researchers to resubmit data directly to them if the same data have already been

submitted to another community database and can be obtained through a coordinated transfer.

Developing and perfecting coordinated direct data submission methods for all genome databases must be a high priority. Attention must be directed both toward developing easy–to–use systems for low–volume data producers and highly efficient bulk transfer systems for major centers. In principle, any biologist should be able to submit research results to multiple appropriate databases with a single electronic transaction. Achieving this goal will require coordinated oversight of all genome–relevant community databases.

Although the HGP does not fund gene function or related studies, such data must be captured in a useful way if the genome project is to be truly successful. Genome researchers, funding agencies, and the entire biological community should ensure that an infrastructure is in place to capture this type of data.

Genome centers

Genome centers have special needs for submitting their data into community databases. Because their data will be stored locally in electronic form, submitting data to community databases through completely automated means must be possible. Tools must be available to assist these organizations in preparing their data and appropriate annotation and documentation for submission. Given the volume of data likely to flow from such centers, community databases must be especially sensitive to the needs of these organizations.

To ensure data quality, many community databases rely upon external editors or curators. Increasingly, community databases are developing means for allowing data submitters to correct, update, and otherwise curate entries they submitted previously. This service is essential now for large centers and will become increasingly important for all researchers.

Linking data publication with paper publication

Database submission of sequences and protein structures is now required by many journals before publication of related papers. At the Chromosome Coordinating Meeting, 1992 (CCM92), some suggested that submission of genes, alleles, and maps be required as well. Enforcing such a mandate requires that professional societies, editors, and journals cooperate and that authors be able to document their submissions. Relevant databases now assign accession numbers to deposited sequence data, allowing authors to cite these accession numbers as proof of deposition. Other databases, such as GDB, have

First published in: *Journal of Computational Biology,* 3: 173–190.

recently implemented the accession number system that is helpful in making data submission a requirement for journal publication.

At the very least, any journal article that can contribute annotation information to database entries should include relevant accession numbers with its keywords and its MedLine reference.

Data curation and quality control

As databases take on a role similar to the primary literature, curation will become increasingly important. Tools are needed to allow and encourage data submitters to take responsibility for the continuing quality of their submissions. Curators must be appointed to oversee long–term quality and consistency of data subsets in community databases. A new professional job category, not unlike museum curators, may develop for these databases. Professional database curators and tools for direct author curation should be supported.

COMMUNITY DATABASE REQUIREMENTS

Considerable discussion was held on basic requirements that must be met by genome databases to support various user needs. Although discussions were wide ranging and vigorous, several consensus concerns emerged.

Database interoperability

Achieving coordination and interoperability among genome databases and other informatics systems must be of the highest priority. We must begin to think of the computational infrastructure of genome research (indeed, of biological research) as a federated information infrastructure of interlocking pieces. The distributed nature of biological research will require the development of multiple software and database projects. However, if we permit the luxury of independent efforts that do not productively interact, financial costs of systems development will be too large. In addition, rogue projects that do not successfully interact with other databases will lead to large scientific losses associated with unlinked data. Users must be able to retrieve related data from multiple databases such as GDB, PIR–International, Medline, PDB, and GenBank without having to make separate queries to the databases and then integrate the results themselves.

References to appropriate literature are a key component of any scientific database. MedLine could provide this information for the genome project, if it were available through an on–line API. However,

the current design and operation of MedLine renders it inadequate as the primary source of scientific literature data in a federated information infrastructure. The lack of an on–line, software–searchable MedLine means that each community database must duplicate some aspects of MedLine data and functionality.

The need for interoperability will increase so that database interoperability within just one research domain *(e.g.,* the HGP) will not be enough. Workers will need integrated access to a variety of biological information. For example, DOE considers studies on gene products and their functions as outside the domain of the DOE genome project. However, if the results of the genome project were never linked to an understanding of the function of gene products, then other researchers might reasonably feel that much of the value of HGP had been lost.

Major modifications to existing computer systems can be expensive. Therefore, if database interoperability is to increase over time without the need for periodic reengineering of existing systems and connections between different systems, community databases should be generically designed for interoperability. With present technology, achieving database interoperability requires both semantic and technical consistency among projects.

For minimum semantic linkage, the same unique identifiers must be used for the same biological objects in all interoperating databases. This is best accomplished if participating databases provide stable, arbitrary external identifiers (accession numbers) for data objects under their curation. References to these objects in other databases should always be made *via* accession numbers, not *via* biological nomenclature. Linking data objects between databases requires that the other objects be identifiable (accomplished *via* accession numbers) and relevant (accomplished *via* semantic consistency). Although perfect semantic consistency is probably unattainable, certain activities would be helpful. Community databases must document the *semantics* of their systems. A recurring problem is the existence of differing semantic concepts in different community databases.

The granularity of database objects also affects semantic linkage. For example, GenBank objects are reported nucleotide sequences and GDB objects are genes. So long as reported nucleotide sequences are about the size of genes, linking objects in the two databases is conceptually straightforward. As reported sequences increase in size, however, problems arise. When multi–megabase (or even whole chromosome) sequences are reported, linking them to GDB genes will provide little information.

First published in: *Journal of Computational Biology,* 3: 173–190.

For minimum technical linkage, the participating systems must present similar APIs to the Internet. At present, this is most cost–effectively achieved when all the interoperating databases are implemented as relational databases that support Internet SQL queries. Ideally, community databases should be (i) *self–documenting* (offer an on–line data dictionary and other documentation), (ii) *stable* (undergo schema change only rarely and then only after ample warning), and (iii) *consistent* (use federation–wide semantics). Note, however, that the goals of stability and consistency are in conflict with that of maximum responsiveness to changing community needs.

At present, local incentives often work against interoperability. Of the major community genome–relevant databases (GDB, GenBank, PIR–International, PDB), no two are funded by the same program, advised by the same advisors, or otherwise coordinated. This poses a great risk to the long–term success of HGP. Coordination among genome–relevant community databases is essential.

Because funding of community databases is always limited, interoperability issues may not rise to the top of local priority lists when so many other needs are pressing. Periodic peer review panels that include end–users who are frustrated by a lack of connections among the data are one incentive, albeit a small one, for database providers to attempt integration of databases. If the interoperability necessary for the success of the genome project is to occur, DOE and others will need to take steps to create high–priority local incentives for interoperability.

Standard data–transfer formats

HGP findings, generated in many laboratories around the world, must be transferred electronically from site to site. Because data will never be stored identically at all sites, some means for electronic data translation and transfer will be needed.

If each pair of sites wishing to exchange information developed a customized exchange process, great inefficiencies would ensue. For example, if customized pair–wise procedures were developed to allow 10 different sites to exchange data, 45 different data–translation procedures would be needed. However, if all sites could agree upon a common data–transfer format, then only 10 translation protocols would be needed. In general, if we let n = the number of sites and t = the number of required translation protocols, then for custom pair–wise procedures the number of required protocols is given by $t = n(n - 1)/2$, but with a common format $t = n$.

The development of appropriate data–transfer formats will facilitate a federated information infrastructure. Appropriate industry

standards (present and likely future candidates) should be considered whenever applicable, but care must be taken in choosing and applying standards so that the genome project will not be hobbled by the selection or misapplication of an inappropriate industry standard.

Data distribution versus data exchange

The development of common data–transfer formats alone, however, will not solve the data–transfer challenges of HGP. To see why, we must first distinguish between mere data distribution and true data exchange. Data distribution involves the unidirectional movement of data, with no expectation that it will ever return to the sending site. In this case, the sending site is responsible for maintaining the data and for distributing it to other sites for use. If the receiving sites cannot use or accommodate all the data or its components, they simply discard the excess. True data exchange, on the other hand, involves a shared responsibility for data maintenance and the expectation of repeated exchange without loss or corruption.

Although a common data–exchange syntax is necessary for data–exchange, it is not sufficient. Using the data effectively requires understanding it, and that requires a common data semantics, without which some (or all) of the information will be lost with each data transfer. True, loss–free data exchange can occur only if participating databases first achieve some kind of semantic parity. The simplest way would be for one of the databases to adopt wholesale the internal semantics of the other. This, however, is often impossible to achieve in practice, because many differences between databases derive from real needs by local users to see and use data in particular ways.

Semantic parity might be achieved while maintaining different local views if databases employ common internal data structures to achieve parity and different external data views to meet local needs. However, this approach may also fail if yet another local view is required. The point here is not that "a lowest common denominator" approach is the solution, but rather that the problem is hard. Achieving semantic consistency in a federated database remains a major challenge–one that can never be met simply by adopting syntactic standards for data transfer. Instead, each participating database will be obliged to incorporate the requirements of federation participation into its local design decisions.

This discussion illustrates some important truths about database design. To meet specific needs of many users, databases may have to employ internal data structures that individual users consider too complex. User needs must guide the design of the system's external behavior, but user opinions should not be allowed to dictate internal

First published in: *Journal of Computational Biology,* 3: 173–190.

data structures, which should be designed to accommodate diversity, to support local requirements for viewing data, and to facilitate participation in a federated information infrastructure.

Appropriate system architecture

If a federated information infrastructure is to emerge, participating systems must follow a common system design plan involving a layered, modular architecture and distributed databases. The infrastructure should permit a client to pass queries transparently through multiple databases at different locations.

The explosive growth of networking systems has shown that modular architectures, with well–defined interfaces between modules, allow great flexibility in developing nicely coupled systems that are also capable of evolution and growth. Layering for data input into genome databases could involve the development of standard data–input systems that work against a standard input file structure. Then, electronic data submission software could be written to produce those standard files as output. On output, providing an API allows various third–party layers to be developed into one or more underlying databases. Ultimately, a knowledge–base layer could reside on top of other layers.

Until distributed database technology matures, community databases must have only a few central sites where data entry and editing are done. To ensure prompt and robust access to data and services, multiple satellite database and server sites will be needed. So that users do not have to "shop" for the most current version of the database, different remote sites will need to be kept updated and current.

To facilitate incorporation into the larger federated information infrastructure, individual participating databases should follow a common architecture and design. At minimum, databases should be a multi–user, networked, client–server system with a stable, documented API for the server. Community databases should be robust and constructed of industry–standard commercial products. Needed support for ad hoc queries is possible only if the system provides some sort of standard query language such as SQL.

At present, only relational database products meet all these criteria. However, the richness of the data models in object–oriented systems has led some to believe that they will be of increasing importance as they mature into commercial products and acquire support for ad hoc queries. On the other hand, the lack of an underlying formal data model, coupled with problems that attend queries cutting across the object hierarchies of an object–oriented schema, lead others to remain

cautious. With relational products acquiring object–like properties and with object–oriented systems able to communicate with relational systems, any shift of genome informatics into object–oriented methodology is likely to be gradual and evolutionary, with hybrid models playing a significant role for some time.

Support for high data volume

The collected data presently in genome community databases represent just a few percent of what will be produced by the completion of the genome project. Community software systems must be designed to scale up gracefully over several orders of magnitude as data flow increases. Community software tools must allow machine–readable input and output, since manual data handling simply will not scale over the expected increases in data flow. In addition, the conversion of data format from that required by one analytical program to that required by another is too labor intensive. We need standardized data–file structures so data may be prepared automatically for analysis by many different software tools.

Improved data models

Historically, informatics projects have developed their own data models to accommodate data being absorbed into their systems. In general, the older the project, the more inadequate its data model. This is true in part because of changes in science rather than inadequacies of the developers. For example, GDB inherited much of its data model from the HGML project at Yale, whose support for Human Gene Mapping workshops GDB was required to duplicate. In consequence, the primary mapping database of the HGP is based upon an inadequate data model that does not appropriately reflect changes in mapping technology that have occurred in the last few years. This is not meant to single out GDB for specific criticism because other community databases also have problems with their data models.

Data models for business data systems must reflect the practices of the business. If a proposed change in policy would be too costly to implement in the business database, the change may be deferred as a matter of policy. No such luxury exists for scientific databases, which need to reflect our understanding of the real world. When research advances change our perception of the real world, our databases must track the change or become inadequate. The effect of these advances can be reduced through careful planning in database design, but they can never be eliminated. Therefore, DOE and other agencies must recognize that all community databases will periodically require major

First published in: *Journal of Computational Biology,* 3: 173–190.

redesigns. The idea that a "properly designed" scientific database should never require modification is simply false.

Many different community databases have overlapping data–modeling problems. For example, all scientific databases need to connect their stored observations to some kind of citation, so they require that a portion of their data models accommodate literature and other citations. Having each community database develop its own solution to these overlapping problems is inefficient. In addition, community databases regularly need to reference the contents of other databases. This is done well only if the other data model is well understood and semantically consistent with the local model.

For these and other reasons, the development of a federated information infrastructure for genome research will require the development of common high–level data–model concepts. Although a single data model spanning all the federated databases is neither possible nor desirable, DOE and other agencies should promote improved, shared data models in those cases where shared concepts are essential for database integration.

Technical standards

Participating systems must be designed for interoperability and portability. This requires adherence to design standards, use of industry–standard hardware and software, and avoidance of bleeding–edge technology. Community databases will need to work with the National Institute of Standards and Technology and the appropriate International Standards Organization committees in developing the required standards in genome data and remote data access (RDA).

DOE and other agencies must assume a leadership role in developing and promulgating standards appropriate to HGP. Standards cannot evolve without centralized attention and efforts to facilitate consensus. Adherence to standards is usually of greater importance for the success of the overall federated system than for that of local systems. Therefore, DOE and other agencies must work together to ensure that local incentives are in place at sites whose participation in the greater federation is essential. The goal must be the adoption of minimum interoperability standards, so that adding a new database to the federation would be no more difficult than adding another computer to the Internet.

Semantic standards

A truly federated information infrastructure cannot be achieved unless some minimum level of semantic consistency exists among

participating systems. No amount of syntactic connectivity can compensate for semantic mismatches.

For example, information about human (ß–hemoglobin can be found in several databases, such as PIR–International, GenBank, GDB, and OMIM. Although it would seem a simple matter for the federated database to provide links that allow the user to traverse these entries easily, data objects in these databases can have fundamental semantic differences. In the past, PIR–International data objects were proteins in the chemical sense so two proteins with same structure were the same protein. Thus, the PIR–International entry for human (ß–hemoglobin actually was also the entry for human, chimpanzee, and pygmy chimpanzee (ß–hemoglobin. Although this policy has been discontinued by PIR–International, it is still evident in Swiss–Prot release 28.0, where entry P02023 is for (ß–hemoglobin for all three species, with cross references to the three different entries in PIR–International. In GenBank, objects are reported sequences, which may or may not correspond precisely with a gene or particular protein. GenBank may have hundreds or thousands of entries of genomic RNA, cDNA, DNA, or even individual exon sequences that relate in some way to human (ß–hemoglobin. In GDB objects include genes, probes, and polymorphisms. There will be one GDB entry for the (ß–hemoglobin *gene,* but multiple entries for associated polymorphisms and probes. In OMIM, objects are essays on inherited human traits, some of which are associated with one locus, some with multiple loci, and some whose genetic component (if any) is unknown.

The concept of "gene" is perhaps even more resistant to unambiguous definition now than before the advent of molecular biology. Our inability to produce a single definition for "gene" has no adverse effect upon bench research, but it poses real challenges for the development of federated genome databases.

Different community databases vary in the richness of their semantic concepts. GDB has more subtleties in its concept of a gene than does GenBank. GenBank's concept of nucleotide sequence is richer than that of other databases. To facilitate federation, participating databases should attempt to accommodate the semantics of other databases, especially when the other semantics are richer or more subtle.

In short, developing a federated information infrastructure will require more effort to ensure semantic consistency across participating systems. The use of controlled vocabularies and common–denominator semantics is important. Support for necessary coordination and communication must be provided by DOE and other agencies.

First published in: *Journal of Computational Biology,* 3: 173–190.

GENOME–WIDE INFORMATICS COORDINATION

Informatics efforts must be coordinated across the entire genome project. The usefulness of any informatics project, no matter how good, is limited if it does not integrate well with other related efforts. The costs of integrating existing, uncoordinated efforts can vary widely. For example, joining two railroads can be as easy as installing some connecting track or as difficult as replacing an entire track system if the roads employ tracks of different gauge. If coordination does not occur early in database development, linking them later can be as challenging as connecting different–gauge railroads. To avoid expensive refitting and to maximize the return on informatics investment, improved coordination, interagency communication, and planning are required.

Because of few timely publishing outlets for informatics work, increased opportunities for interactions among informatics developers and users are needed. The computer demonstrations at the 1993 Santa Fe DOE contractor and grantee workshop seemed useful, and many felt that similar, genome–wide NIH–DOE meetings might be organized. Data and software fairs at the Hilton Head sequencing meeting were suggested, as well as an NIH–DOE genome informatics workshop with informatics experts from genome centers and major databases.

Although such short–term interactions are valuable, even more important are opportunities for more extensive interactions among informatics practitioners, such as sabbaticals or other similar interchange between centers. Programs should be established to encourage this short–term exchange of personnel.

One person noted that the previous 5–year goals for genome informatics focused on the community domain and did not specifically address the core–support and local–user domain. The development of genome–specific groupware could be important and may not be receiving enough emphasis. Dispersed collaborations may be needed to complete both mapping and sequencing. Robust connections may be needed for efficiency in linking various mapping centers to sequencing centers.

DOE needs to increase the rate of software implementation and consider strategies for funding more. Funding should be considered for maintaining resource databases and servers at various sites. DOE should also consider asking for proposals that integrate diverse existing software into common sets of tools.

Discussion continued about how supercomputer centers may assist genome and structural biology informatics. Although some of these centers may not be effective for the bioinformatics community, their

state–of–the–art technology and emerging role as "national computer rooms" makes them of continuing interest.

TECHNOLOGY TRANSFER AND SOFTWARE SHARING

Although there was some discussion of technology transfer and software sharing, this area will need further discussion, clarification, and perhaps formulation of action items.

Allowing for open and sharable systems and designs, when possible, is an important goal. At present, there are varying degrees of transfer and sharing, ranging from exchanging experiences and lessons learned in software and database development to sharing source code and schemas. However, informatics core–support systems (such as databases and software tools) developed at one genome center often cannot easily be transferred to other centers. Experimental work specifics are usually deeply embedded in software designs and database schemas.

From a strictly local perspective, developing specific, nongeneric systems is usually more cost–effective and may lead to more rapid support of local end–users. In addition, requests for assistance in code porting or in providing off–site user support can be disruptive to local informatics activities. At genome centers, the priority for informatics core support should be to support local biologists and their specific needs. However, other informatics projects have been explicitly funded to provide resources to the wider community and produce nongeneric results that can be readily shared.

Clearly, the entire community and individual centers will sometimes need to exchange data and analytical methods, and the need for this sharing and data–flow integration will only increase as HGP progresses. Thus, OHER/DOE should consider addressing the growing need for standardized, integrated, sharable software–analysis tools. A simple first step was suggested to encourage one or more sites to establish servers that integrate a suite of different analysis tools developed by different research projects. Some effort would be required to integrate these diverse tools into something resembling a coherent suite. Also, the development of software libraries and resource listings could assist in making results of individual projects more widely available.

The group considered the role of industry and government–industry programs like Small Business Innovative Research (SBIR), CRADAs, and the Advanced Technology Program (NIST/Commerce) in bioinformatics research and results. While recruiting industrial

First published in: *Journal of Computational Biology,* 3: 173–190.

partners may be essential for long–term success, directing these efforts effectively toward informatics research presents problems. Useful genome–related software seems best developed in close conjunction with bench research, and commercial efforts should bear this general observation in mind. Should commercial partners become involved in systems development, intellectual property issues may become even more important.

MISCELLANEOUS TOPICS

A few other subjects were discussed that are relevant to this summary but do not fit into other sections of the report. They are included here for completeness.

Productivity effects

The quality of a site's local informatics efforts can affect its bench–research output. Although costly, robust systems integration of present sequencing technologies and the automation of information flow at a center can increase sequencing throughput even without drastic changes in sequencing technology. Because the Human Genome Project must maximize its overall return on investment, genome centers must consider informatics carefully when planning local budgets.

Training needs

Human resources are still too often a limiting factor. Staff with joint biological and computational expertise are a great asset but are in very short supply. Training programs, particularly institutional training grants that permit sites to develop courses, support students, *etc*., are necessary to produce the multidisciplinary people who support bioinformatics. Individual fellowships should be maintained or increased.

DOE informatics review

In 1982, an independent programmatic peer review of the OHER/DOE genome informatics research was conducted. This review focused on the program as a whole and thus differed from standard peer review of individual research projects and centers. The review extended over a week, occurred at four different locations, and included both site reviews and reverse site visits. Clear benefits resulted from this in–depth and completely independent review of genome informatics.

Over the long term, such in–depth, extensive reviews create a far stronger overall enterprise. External program–wide independent peer review is essential to ensure the best bioinformatics research and development and the proper allocation of scarce resources.

Interactions with other organizations

A discussion on how private commercial organizations might assist genome bioinformatics highlighted several concerns, including: (i) These groups have tended to work on large projects where fairly stable specifications can be drawn up. Genome informatics, on the other hand, is still rapidly evolving. Since costs rise dramatically when specifications change, such organizations may find participation difficult until more stability in requirements is achieved. (ii) Maintaining the interest of these organizations may be difficult in the long run, because genome budgets are quite small compared with defense and other projects familiar to these organizations. Also, genome funding of peer–reviewed, investigator–initiated projects may be an unfamiliar mechanism to the groups. (iii) Making a significant contribution to genome informatics requires a good understanding of genome biology. Training requirements will be considerable for these groups.

SPECIFIC RECOMMENDATIONS

Several specific recommendations emerged from the discussions. Some are conceptual injunctions and others call for specific actions from participating informatics projects. Some call for specific actions from DOE or other funding agencies. The recommendations are presented below. The overall recommendation was: Successful HGP data management will require the development of a federated information infrastructure, with data flowing electronically over networks from producers to databases to users. Achieving this must be a top priority.

- In the HGP, community databases must be recognized as tools that support research, not as mere archives for the storage of information.
- The use of the Internet must be assumed for genome informatics. Agencies must support improved access to the Internet for genome researchers. Informatics projects should not be forced to implement technology to accommodate those without Internet access.

First published in: *Journal of Computational Biology,* 3: 173–190.

- Community databases should be available over the Internet through stable and well–documented APIs.
- Improved support for complex, multi–database queries is needed. This will require major efforts toward improving integration and interoperability of community databases. Community database providers should work to meet these needs, and agencies should support their efforts. Incentives must be provided to the operators of community databases for their local efforts to ensure database interoperability.
- Support must be given to improving data quality. Professional data curators should be supported for community databases and, in addition, support for direct author curation should be developed.
- Increased communication and interaction among informatics practitioners must be encouraged and supported including more opportunities at meetings and support for "sabbatical" work at other sites.
- Informatics efforts across the HGP (and beyond) must be coordinated. DOE and other agencies should exert leadership and provide support for achieving this integration.
- Access to on–line citation data is needed. DOE should work with the National Library of Medicine and other organizations to ensure that this can be developed.
- Direct data submission into community databases must be recognized as the preferred form of data entry. Direct data submission must be encouraged, perhaps through the mandatory linking of submission with publication for all genome data, not just sequences.
- Professional data curators should be supported for community databases and, in addition, tools for direct author curation should be developed.
- Steps must be taken to reduce the human resources shortfall in bioinformatics expertise.
- We must begin to think of the computational infrastructure of genome research as a federated information infrastructure of interlocking pieces, including both data resources and analytical tools. Minimum interoperability standards must be defined, so that adding a new participating project will be no more difficult than adding another computer to the Internet.

- Efforts should be made to integrate independent analytical software tools into application suites. These should work readily with community databases.
- Any biologist should be able to submit research results to multiple appropriate databases with a single electronic transaction.
- Genome information resources should be embedded in a larger federated information infrastructure for all of biology. Working with other agencies and the community, DOE should exert leadership and provide support for achieving this integration.

First published in: *Journal of Computational Biology,* 3: 173–190.

APPENDIX I:

SAMPLE QUESTIONS FOR A FEDERATED DATABASE

Continued HGP progress will depend in part upon the ability of genome databases to answer increasingly complex queries that span multiple community databases. Some examples of such queries are given in this appendix.

Note, however, until a fully atomized sequence database is available (*i.e.*, no data stored in ASCII text fields), none of the queries in this appendix can be answered. The current emphasis of GenBank seems to be providing human–readable annotation for sequence information. Restricting such information to human–readable form is totally inadequate for users who require a different point of view, namely one in which the sequence is an annotation for a computer–searchable set of feature information.

- Return all sequences that map `close' to marker M on human chromosome 19, are putative members of the olfactory receptor family, and have been mapped on a contig map of the region; return also the contig descriptions. (This is nominally a link between GenBank, GDB, and LLNL's databases.)
- Return all genomic sequences for which *Alu* elements are located internal to a gene domain.
- Return the map location, where known, of all *Alu* elements having homology greater than "h" with the *Alu* sequence "S."
- Return all human gene sequences, with annotation information, for which a putative functional homologue has been identified in a nonvertebrate organism; return also the GenBank accession number of the homologue sequence where available.
- Return all mammalian gene sequences for proteins identified as being involved in intracellular signal transduction; return annotation information and literature citations.
- Return any annotation added to my sequence number #### since I last updated it.
- Return the genes for zinc–finger proteins on chromosome 19 that have been sequenced. (Note that answering this requires either query by sequence similarity or uniformity of nomenclature.)

- Return the number and a list of the distinct human genes that have been sequenced.
- Return all the human contigs greater than 150 kb.
- Return all sequences, for which at least two sequence variants are known, from regions of the genome within ± one chromosome band of DS 14###.
- Return all publications from the last 2 years about my favorite gene, accession number X####.
- Return all G 1/S serine/threonine kinase genes (and their translated proteins) that are known (experimentally) or are thought (by similarity) also to exhibit tyrosine phosphorylation activity. Keep clear the distinction in the output.

First published in: *Journal of Computational Biology,* 3: 173–190.

APPENDIX II:

MEETING ATTENDEES

Author/participants:

Helen Berman, Rutgers University
Elbert Branscomb, Lawrence Livermore National Laboratory
Peter Cartwright, University of Utah
Michael Cinkosky, Los Alamos National Laboratory
Dan Davison, University of Houston
Kenneth Fasman, Johns Hopkins University
Chris Fields, Institute for Genome Research
Paul Gilna, Los Alamos National Laboratory
David Kingsbury, Johns Hopkins University
Thomas Marr, Cold Spring Harbor Laboratory
Robert J. Robbins,* Johns Hopkins University†
Thomas Slezak, Lawrence Livermore National Laboratory
F. Randall Smith, Baylor College of Medicine
Sylvia Spengler, Lawrence Berkeley Laboratory
Ed Uberbacher, Oak Ridge National Laboratory
Michael Waterman, University of Southern California

Agency Observers:

David Benton, NIH/NCHGR
John Wooley, DOE/OHER
David Smith, DOE/OHER
Jay Snoddy, DOE/OHER

* Attended portions of the meeting.

† Now on leave from JHU, serving as Bioinformation Infrastructure Program Director at DOE/OHER.

APPENDIX III:

RECOMMENDATIONS FROM READERS

A draft of this document and a call for comments was in general circulation from September, 1993, through March, 1994. The final draft of the report has benefitted from the comments provided by many readers. Several interesting suggestions went beyond the scope of the original meeting and therefore could not be accommodated within the meeting report itself. Some examples of these suggestions are given here.

- Meaningful analyses of genomic data requires that the biological source of the genomic material used in producing the data is well documented. Most of the "official" model organisms for the genome project have well–recognized stock centers, with associated databases. Materials obtained from these stock centers, from service culture collections such as ATCC, and from research collections can be documented by name and by unique accession number. Other biological materials, not as easily documented, require the online availability of taxonomic and terminological databases to which organisms and their components may be mapped.
- The report does not note that its recommendations span a wide range of difficulty, from relatively straightforward implementation using current technology to outstanding research challenges in computer science. If the more difficult challenges are to be addressed, the genome project should direct some of its informatics support toward more basic computer science research.
- The report omits any mention of knowledge representation (KR) technology, which is a very useful approach to information management. KR technology may be the best current technology for building the type of shared data models discussed in the report.
- Current researchers in genome informatics rarely publish on the details of their methodology or on the lessons learned from their research. This reticence hinders the development of the field and efforts should be made to secure better exchange of general ideas in bioinformatics. Regular workshops that emphasize general principles and methodology would be

First published in: *Journal of Computational Biology,* 3: 173–190.

helpful, especially in the development of improved data models and better user interfaces.

- Computer scientists interested in genome informatics would benefit from access to more general descriptive material about the computational needs and computational activities of the genome project. Directories of researchers and facilities would be helpful, as would collected bibliographies of informatics publications and meeting reports.
- Efforts to develop a long–term strategic plan for genome informatics should be encouraged. Among other things, such an effort might involve interactions with vendors to ensure that coordinated feedback on genome informatics needs could be communicated to commercial software vendors.
- A common misunderstanding is that the lack of exchange formats has significantly hampered database intercommunica–tion in genome informatics. In practice, the semantic differences among the databases greatly overshadow issues of physical data *format.* Having the data expressed in the same format is of limited use if the conceptual database schemas are incompatible. Practical experience shows that the major effort is involved in understanding and cross–mapping semantic differences between databases and in developing tools to transform the data accordingly (often this cannot be done in a fully automated manner). In comparison, the effort to develop syntactic parsers is insignificant. Currently, it is not possible for groups not associated with a particular database to understand the semantics of that database well enough to cross–map semantic content accurately.
- While the report supports increasing development of client–server facilities of the Internet, there is little mention of widely used public–domain systems such as WAIS, Gopher, Mosaic, or other such systems. These systems can be set up and maintained for relatively little cost and their use in support of genome informatics should be encouraged.
- Many community databases are significantly under–funded, given the expectations of the community. If a proper information infrastructure for biology is to be developed, appropriate levels of support must be identified and maintained.
- Direct submission of data will not entirely replace published literature as a source of data input until the bench scientist

approaches the publication of a data submission with the same care as the preparation of a traditional publication.

Note added in proof. – Prior to this meeting, the GenBank nucleotide sequence database effort in the United States consisted of two components: one operated by the National Center for Biotechnology Information (NCBI) and charged with overall management and with the distribution of database materials to the general biological community, and the other operated at the Los Alamos National Laboratory and charged with accepting and processing submitted data and with building the database itself.

Since this meeting, these two components have begun to operate separately. NCBI has retained the GenBank name and has added the processing of data submissions to its activities. The group in New Mexico are now operating under the name Genome Sequence DataBase (GSDB) and have added the development of tools and procedures specifically aiming at meeting the needs of the Human Genome Project to their activities.

These actions will help alleviate some of the concerns outlined in this meeting report and they also represent movement toward meeting the recommendations of the 1988 National Academy study (Committee on Mapping and Sequencing the Human Genome, 1988, Mapping and Sequencing the Human Genome. Washington, DC: National Academy Press) that specifically called for the creation of a dedicated sequence database to meet the needs of the genome community. Steps have been taken to ensure that these changes will not adversely affect the activities of the international collaboration now overseeing the development of coordinated sequence databases worldwide.

Address reprint requests to:

*Dr. Robert J. Robbins**
Bioinformation Infrastructure Program
Office of Health and Environmental Research
ER–72 GTN
U.S. Department of Energy
Washington, D.C. 20585

Received for publication June 7, 1994; accepted June 14, 1994.

* ESP NOTE: Current (Jan, 2000) address: Fred Hutchinson Cancer Research Center, 1100 Fairview Ave, North, J4-300, Seattle, WA 98109; also rrobbins@fhcrc.org

First published in: *Journal of Computational Biology,* 3: 173–190.