Invited manuscript: submitted to *IEEE Engineering in Medicine and Biology* for a special issue on genome informatics.

INFORMATICS AND THE HUMAN GENOME PROJECT

Robert J. Robbins
US Department of Energy
robbins@er.doe.gov

David Benton
National Center for Human Genome Research
benton@nchgr.nlm.nih.gov

Jay Snoddy
US Department of Energy
snoddy@er.doe.gov

TABLE OF CONTENTS

Invited manuscript: submitted to *IEEE Engineering in Medicine and Biology* for a special issue on genome informatics.

INFORMATICS AND THE HUMAN GENOME PROJECT[1]

ROBERT J. ROBBINS, DAVID BENTON, AND JAY SNODDY

INTRODUCTION

Information technology is transforming biology and the relentless effects of Moore's Law (discussed later) are transforming that transformation. Nowhere is this more apparent than in the international collaboration known as the Human Genome Project (HGP). Before considering the relationship of informatics to genomic research, let us take a moment to consider the science of the HGP.

It has been known since antiquity that like begets like, more or less. Cats have kittens, dogs have puppies, and acorns grow into oak trees. A scientific basis for that observation was first provided with the development of the new science of genetics at the beginning of this century. More recently, the techniques of molecular biology have shown that information is passed from parent to offspring in the form of large molecules of deoxyribonucleic acid, or *DNA*. DNA contains four different kinds of subunits (known as *nucleotides*, or *bases*) arranged in a linear order. The individual bases are generally indicated by the first letters of their full chemical names adenine, thymine, cytosine, and guanine: A, T, C, or G.

A molecule of DNA is actually two different, but complementary and fully redundant, linear sequences of nucleotides.

```
· · ·→ T–A–G–C–C–G–C–A–T–A → · · ·
       | | | | | | | | | |
· · ·← A–T–C–G–G–C–G–T–A–T ← · · ·
```

Whenever "A" appears in the top strand, a "T" appears in the lower strand, and vice versa. The same relationship holds for C's and G's. Because the two strands are completely redundant, all of the information in a given DNA molecule can be expressed by writing out the sequence of only one of the two paired strands.

Information in DNA is stored in the form of a digital code, a sequence of pairs of A's, T's, C's, and G's, that is read out by the cell's molecular machinery to control the synthesis of new molecules. Functional blocks of sequence are called

[1] The ideas in this paper are the opinions of the authors and do not necessarily represent the views of the US Department of Energy, the National Institutes of Health, or of any other Federal agency.

genes. If DNA molecules are the mass–storage devices of life, then genes are the files on those devices.

At conception, each organism begins as a single cell, carrying one full copy of digital instructions from its mother, one from its father. The full set of information obtained from both parents is known as the organism's *genome* (or sometimes *diploid genome*, to distinguish this double set of instructions, from the single, or *haploid*, set provided by each parent individually). As the programs in the genome are decoded and executed, the single cell becomes either a cat, or a dog, or an oak tree, according to the information stored in the genes.

That genes must act as some sort of encoded instruction set was recognized well before DNA was known to be the hereditary material [16]:

> [The] chromosomes ... contain in some kind of code–script the entire pattern of the individual's future development and of its functioning in the mature state. ... [By] code–script we mean that the all–penetrating mind, once conceived by Laplace, to which every causal connection lay immediately open, could tell from their structure whether [an egg carrying them] would develop, under suitable conditions, into a black cock or into a speckled hen, into a fly or a maize plant, a rhodo–dendron, a beetle, a mouse, or a woman.

Schrödinger's essay, *What is Life*, is credited by many with stimulating the interests of those who established molecular biology in the 1950's. (James Watson has written [22]: "As an undergraduate at Chicago, I had already decided to go into genetics... Population genetics at first intrigued me, but from the moment I read Schrödinger's 'What is Life' I became polarized toward finding out the secret of the gene.") Schrödinger's suggestion that the code–script must, in principle, be understandable provides the intellectual underpinnings of genome research even today.

Genome projects are essentially efforts to obtain and reverse engineer the DNA code–script of humans and other species. Sequencing the genome, then, is equivalent to obtaining an image of a mass storage device. Mapping the genome is equivalent to obtaining a file allocation table for the device. Understanding the genome, when it happens, will amount to reverse engineering the genetic programs all the way back to design and maintenance specs.

Research and technology development efforts aimed at sequencing and mapping entire, or large portions of, genomes are known as genome projects. Although the singular term "The Human Genome Project" is often used, in fact there is no single human genome project, but instead many different projects carried out at different facilities around the world. However, great efforts have been made to coordinate these many projects worldwide, especially those operating with governmental sponsorship, so the notion of a singular international HGP is not entirely unreasonable.

Within the United States, two agencies have primary responsibility for supporting research relevant to the Human Genome Project (which includes work on selected model organisms such as mice, fruit flies, nematode worms, yeast, and bacteria): the Office of Health and Environmental Research of the Department of

Energy (DOE) and the National Center for Human Genome Research (NCHGR) of the National Institutes of Health (NIH). The United States Department of Agriculture (USDA), the National Science Foundation (NSF), and DOE also support genome research on other organisms.

Information and Genome Projects

By their nature, genome projects involve the generation and management of large amounts of highly complex, interrelated information. Just one human haploid genome (i.e., the DNA in a single sperm cell) contains over 3 billion base pairs of DNA. Typed out in 10–pitch type, this sequence would be thousands of miles in length. Keeping track of functional and other annotations associated with regions of DNA sequences increases the information–handling requirements greatly. For example, the gene for human β–hemoglobin is only a few thousand base pairs in length, yet nearly a megabyte of text about β–hemoglobin and its role in human biology is stored in sequence, map, and functional databases collectively.

Handling all of this information, especially the complex and often provisional relationships among pieces of the information, cannot be done without appropriate information technology. From the beginning, the importance of information management to genome research has been clearly recognized. In 1988, for example, the Office of Technology Assessment (OTA), at the request of the House Committee on Energy and Commerce, carried out a study on the feasibility of genome projects. The OTA report [5] of that study placed data management first in the list of objectives of genome projects:

> Genome projects have several objectives:
>
> - to establish, maintain, and enhance databases containing information about DNA sequences, location of DNA markers and genes, function of identified genes, and other related information;
> - to create maps of human chromosomes consisting of DNA markers that would permit scientists to locate genes quickly;
> - to create repositories of research materials, including ordered sets of DNA fragments that fully represent DNA in the human chromosomes;
> - to develop new instruments for analyzing DNA;
> - to develop new ways to analyze DNA, including biochemical and physical techniques and computational methods;
> - to develop similar resources for other organisms that would facilitate biomedical research; and possibly
> - to determine the DNA sequence of a large fraction of the human genome and that of other organisms.

In addition, the report noted that many of the management challenges facing agencies that support genome research are problems of resource allocation:

> Most issues that need to be addressed regarding genome projects are variations on the problem of the commons: how to create and maintain resources of use to all.

> ... The core issue concerning genome projects is resource allocation. What priority should be given to funding databases, materials repositories, genetic map projects, and development of new technologies?

A special OTA workshop on the costs of genome of genome research recommended that at least fifteen percent of the genome budget be dedicated explicitly to informatics, "*in addition to* continued support of existing databases and computer facilities." [5, Appendix B]

A National Research Council study [20] was more explicit, asserting that much of the value of genome research would *depend* upon proper information management

> Considerable data will be generated from the mapping and sequencing project. Unless this information is effectively collected, stored, analyzed, and provided in an accessible form to the general research community worldwide, it will be of little value.

and it called for the creation of new multi–million–dollar map and sequence databases, specifically designed to meet the needs of genome research.

Thus, even before DOE and NCHGR support for the HGP officially began, there were explicit recommendations from high–level advisory committees that appropriate information management and analysis would be crucial to HGP success. These were accompanied with explicit recommendations that funding agencies commit significant resources to the creation, operation, and maintenance of appropriate facilities and software.

The Nature of Information Technology

In the decade since these recommendations were drafted, major changes have occurred in information technology (IT) and studies abound that describe how IT is transforming the way institutions work (e.g., [6], [9], [17], [19]). IT reduces the effects of distance, time, and complexity in both the *performance* and the *management* of tasks.

No other infrastructure technology combines the ability to carry out activities with the ability to assist in their organization and operation. In genome research, for example, computers assist in the carrying out of research (robotics), in the analysis of results (map or sequence assembly software), in the management of reagents (inventory control), in the integration of findings from multiple sites (community databases), and in the preparation of research publications (word processing, statistical analysis, graphics design).

Computer hardware regularly shows an annual cost–performance improvement on the order of 30–50%. Over one year, such an improvement is merely a convenience. Compounded over a decade or more, however, such exponential changes profoundly change the way IT can affect genome research, or any other human activity. Such is the relentless effect of Moore's Law.

Moore's Law

In the late 1970's, Gordon Moore, co–founder of Intel, observed that technological improvements allowed a doubling, every eighteen months, of the number of transistors that can be placed on a chip. This has held true for nearly two decades, resulting in a sustained exponential improvement in the cost–performance ratio of computing equipment. Doubling the transistor count every year and a half is equivalent to a 58% per year improvement in the cost/performance ratio of computer hardware. Intel itself has shown about an average 50% improvement per year over the succession of its CPU chips. Similar results are seen with other architectures.

Such exponential improvement in cost performance means either a tremendous improvement in performance for constant cost, or an equally dramatic decline in cost for constant performance (Figure 1).

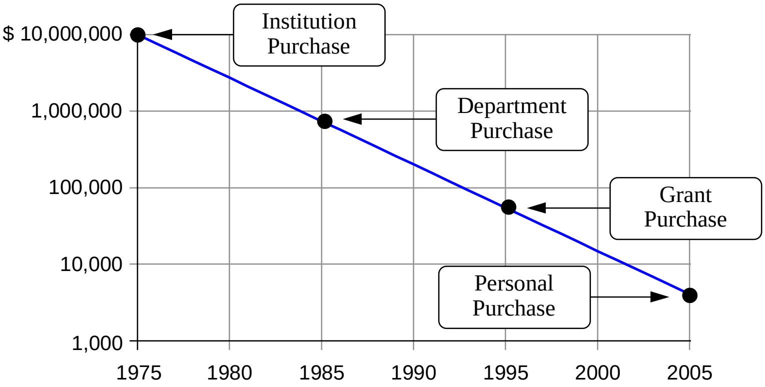

Figure 1 Relative cost, over time, of the same amount of computing power. Over time, computing resources once affordable only by large institutions become so inexpensive that they may be acquired even by individuals of modest means.

In 1975, a major research university would have had to spend perhaps 10 million dollars to obtain a state–of–the art central computer system. With Moore's Law driving down cost, by 1985, that same computing power was within the purchasing reach of a major research department. Now, the price has dropped to less than $100k, allowing such systems to be acquired with a single–investigator grant. By 2005, if current trends continue, the price will be low enough that a researcher could consider acquiring one personally, out of pocket.

Although Moore's Law reduces hardware costs, the majority of information–management expenses stem from software development and data–acquisition. Moore's Law can also lead to reductions in the cost of software, either directly by allowing programming to be done at a higher level of abstraction, or indirectly as inexpensive hardware allows commercially produced software to reach a larger market. The combination of less expensive hardware and mass–market–priced software can greatly reduce the overall costs of implementing data management systems.

At the same time, the costs of custom software development remain high. Taken together, this suggests that, to be maximally cost effective, genome informatics should (a) rely upon commercial software whenever practical, (b) move to less expensive hardware platforms when feasible, and (c) reduce the overall costs of custom software development through shared efforts and the implementation of interoperable component systems.

INFORMATICS

Sometimes "bioinformatics" and "computational biology" are used almost interchangeably. Other times they are used to distinguish data management (informatics) from data analysis (computational biology). In this essay, however, *bioinformatics* (or informatics) will be used generally to refer to the entire collective of information–management systems, analysis tools, and communication networks that support biology in general and genome projects in particular.

Although bioinformatics (the application of computers to biological information management) is part of the infrastructure that supports biological investigations, it is not just another infrastructure component, no more deserving of special consideration than, say, biomicroscopy (the application of magnification to biological investigations). Instead, the need for interoperability among different projects makes informatics a special case.

With the spread of global networking, biological information resources, such as community databases and analytical tools, must be capable at some level of working together, of *interoperating*, so that researchers may interact with them collectively as a *federated information infrastructure*. In contrast, much enabling infrastructure for other science, such as particle accelerators or orbiting telescopes, may operate usefully as stand–alone facilities. Researchers interact with them, carry out work, and take the results back to their desks (or computers).

This requirement of interoperability means that mere excellence as a stand–alone facility is not good enough. Informatics projects must also be excellent components in a larger, integrated system. This can only be achieved as a result of coordination among those who develop the systems, among the professional societies and other advisory bodies that help guide the projects, and among the agencies that support the work. The required level of coordination in maintaining these facilities is much greater than that seen in most other sponsored research or research infrastructure activities.

Informatics Enables Big Science

Informatics has become an enabling technology, the technical *sine qua non*, without which big biology cannot be done. Informatics is also becoming a *sine qua non* for commercial biotechnology activities. For example, The Institute for

Genome Research (TIGR) reportedly spends more than 25% of its budget on informatics and Craig Venter of TIGR has asserted that informatics is now a limiting factor for large–scale sequencing.

Many pharmaceutical and other companies are investing tens or even hundreds of millions of dollars to create or to gain access to private databases of genomic information. The value of genomic data is now clearly recognized.

The international human genome project, recognized in the popular, scientific, and business press as "ahead of schedule and under budget" [10], exemplifies the importance of informatics to successful big–science biology projects. Most of the genome gains already made could not have been done without informatics support and much of the work remaining will depend upon further advances in the underlying informatics.

The Intellectual Standing of Informatics

Is informatics an intellectual discipline in its own right, or does it represent interdisciplinary research between biology and computer science, or is it merely some kind of applied computation? Evidence suggests that a new discipline of information science may be emerging from the interaction of domain sciences with computer science, with library and information science, and with management science. A recent workshop report [12] asserted a need for a new training discipline in informatics, and similar claims are increasingly seen in the business and technical literature.

Bio–informatics itself is neither computer science nor biology, occupying instead some middle ground. One might envision a conveyor belt carrying ideas from computer science (CS) to biology (Figure 2). The extensive refinement that occurs along the way is perhaps the essence of informatics as a discipline.

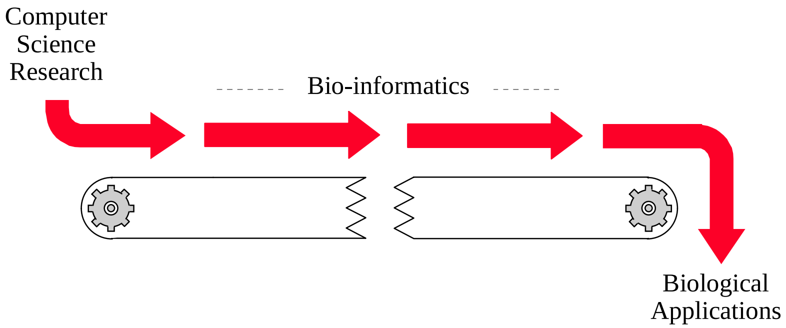

Figure 2. Informatics is the process by which the results of general research in computer science become transformed into applications of use to practicing biologists.

This refinement is increasingly informed by notions from library science and information science, with their expertise in making information resources usefully available. As informatics projects become larger, systems analysis and management science play increasingly significant roles. Informatics is similar to engineering, in that it involves the scientific application of known principles to solve real problems under constraints of both budget and time. (Figure 3)

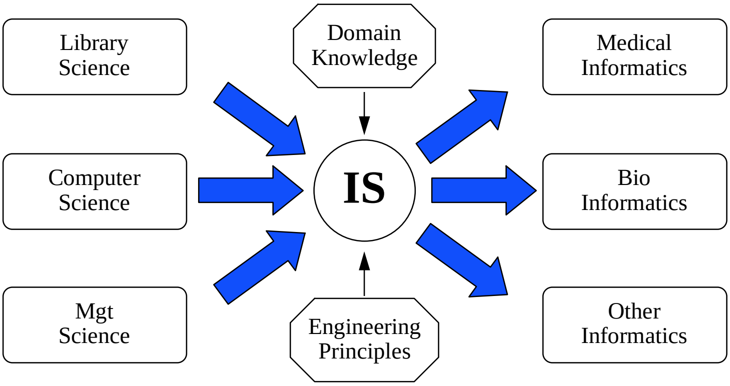

Figure 3. Information Science (IS) lies at the crossroads connecting a variety of information–management fields (library science, computer science, and management science) with other disciplines. Activities at the crossroads must be significantly informed by the domain knowledge of the subject area (e.g., biology) and by the techniques and principles of quality engineering. Domain–specific informatics are the result—medical informatics, bio–informatics, etc.

Information science, should it emerge, would likely be similar to statistics or engineering, in that it would train a mixture of practitioners and theoreticians. The emphasis on working applications would enforce an engineering mind set.

AGENCY COMMITMENTS

Both DOE and NCHGR recognize the importance of quality informatics to accomplishing genome research and both agencies are committed to providing appropriate support for genome informatics. When funding for dedicated informatics activities is combined with that estimated as the informatics components of bench projects, the U.S. HGP currently allocates more than 20% of its budget to informatics.

Community Recommendations

Like other federal agencies that support basic research through competitive review, both NCHGR and DOE attend to advice from the scientific community, both in evaluating individual proposals and in setting guidelines and priorities for research and development programs. Summaries from a number of relevant advisory workshops and panels are presented here.

OTA Report

The OTA report helped set the stage for defining genome informatics needs. In particular, it emphasized the need for multiple, interoperable databases:

> The many types of information that are produced in molecular biology necessitate the maintenance of a variety of specialized databases. At the same time, however, the information in different databases must often be combined in order to understand the full dimensions of any specific research problem. It is crucial for the scientific community to be able to access information on a topic of interest from a variety of databases that may handle different aspects of the problem.

and for appropriate analytical tools that use high–end computing hardware:

> Development of analysis methods to search for and compare sequence information, to predict sequences that code for proteins and the structures of those proteins, and to aid in other aspects of the analysis of data from genome projects will eventually need to utilize parallel processing techniques and the capacity of supercomputers. Most researchers agree that the hardware to tackle the complex problems of sequence analysis and comparison already exists but that satisfactory software must be developed.

Baltimore White Paper

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, 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 report has recently been published [15] and will be discussed here only briefly. The report reiterated that HGP success depends upon quality informatics

> In the future, the success of the genome project will increasingly depend on the ease with which accurate and timely answers to interesting questions about genomic data can be obtained.

and noted that integration of analytical tools with information resources is needed.

> To use the information in community databases, users require software for analysis and synthesis. These should be incorporated into suites of software tools... 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.

Database and tool interoperability was scored as essential, with the added notion that ultimately biologists should be able to interact with genome information resources without having to know precisely where the data were located.

> As multi–database queries become more important, users will require access to integrated views of data from multiple databases. ... 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. .... Ultimately, biologists should not have to know where data are located.

The need for managerial coordination of genome informatics projects was noted and a vision for the future of genome informatics was provided:

> 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.

GeSTeC Report

At its January 1994 meeting, the National Advisory Council for Human Genome Research requested that NCHGR staff determine the informatics capabilities that will be required to support the mapping and sequencing research needed to achieve the goals published in “A New Five–Year Plan for the U.S. Human Genome Project” [4]. In addition, NCHGR was asked to determine current capabilities and those areas which will need additional research and development work in order to meet the requirements of the five–year plan.

As a first step in this process, the NCHGR invited Genome Science and Technology Center (GeSTeC) directors to a one–afternoon meeting with the goal of determining the informatics requirements of the genome project over the next five–year period. The meeting participants recognized the strong dependence of the design of data management tools and analytical methods on the mapping and sequencing experimental methods they are intended to support and further recognized the difficulty of predicting far in advance the experimental methods that would be employed by genome mapping and sequencing projects.

The report [7] gave highest priority to “software reuse and interoperability” and to “the integration of genome and genome–related databases.” The lack current of software reusability was seen as problematic:

> [W]ith perhaps one exception, each of the genome centers had built its laboratory database from scratch, but that even now medium–sized projects cannot take advantage of this investment by importing a database system (or components thereof) developed in a center.

The report did acknowledge that building reusable software is more difficult and more expensive than building single–use, stand–alone systems (cf. [1]), but even with that caveat, the attendees felt that increased efforts should be made at developing more reusable software components.

Meeting attendees noted that database integration among heterogeneous systems is still an open research question in computer science, but nonetheless hope was expressed that genome informatics facilities might take advantage of work done elsewhere to improve the connectivity among genome–relevant information resources. The report also observed that technical connectivity between databases merely *enables* data element cross referencing. Achieving data connectivity requires that specific information be present in all relevant databases. Locating and loading such data is neither trivial nor inexpensive.

Some outstanding problems with current databases and laboratory data management systems were noted by meeting participants:

1) richer (not simpler) data models are needed, whether relational, object, or ASN.1.

2) database evolution: database schemata need to evolve to accommodate new experimental approaches. This is not straightforward with current DBMS systems; there is a need for tools to facilitate database evolution.

3) databases to support large–scale sequencing efforts will be required. These will need to incorporate local data and data from multiple public sources.

4) computer–readable (as contrasted with human–readable) databases are required.

5) data quality should be a primary concern of the public databases.

Workshop on Database Interoperability

In June of 1994, a Workshop on Database Interoperability was held in Gaithersburg, Maryland, co–sponsored by DOE, The Institute for Genomic Research (TIGR) and MasPar. Attendance was limited to groups working on relational databases in order to focus the discussion on the practical issues of establishing cross–database query capabilities using SQL.

This meeting was held shortly after a similar meeting held at UC Berkeley which addressed infrastructure requirements and design considerations for a federation of botanical specimen databases. The presence at both meetings of representatives from major funding agencies underscores the increased recognition that no one database can accurately reflect all biological knowledge, nor can a centralized database system or single institution begin to address the explosion of biological information and the needs of the multiple constituencies which seek access to that information.

The report noted:

> The representation of biological information in multiple ... databases is a growing phenomenon. Databases of DNA and protein sequences, genetic and physical maps, biochemical data, phenotypes and strains, biogeographical data, museum collections information, and other types of data already exist; many others are under development. As the volume of information electronically available has skyrocketed, new thinking about distributed information systems has generated a strong interest in implementing software tools and database structures to enable true database interoperability.

Meeting on Interconnection of Molecular Biology Databases (MIMBD)

Another meeting on database connectivity took place at Stanford University in August 1994. The meeting brought together biologists, computer scientists, and bioinformatics researchers interested in the problem of interoperation of the growing number of distributed, heterogeneous databases and knowledge bases that serve molecular–biology and genome researchers:

The premise behind this meeting was that the roughly 100 existing molecular biology databases would be of much greater value to molecular biologists when interconnected than in their current isolated states.

Interoperable Software Tools

Although the majority of advisory meetings and workshops over the past five years have addressed problems of database interoperability, there is a growing recognition that tool interoperability and software re–use are also important. David States of Washington University organized a discussion of these issues in conjunction with the annual Cold Spring Harbor meeting in 1994 and follow–on discussions, led by Ed Uberbacher of Oak Ridge National Laboratory were held at the National Center for Genome Resources during the DOE genome contractors meeting in Santa Fe in late 1994. Nat Goodman of the Whitehead Institute in Boston is emerging as a leading spokesman for a *componentry* based approach to genome informatics.

Summary Needs

Common concerns are apparent in the reports of these workshops and advisory bodies: interoperation of databases and software tools is essential for continued advances in molecular biology and genomics and the development of more generic, sharable, reusable information–resource components is greatly needed.

Achieving tightly coupled database interoperability among heterogeneous systems is an active research area in computer science (see reviews in [8], [18]), with some workers expressing doubt that many of the current research programs have much likelihood of practical success [2]. Building software that is to function as a component in a larger programming environment is recognizably more difficult than building stand–alone systems [1], but it is clear that efforts must be made in this direction.

As similar problems are faced by other funding agencies in non–genome areas, a government–wide approach to reconsidering methods for supporting information infrastructure might be useful.

Current Trends

Although tightly coupled database interoperability is an unsolved problem for heterogeneous systems, recent experience with gopher, World–Wide Web, and Mosaic has shown that loosely coupled read–only systems can provide tremendous utility, as evidenced by their ability to attract users. Since WWW technology first became available in early 1993, WWW traffic on the NSFnet backbone has increased almost 20,000–fold, while overall traffic has gone up only four–fold.

Many genome sites now use WWW technology to distribute their findings, and the built–in capability for inter–server cross–referencing has allowed the development of considerable utility at relatively little cost. The use of middleware approaches can greatly leverage the value of WWW resources. For example, Johns Hopkins is now providing WWW access to a sequence–analysis package running in Oak Ridge, with the results of the analysis being returned as a WWW document with live hot links to all referenced objects from other databases.

Future Support

As the HGP continues, the need for focused, interacting and interoperable informatics activities will increase, requiring a change in proposal evaluation criteria. Early on, genome informatics projects were judged, in part, by the following informal criteria: (a) is there a need for such an activity, (b) will the proposed activity meet that need, (c) can the applicants deliver the project on time and within budget, and (d) is it worth it? Now, additional criteria must also be attended: (a) does it adhere to standards, (b) will it interoperate, and (c) is there commitment to federation?

The guiding principles behind future informatics work may prove to be:

- A *global value explosion* occurs when multiple information resources are interconnected. Therefore, interoperability is a high priority.
- Sharable, reusable software must become a part of the genome informatics culture. A commitment to *componentry* is essential.
- Projects must not only work in pilot mode, but must also scale gracefully with an exponential increase in data volume and user access. Plans for scalability must be built in to genome informatics work, and must include both *technical scalability* (systems must be able to grow without failure or major loss of efficiency) and *social scalability* (systems should be designed to avoid human–participation bottlenecks).
- *Anonymous interoperability*—interoperability that does not require that interacting partners know of each other's existence—is highly desirable, especially as it facilitates social scalability.
- Third–party value–adding activities greatly increase the value of information resources. Genome systems should be designed from the beginning to support such *value additivity*.

OPEN ISSUES

Many other yet unsolved technical and social issues in informatics need addressing. As the number of information resources grows, the problems first of *resource discovery* (how do I find data relevant to my needs) and then of *resource filtering* (how do I eliminate data not relevant to my needs) will grow. Better methods for organizing global, networked information resources will be required. Some solutions may develop from work on digital libraries, others from efforts to extend the current networking naming protocols to include information resources and individual data elements within those resources.

The problem of data standardization and data indexing will grow. A recent comparison of data in several gene map databases found over 1800 genes with the names of associated proteins and the protein's EC numbers. (EC number: a number derived from a system developed by the Enzyme Commission to identify catalyzed reactions. Since enzymes act as catalysts, they are often assigned the EC numbers associated with the reactions that they catalyze. This does not provide a unique identifier for a protein (some proteins catalyze more than one reaction and some reactions are catalyzed by more than one protein), but it does provide an unambiguous identification of its catalytic capabilities.) However, only a few hundred of those protein names matched the canonical name associated with the EC number given for the protein. Such inconsistencies will make collecting all relevant data from large electronic databases increasingly difficult.

New social processes affecting data resources will need to be developed. Databases are becoming a new scientific literature [3], [14]. The communication role of genome databases has been explicitly recognized by leading genome researchers in a recent review [11]:

> Public access databases are an especially important feature of the Human Genome Project. They are easy to use and facilitate rapid communication of new findings (well in advance of hard–copy publications) and can be updated efficiently.

Traditional publishing provides many functions beyond the simple communication of findings from one researcher to another. For example, print journals provide evidence of primacy, editorial oversight and thus quality control, citability of results, archival preservation, and many other functions. Libraries provide organization, classification, maintenance, and access functions for print literature. As databases become ever more literature–like, means for implementing those other functions will be needed. Professional societies should become increasingly involved, both to help guide the processes and possibly to offer the beginnings of scholarly electronic publishing. A 1994 meeting of representatives of various professional societies organized by FASEB suggests promising movement in this regard.

Several important policy issues relevant to genome informatics are yet unresolved. Intellectual property rights, data sharing, and information access will continue to need thought. Dealing with this across national borders, and thus across differing legal and social traditions will make the problem more challenging.

The best means for providing long–term support for information resources will need additional thought. If databases become more literature–like in their social role, perhaps they should become more literature–like in their means of support. But even if some databases become self supporting, there will likely remain long–term needs for government–supported resources. How should these be identified, and how should priorities be set? With databases now often supported by means similar to those for original bench research, there has historically been something of a first–come, first–served aspect to database support. It is not clear that this is the best means for allocating infrastructure resources.

At present, nearly all public information resources are operated independently, with very few funded by the same organization or sharing the same advisors. With the requirement of interoperability among these resources increasing dramatically, this will cause increasing difficulties. Coordinated international efforts to facilitate cooperation among informatics resources should help minimize those difficulties.

BIBLIOGRAPHY

1. Brooks, FP, Jr: *The Mythical Man–month*. Addison–Wesley Publishing Company, Reading, MA, 1982.
2. Chorafas, DN, and Steinmann, H: *Solutions for Networked Databases: How to Move from Heterogeneous Structures to Federated Concepts.* Academic Press, Inc., New York, 1993.
3. Cinkosky, MJ, Fickett, JW, Gilna, P, and Burks, C: Electronic data publishing and GenBank. *Science*, 252:1273–1277, 1991.
4. Collins, F, and Galas, D: A new five–year plan for the U. S. human genome project. *Science*, 262:43–46, 1993.
5. Congress of the United States, Office of Technology Assessment: *Mapping Our Genes–Genome Projects: How Big, How Fast?* Johns Hopkins University Press, Baltimore, 1988.
6. Daniels, NC: *Information Technology: The Management Challenge*. Addison–Wesley Publishing Company, New York, 1994.
7. GeSTeC Directors: Report: NCHGR GeSTeC Director's meeting on genome informatics, 1994. (Available electronically from Johns Hopkins WWW server, http://www.gdb.org/Dan/nchgr/report.html.)
8. Hurson, AR, Bright, MW, and Pakzad, SH (Eds.): *Multidatabase Systems: An Advanced Solution for Global Information Sharing*. IEEE Computer Society Press Los Alamitos, CA, 1994.
9. Keen, PG: *Shaping the Future: Business Design Through Information Technology.* Harvard Business School Press, Boston, 1991.
10. Koshland, DE, Jr: Ahead of schedule and on budget (editorial). *Science,* 266:199, 1994.
11. Murray, JC, Buetow, KH, Weber, JL, Ludwigsen, S, Scherpbier–Heddema, T, et al: A comprehensive human linkage map with centimorgan density. *Science,* 265:2049–2054, 1994.
12. National Science Foundation, Informatics Task Force (M. C. Mulder, Chair): *Educating the Next Generation of Information Specialists: A Framework for Academic Program in Informatics,* report of workshop held 4–7 November 1993 in Alexandria, Virginia.
13. Robbins, RJ: Database and computational challenges in the human genome project. *IEEE Engineering in Medicine and Biology Magazine*., 11:25–34, 1992.
14. Robbins, RJ: Biological databases: A new scientific literature. *Publishing Research Quarterly*, 10:1–27, 1994.
15. Robbins, RJ (Ed.): Genome informatics I: Community databases. *Journal of Computational Biology,* 3:173–190, 1994b.
16. Schrödinger, E: *What is Life*. Cambridge University Press, Cambridge, 1944.
17. Scott Morton, MS: *The Corporation of the 1990s: Information Technology and Organizational Transformation.* Oxford University Press, New York, 1992.

18. Sheth, AP, and Larson, JA: Federated databases systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22:183–236, 1990.

19. Tapscott, D., and Caston, A. *Paradigm Shift: The New Promise of Information Technology*. McGraw Hill, Inc., New York, 1993.

20. United States National Academy of Sciences, National Research Council, Commission on Life Sciences, Board on Basic Biology, Committee on Mapping and Sequencing the Human Genome: *Mapping and Sequencing the Human Genome.* National Academy Press, Washington, DC, 1988.

21. United States National Academy of Sciences, National Research Council, Commission on Physical Sciences, Mathematics, and Applications, Computer Science and Telecommunications Board, NRENAISSANCE Committee: *Realizing the Information Future: The Internet and Beyond.* National Academy Press, Washington, DC, 1994.

22. Watson, JD: Growing up in the phage group. In Cairns, J., Stent, GS, and Watson, JD (Eds.): *Phage and the Origins of Molecular Biology*. Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 1966.