Manuscript prepared at the request of James Edwards of the National Science Foundation for presentation to the Organization for Economic Coordination and Development, Paris.

# BIOINFORMATICS: ESSENTIAL INFRASTRUCTURE FOR GLOBAL BIOLOGY

**Robert J. Robbins**

Johns Hopkins University
US Department of Energy

rrobbins@gdb.org
robbins@er.doe.gov

# TABLE OF CONTENTS

Manuscript prepared at the request of James Edwards of the National Science Foundation for presentation to the Organization for Economic Coordination and Development, Paris.

# BIOINFORMATICS: ESSENTIAL INFRASTRUCTURE FOR GLOBAL BIOLOGY[1]

**Robert J. Robbins**

## INTRODUCTION

Bioinformatics[2] (the application of computers to biological information management) is part of the information infrastructure that supports biological investigations. However, bioinformatics is not just another infrastructure component, no more deserving of special consideration than, say, biomicroscopy (the application of magnification to biological investigations). Instead, bioinformatics is a special case, requiring coordinated attention by members of the research community, by representatives of professional societies, and by funding agencies.

With the spread of global networking, biological information resources, such as community databases, must be capable at some level of working together, of *interoperating*, so that users may interact with them collectively as a *federated information infrastructure*. In contrast, enabling infrastructure for other science, such as particle accelerators or orbiting telescopes, may operate usefully as essentially 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—bioinformatics projects must also be excellent

---

1 The ideas in this paper are the opinions of the author and do not necessarily represent the views of the US Department of Energy or of any other Federal agency.

2 The terms “bioinformatics,” “computational biology,” and “biological information infrastructure” are sometimes used almost interchangeably. In this essay, however, *bioinformatics* will refer to database-like activities, involving persistent sets of data that are maintained in a consistent state over essentially indefinite periods of time, *computational biology* will denote the use of algorithmic tools to facilitate biological analysis, and *bioinformation infrastructure* will mean the entire collective of electronic information-management systems, analysis tools, and communication networks that support biology.

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.

This essay presents a brief overview[3] of bioinformatics activities, calling attention to those aspects that would most benefit from coordinated international attention. Issues presented are drawn more from interactions with community researchers and from many recent reports of community workshops[4] on bioinformatics and much less from compiled statistics on supported activities. Several examples are drawn from the genome project, since it is a successful, large–scale international biological project with a major informatics component.

## BIOINFORMATICS IN THE UNITED STATES

Cataloging *all* bioinformatics activities in the United States, however large or small, is simply impossible, since the computer management of biological information is now sufficiently routine that many relevant activities occur as undocumented components of basic biological research and thus cannot be identified and tallied. Like any good infrastructure, much of bioinformatics is becoming invisible.[5]

What is clear, however, is that the largest, most ambitious biological projects at every US agency have explicit and essential informatics components. Examples include the National Biological Survey at the Department of Interior, Long Term Ecological Research at the National Science Foundation, Genome Projects at the Department of Energy, the National Institutes of Health, and the Department of Agriculture. None of these projects would be possible without bioinformatics and

---

3 A thorough examination of bioinformatics would require book–length treatment.

4 Workshops that influenced this report include (1) *Scientific Data Management*, Charlottesville, Virginia, March 1990, (2) *Data Management at Biological Field Stations and Coastal Marine Laboratories*, W. K. Kellogg Biological Station, Janurary 1992, (3) *Genome Informatics I: Community Databases*, Baltimore, Maryland, April 1993, (4) *Arabidopsis Database Requirements*, Dallas, Texas, June 1993, (5) *A Biological Survey for the Nation*, several locations, 1993, (6) *Brain Map '93*, San Antonio, Texas, December 1993, (7) *Infrastructure Requirements and Design Considerations for a Federation of Interoperable Botanical Specimen Databases*, June 1994, (8) *FASEB Meeting on Biomolecular Databases*, Bethesda, Maryland, June 1994, (9) *Interoperability of Biological Databases*, Gaithersburg, Maryland, June 1994, and (10) *Interconnection of Molecular Biology Databases*, Stanford, California, August 1994.

5 When an infrastructure is new, it may attract positive attention as a novelty. However, with maturity, it rarely attracts notice, except when not working.

all of them combine strong support for informatics activities as components of basic biological projects with explicit support for stand–alone community information resources.

Bioinformatics has become an enabling technology, the technical sine qua non, without which big biology cannot be done. Bioinformatics 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 has asserted that informatics is now the limiting factor for large–scale sequencing. Smith-Kline Beecham invested more than $100,000,000 in Human Genome Sciences (the parent organization of TIGR), apparently motivated at least in part by access to the intellectual property in TIGR's databases.

Biology is inherently an information–rich discipline, with a great need to maintain considerable information about specific biological entities such as clones, probes, ecosystems, locations, specimens, species, and even individual organisms. Biology's claim to special needs in information–management systems is real. In chemistry and physics all things of interest in a particular class (hydrogen atoms, electrons, quarks, etc.) are held to be genuinely, not metaphorically interchangeable. All living things, on the other hand, are truly unique, and the properties of individual living things are determined in significant part by the unique, frequently contingent historical events that happened to each of their unique ancestors.

The number of living things that now exist, that have existed, or that ever will exist is sufficiently small in relation to their information content, that we will never be able to apply some sort of law of large numbers so that they could be described in all interesting ways as essentially, if not actually, interchangeable items. Understanding biology will depend in part on managing information in a way that preserves the individuality of the subjects.

Business information management also requires attention to individuality (of customers, employees, products, etc.) and thus solutions to biological information–management problems are likely to be highly relevant in the commercial sector. All of the basic statistical tests, now applied in fields ranging from quality control in mass manufacturing to traffic analyses in transportation and communication, were originally developed to solve biological problems.[6] Investment in better methods for biological data management are also likely to yield general economic benefits.

The international human genome project, increasingly recognized in the popular, scientific, and business press as a success that is "ahead of schedule and under budget," exemplifies the importance of informatics to successful big–science biology projects. Most of the genome gains already made could not have been

---

6 Galton devised regression analysis to study the correlation between parents and progeny, Pearson developed chi-squared methods to study the distribution of different morphs in wild populations, and Fisher invented analysis of variance to tease apart factors affecting the inheritance of traits.

done without informatics support and much of the work remaining will depend upon further advances in the underlying informatics. The continuing success of all major genome research centers, from Genethon to the Sanger Centre to the Whitehead Institute to Lawrence Livermore National Laboratory, depends upon local bioinformatics projects and access to public data repositories.

## THE INTELLECTUAL STANDING OF BIOINFORMATICS

Is bioinformatics 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 informatics, or some new discipline of information science or information engineering may be emerging from the junction of domain sciences with computer science, with library and information science, and with management science. A recent workshop report[7] asserted a need for a new training discipline in informatics, and similar claims are increasingly seen in the business and technical literature.

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 necessary emphasis on working applications would enforce an engineering mind set.

Bioinformatics itself is neither computer science nor biology, occupying instead some middle ground between the two, with bits of other fields thrown in. One might envision a conveyor belt carrying ideas from computer science (CS) to biology: At the biology end, what falls off are biological applications, to be judged purely on their utility to immediate biological problems. What gets loaded on at the far end are basic CS research ideas. The extensive refinement that occurs in between is perhaps the essence of bioinformatics as a discipline.

This refinement is increasingly informed by notions from library science and information science, with their expertise in making information resources usefully available. As bioinformatics projects become larger, systems analysis and management science play increasingly significant roles in successful activities. Bioinformatics also has much in common with engineering, in that it involves the scientific application of known principles to solve real problems under constraints of both budget and time.

Bioinformatics projects often have trouble obtaining support. Work in the middle of the conveyor belt may seem too much on the CS side, without a visible guarantee of a biological payoff, to make it comfortable for purely biological-oriented agencies to fund it. Yet it may have too much in the way of application-

---

7 NSF Informatics Task Force (M. C. Mulder, Chair). 1993. *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.

driven aspects to make it comfortable for pure CS programs to provide support. Despite this, the advance of bioinformatics is essential for much of biology.

---

⇒ **Needed:** Coordinated, international efforts to develop better means for supporting worthy bioinformatics activities.

---

## THE CHALLENGE OF INTEROPERABILITY

Several recent reports, in areas ranging from herbarium data management to biological surveys to neuroanatomy information resources to molecular biology and genomics, have singled out information–resource interoperability as the biggest problem currently facing bioinformatics. In the 1980s, the databases were falling behind the rate of data production and a crisis of data acquisition was recognized (Lewin, 1986), with the problem in molecular biology especially acute. Figure 1 shows the growth in the world's sequence databases from the first release of GenBank to 1994.

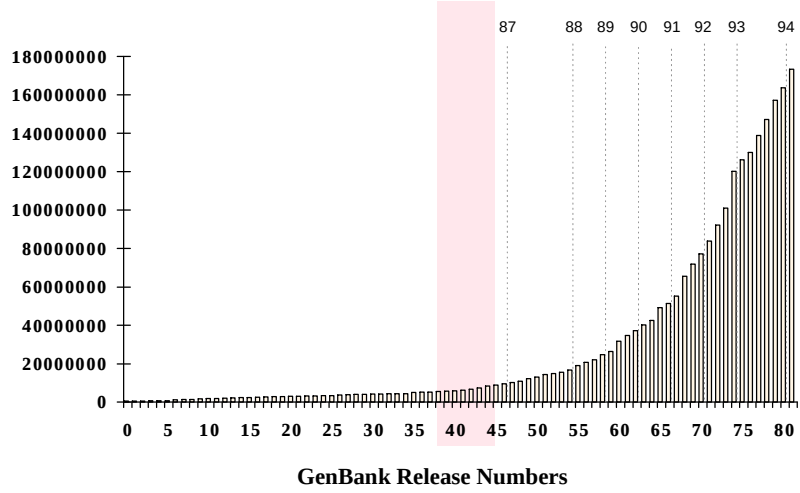


**Figure 1.** Growth in the world's collection of nucleotide sequence data, shown as the number of bases contained in every release of GenBank from 1 through 82. The numbers at the tops of the dotted lines show years (which do not necessarily coincide with a particular number of releases). The shaded bar in the middle represents the period in the mid 1980s when the data volume was, for a time, more than the databases could handle. (Data supplied by Michael Cinkosky and Dennis Benson.)

Although the data volume is still increasing exponentially, with a doubling time less than two years, merely keeping up is no longer a problem. Technical and

sociological advances now allow the databases to absorb easily a far greater amount of new information than previously conceivable. In 1986, 13 *months* elapsed between the publication of a sequence and its appearance in the databases. Now, the Genome Sequence Data Base processes a typical submission within 13 *hours*. Every 2-4 weeks, more sequence data enter the databases than did so in the first five years of their existence.

With the crisis of data acquisition resolved, we face a new crisis of data integration. Much of the value of the great masses of biological information now being compiled electronically will be lost, unless the data in one information resource can be meaningfully linked to relevant data in other resources: sequence data must be linked to map data; protein structures must be connected to metabolic function; species data must be connected with ecosystem data, and on and on.

## Database Interoperability

Today's crisis of data integration cannot be resolved through data consolidation (the collection of all relevant data in one facility), since the number of relevant information resources is large and growing.[8] Nor can it be solved by creating distinct, officially sanctioned subsets of data resources relevant to individual research areas, since it is simply impossible to identify a set of information resources that are all relevant to one, and only one, biological community (Figure 2).

Biological information resources dynamically group and regroup into transient overlapping collections of resources, with each collection being of special interest for some research discipline, or some individual researcher, at some time. As certain key databases (e.g., nucleotide sequence collections) play crucial roles in many such dynamic groups, physical or even administrative consolidation holds little prospect as a solution. Rather, advances will be required to allow autonomous data resources to interoperate productively. The challenge will be creating collections of data resources that are perceived by users to be functionally integrated, yet with each resource maintaining its autonomy, especially in the basic creation and maintenance of its data resources.

---

⇒ **Needed:** Coordinated, international efforts to further the development of a federated information infrastructure for biology.

---

8 The problems are as much social as technical: would a scientific community tolerate the requirement that all publication in a given field must occur in only one journal? As electronic biological publications become easier to build, we can expect a general increase, not a reduction, in their number.

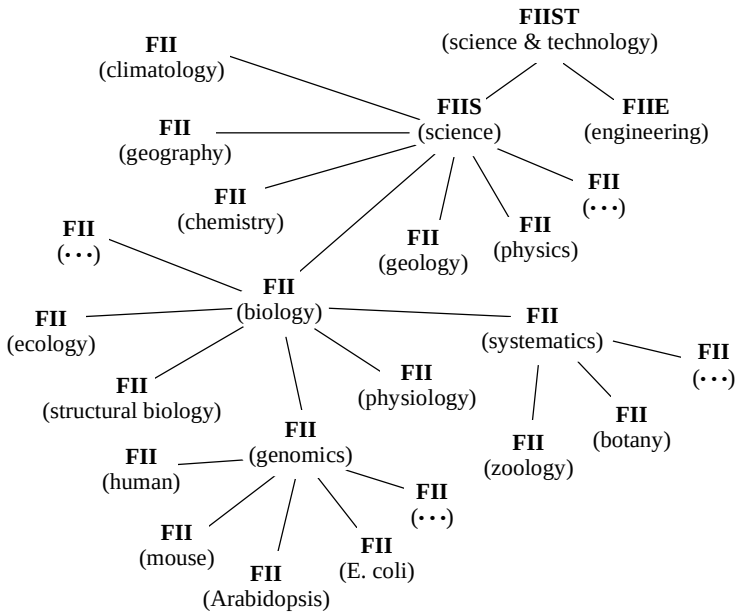


**Figure 2.** Many different fields in biology and other sciences are becoming increasingly dependent upon access to a coherent information infrastructure of electronically published text and data. Interoperability among different electronic resources is required, at least at the level of a loose federation. None of these subgroups is completely independent of any other, and this is true at all levels in the hierarchy. Understanding the genome will ultimately require integrating genome findings with protein structure (structural biology) and metabolic information (physiology). Comparative genomics involves systematics and other areas of comparative biology. This non-independence can involve merely the need to cross reference objects in other databases or the mutual need to access shared resources or a parallel need for similar resources (e.g., bibliographic information, geographic reference data, molecular structure, etc.)

## The Genome Example

The importance of integrating genome information resources has been publicly recognized in reports from groups of leading biologists (e.g., the Genome Science and Technology Center directors; GeSTeC Directors, 1994) and of informatics experts (an invitational meeting held in Baltimore in April, 1993; reported in Robbins, 1994c):

> A...major...goal of genome informatics should be the integration of genome and genome–related databases. (GeSTeC report)

> 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...as a federated information infrastructure of interlocking pieces. (Baltimore report)

For a variety of historical and operational reasons, genome data are now, and will continue to be, housed in several independent data resources. Already, the lack of interoperability among these resources makes answering simple questions overly difficult, leading the Baltimore report (Robbins, 1994c) to observe:

> An embarrassment to the Human Genome Project is our inability to answer simple questions such as, "How many genes on the long arm of chromosome 21 have been sequenced?"

Removing this embarrassment will require several interoperability improvements:

- *Technical interoperability* must be achieved, so that minimum functional connectivity can be assumed among participating information resources.
- *Semantic interoperability* must be developed, so that meaningful associations *can* be made between data objects in different databases.
- *Social interoperability* must occur, so that meaningful associations *are* made between data objects in different databases. Each asserted link is an act of scientific creativity, not merely the result of computations on existing data. Therefore, social changes must occur to stimulate the creation and entry of this information.

These three advances will likely occur in the order given. Without semantic interoperability, it is difficult to define, much less enter links between objects. Without technical interoperability, the motivation for providing semantic interoperability is lacking.

---

⇒ **Needed:** Coordinated, international efforts to improve the technical, semantic, and social interoperability among biological information resources.

---

Another embarrassment is the length of time that genomic databases have been promising, but not delivering, connectivity with other information resources. The problem has been, not lack of good intentions or of hard work, but rather a simple absence of the technical interoperability infrastructure necessary to enable and motivate the remaining work. However, recent advances such as World-Wide Web (WWW) and Mosaic (Berners-Lee, et al., 1994; Schatz and Hardin, 1994; Vetter, et al., 1994)) now promise that solutions may soon be at hand. This essay will describe some of those recent advances and will comment on the remaining steps to be taken. For reasons of space, the essay will not treat either semantic or social interoperability. Semantic compatibility and other aspects of genome informatics have been discussed elsewhere (Robbins, 1992, 1993, 1994a, 1994b, 1994c).

## Achieving Interoperability

Achieving full interoperability among distributed databases is a hard problem, and no simple, vendor–supplied solution is available. The overall difficulty is affected by many factors that can be arrayed along several dimensions, such as site autonomy, system heterogeneity, and physical distribution. Unfortunately, the integration of biological databases nearly always involves high levels of difficulty on all dimensions.

### *Interoperating Databases Still a Research Problem*

It has generally been held that achieving full read–and–write interoperability across multiple databases requires the development of an integrated data model, or schema, spanning the participating information resources. A recent refinement is the integration of only portions of the local schema, which may be specially modified to facilitate integration. These modified subschemas are known as export schemas (Figure 3). (A collection of research papers on database interoperability may be found in Hurson, Bright, and Pakzad, 1994.)

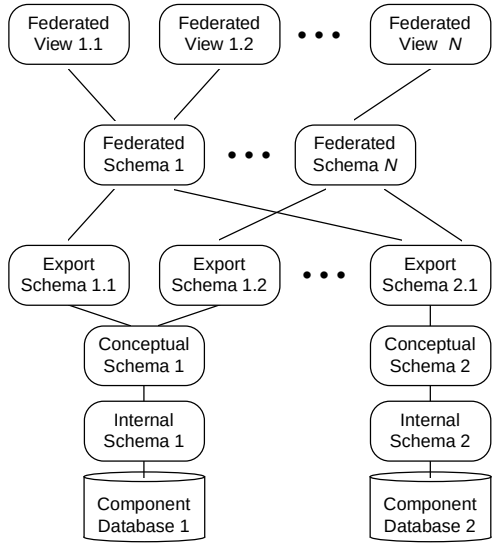


**Figure 3.** Many researchers believe that an essential first step toward database interoperability is the preparation of one or more export schemas by each participating database. The export schemas are then integrated into one or more federated schemas, which serve as the basis for one or more federated views into the underlying integrated information resource. (Figure adapted from Sheth and Larson, 1990).

Export schemas buffer against changes in the underlying databases, but only if the export schemas themselves are stable. Ultimate fragility due to inevitable changes in the underlying systems has led Chorafas and Steinmann (1993) to

dismiss global schema integration as impractical, requiring too much managerial coordination, and to characterize such attempts as an "approach which has been tried and failed since 1958". These authors claim that integration efforts can be arranged along a continuum bounded by unfeasibility (full schema integration) and unacceptability (do nothing). Such pessimism notwithstanding, recent experience with information publishing systems such as WWW have shown that useful federations can be built upon loosely coupled systems.

### *Loosely Coupled Data Publishing*

In database research, different kinds of interoperating databases have been described as falling into a taxonomy (Figure 4). Loosely coupled systems have not been significantly pursued in the database community, because joint updates across such systems are widely judged impossible. Tightly coupled systems are not practical across diverse biological information resources, because too high a level of integrated management is required. In consequence, biological information resources have for years promised, but not delivered, interoperating systems.

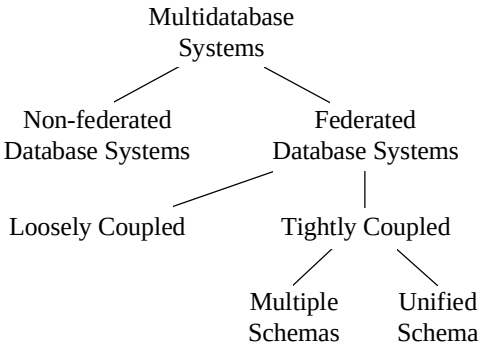


**Figure 4.** A taxonomy of multidatabase systems, according to Sheth and Larson (1990).

In the past 18 months, however, WWW and Mosaic have swept across the world of networked computing, demonstrating the tremendous power of loosely coupled, read–only information resources. Over 20,000 different WWW servers now exist, and any user with one copy of some generic browsing software, such as Mosaic, can access any one of them simply by knowing its name, or by following cross references from other servers.

Providers of data can easily link their information to that in other WWW servers, simply by embedding the "name" of the other data objects in the local information file. The power of WWW technology has rapidly led nearly every major provider of biological data to adopt WWW as *part* of their local interoperability strategy. With the advent these systems, the dichotomy between tightly and loosely coupled systems now appears more as a continuum (Figure 5).

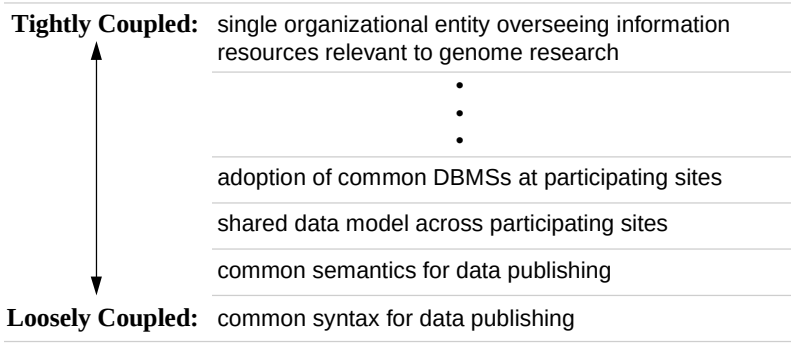


**Figure 5.** The distinction between tightly coupled and loosely coupled systems, seen as designating the ends of a continuum of relationships among database publishing systems. The tightest level of coupling yields a completely integrated, single management structure. The loosest level of coupling involves a collection of wholly independent organizations that share in common only a willingness to publish their data in a common syntax.

### *Lessons from the Genome Project*

A genome–informatics advisory group offered a challenging goal (reported in Robbins, 1994c) for a federated information infrastructure for biology:

> Adding a new database to the federation should be no more difficult than adding another computer to the Internet.

Achieving such interoperability will be a multi–part process, with some effort being devoted to developing necessary specific applications and other effort being devoted to the development of appropriate underlying enabling technology.

The suggestion of a networking metaphor proved remarkably prescient—the period since that meeting has seen the tremendous growth of WWW as an information-delivery system that is based in part on extensions to existing networking protocols.

## WWW and Mosaic are not Enough

Although WWW and Mosaic (and also gopher – Anklesaria et al., 1993) have been employed to good use in the distribution of structured data by several major biological databases, they are not capable of meeting all of the needs of the biological database community. These projects have intellectual ties with information retrieval (IR), not database development, and many differences exist between the needs of database users and the services delivered by IR systems in general, and gopher or WWW:Mosaic systems in particular:

- IR query systems support ambiguous queries and resolve them using probabilistic retrieval systems. Databases, on the other hand, hold

structured data and provide exact answers to well–formed, structured queries.

- Hypertext supports flexible linkages between objects, but more structured linkages, with defined semantics (such as a foreign key to primary key reference), are required for structured data.
- Gopher and WWW servers present their data objects one at a time. A menu choice retrieves one object in gopher, a click on a hypertext link retrieves one more HTML document. In database queries, users frequently want to obtain *sets* of objects that match their request.
- Hypertext links are available as paths the user may or may not choose to follow. Active steps must be taken to follow any particular step. Database queries frequently involve requested "joins" among data objects, in which the user wants to specify in advance what related objects are to be retrieved and in what connected configuration. Single database queries should be capable of returning large sets of joined objects, not merely the "option" of following what might be hundreds or thousands of hypertext links one mouse click at a time.
- Hypertext browsers are intended for human usability, with the assumption that they will present multiple navigation options to a human user. Database users frequently need a computational application programming interface with which to interact, so that they can direct an application program to extract and analyze data sets, then return the analytical results.

The list could be extended. But, the goal here is to offer neither the definitive characterization of the problem nor the definitive solution. Instead, we wish to establish that, *in their present form*, the widely available tools for easily fetching text and hypertext do not adequately meet the needs of those who desire integrated access into structured databases.

Many groups are working to extend WWW:Mosaic technologies to handle more structured data, in varying degrees of generality. A good solution would do for databases what WWW has done for hypertext: provide an easy way to deliver transparent navigation through the holdings of information resources. The WWW approach involved a new data model (HTML documents), new protocols (e.g., HTTP), and, most importantly, a new vision for how information should be represented, organized, and delivered. It is presently an open question whether the needs of structured database users can be met through clever additions to the WWW:Mosaic system, or whether substantial new database equivalents of HTML and HTTP will need to be developed.

---

⇒ **Needed:** Coordinated, international efforts to develop appropriate methods for supporting loosely coupled access to structured data.

---

Indeed, resources spent in successful pursuit of this end would likely produce a higher return on investment than any other such commitment to bioinformatics.

## FUTURE NEEDS

Many other yet unsolved technical and social issues in bioinformatics 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. 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 (Cinkosky, et al., 1991; Robbins, 1994a). The communication role of genome databases has been explicitly recognized by leading genome researchers in a recent review (Murray, *et al.*, 1994):

> 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 into 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.

Several important policy issues relevant to bioinformatics 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.

In addition, the present methods for reviewing bioinformatics projects tends to confuse the question of "is such a resource needed?" with "is this the facility to deliver the resource?" so that reviewers can be faced with the choice between eliminating a needed resource and supporting a poorly run operation. This leads to a vicious circle, with an unwillingness to cancel the project, coupled with an unwillingness to provide significant funding. Projects that fall into this condition have great difficulty extricating themselves.

---

⇒ **Needed:** Coordinated, international efforts to establish information-infrastructure priorities and to review bioinformatics projects.

---

At present, nearly all public bioinformation 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 dificulties.

---

⇒ **Needed:** Coordinated international efforts to facilitate cooperation among bioinformatics resources.

---

## BIOINFORMATICS AND THE GLOBAL INFORMATION INFRASTRUCTURE

A recent report from the National Research Council[9] notes that the National Information Infrastructure (NII) may be divided into analog and digital components, based on method of delivery, and into commercial and non-commercial sectors, based on usage. The report further subdivides commercial use into ETC (entertainment, telephone, and cable) and other categories, and subdivides non-commercial use into education, libraries, and research (Figure 6).

The report goes on to note that over the next ten years, the commercial sector will likely spend between $10 and $20 billion building a new communications infrastructure that will move most ETC commercial usage away from analog and to a digital substrate. The vastness of replacing an entire national communication

---

9 *Realizing the Information Future*, prepared by the NRENaissance Committee.

infrastructure, especially the "last mile" components[10] is such that it can only be repeated a few times each century.

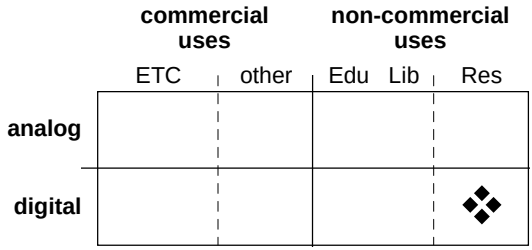


**Figure 6.** The current information infrastructure may be categorized by delivery mechanism (analog vs. digital) and by usage (commercial vs. non-commercial). Bioinformatics usage falls nearly completely in the non-commercial, digital category. Until recently, public digital networking was associated with the Internet, with commercial digital networking being carried out over private leased lines.

The implications for bioinformatics of this coming major transition are significant. If the new infrastructure results in a loss of function now available, biologists will have little choice but to accept the consequences since it would likely be very difficult to justify spending millions of public dollars to "fix" a multi-billion dollar private investment.

Instead, the fix should occur in advance, with biologists documenting their information–infrastructure needs and showing how an NII that meets these needs will also meet key, if as yet unrecognized, needs in the private sector. The pressure to move quickly in developing and sharing this documentation is underscored in the NRC report:

> The challenge for the country is to shape the architecture of the network so that the NII that results meets not just short-term commercial objectives, but also longer-term societal needs. It is important to appreciate this differences in outlook now, since progress dictates that rough agreement on an NII vision be achieved sooner rather than later. (NRENaissance Committee, 1994)

To this end, the report presents an extensive argument in favor of an Open Data Network model (Figure 7), in which the underlying communication infrastructure is designed to support the functionality of all "information appliances"—a term used to describe any electronic device that operates on information, whether that be obtaining and viewing a movie or manipulating complex data. The report concludes:

---

10 *Last mile* components refer to those parts of the communication infrastructure necessary to connect each individual user to the central system.

> The NII initiative presents exciting opportunities for the federal government to reap far greater returns from the NREN program than those experienced to date and to meet a broad range of social and economic needs. The NSF and other HPCC agencies have opportunities to lead in the development of general and flexible architectures and to experiment with their implementation. NIST and other agencies have opportunities to promote more effectively the kind of standards that will be needed to assure the broad interoperability characteristic of the Open Data Network described by this committee. Above and beyond the roles that seem obvious for individual agencies is a need for sustained leadership and effective coordination—for management in the best sense, reflecting the recognition that the federal role is one of catalyst rather than performer for most of the actions necessary to implement the NII.

Coming changes in information infrastructure will not be restricted to the United States. Indeed, we are already seeing the growth of a truly Global Information Infrastructure.

---

⇒ **Needed:** Coordinated, international efforts to ensure that the needs of the bioinformatics community are addressed during the coming communications revolution .

---

## SUMMARY

Bioinformatics, the use of computers to support biological information management, has become an enabling technology, essential for the success of big–science projects in biology. Not yet a true discipline of its own, bioinformatics occupies space between biology and computer science, with interests in library and information science, engineering, and management as well. The interdisciplinary nature of bioinformatics can make it difficult for projects to gain support from agencies focused either on biology or on computer science.

With the growth of global networking, achieving interoperability among biological information resources is now one of the most pressing challenges in bioinformatics. Technical, semantic, and social advances will be required for success to occur.

Although the great success of Mosaic and WWW in providing a loosely coupled, distributed information delivery system has finally proved the tremendous utility of a federated information infrastructure, WWW technology itself is not sufficient to meet the needs of those who need coordinated access into robust, structured data.

Tools for resource discovery and resource filtering loom as unmet needs. Better data standardization and data indexing will be required as the resources continue to grow exponentially. As databases become more like scientific literature, new infrastructure functionality must be added.

Intellectual property rights, data sharing, and data access remain as challenging policy issues, complicated by differing national approaches. Improved and coordinated approaches to providing long–term support for bioinformatics projects are needed.

As the Global Information Infrastructure takes shape, international agencies with an interest in bioinformatics should work together to ensure that advances in the commercial sector are accompanied with support for needed functionality in the research community.

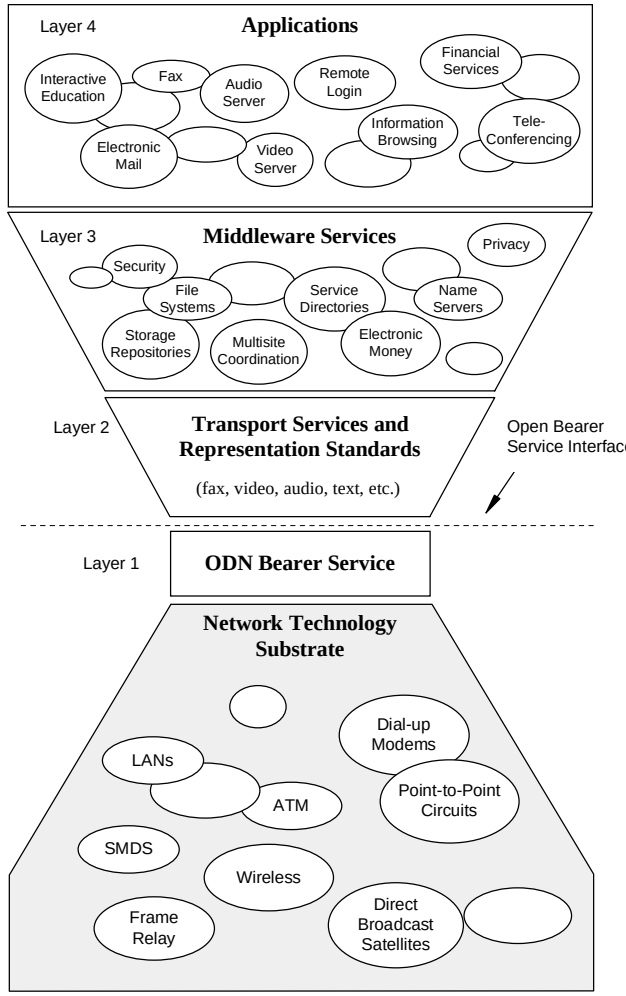


**Figure 7.** A four-layer model for an Open Data Network. (adapted from NRENaissance Committee, 1994, *Realizing the Information Future.* Washington, D.C.: National Academy Press.)

## RECOMMENDATIONS

Coordinated, international efforts are needed to:

- develop better means for supporting worthy bioinformatics activities.
- further the development of a federated information infrastructure for biology.
- improve the technical, semantic, and social interoperability among biological information resources.
- develop appropriate methods for supporting loosely coupled access to structured data.
- establish information-infrastructure priorities and to review bioinformatics projects.
- facilitate cooperation among bioinformatics resources.
- ensure that the needs of the bioinformatics community are addressed during the coming communications revolution.

# BIBLIOGRAPHY

Anklesaria, F., McCahill, M., Lindner, P., Johnson, D., and Torrey, D. 1993. F.Y.I. on the Internet Gopher Protocol

Berners–Lee, T., Cailliau, R., Luotonen, A., Nielsen, H. F., and Secret, A. 1994. The World–Wide Web. *Communications of the ACM*, 37(8):76–82.

Chorafas, D. N., and Steinmann, H. 1993. *Solutions for Networked Databases: How to Move from Heterogeneous Structures to Federated Concepts.* New York: Academic Press, Inc.

Cinkosky, M. J., Fickett, J. W., Gilna, P., and Burks, C. 1991. Electronic data publishing and GenBank. *Science*, 252:1273–1277.

Committee on the Formation of the National Biological Survey. 1993. *A Biological Survey for the Nation.* Washington, D.C.: National Academy Press.

GeSTeC Directors. 1994. Report: NCHGR GeSTeC Director's meeting on genome informatics. Available electronically from Johns Hopkins WWW server, http://www.gdb.org/Dan/nchgr/report.html.

Hurson, A. R., Bright, M. W., and Pakzad, S. H. (Eds.) 1994. *Multidatabase Systems: An Advanced Solution for Global Information Sharing*. Los Alamitos, California: IEEE Computer Society Press

Lewin, R. 1986. DNA databases are swamped. *Science*, 232:1599.

Murray, J. C., Buetow, K. H., Weber, J. L., Ludwigsen, S., Scherpbier–Heddema, T., Manion, F., Quillen, J., Sheffield, V. C., Sunden, S., Duyk, G. M., Weissenbach, J., Gyapay, G., Dib, C., Morrissette, J., Lathrop, G. M., Vignal, A., White, R., Matsunami, N., Gerken, S., Melis, R., Albertsen, H., Plaetke, R., Odelberg, S., Ward, D., Dausset, J.,Cohen, D., and Cann, H. 1994. A comprehensive human linkage map with centimorgan density. *Science*, 265:2049–2054.

NRENaissance Committee. 1994. *Realizing the Information Future: The Internet and Beyond.* Washington, DC: National Academy Press.

Robbins, R. J. 1992. Database and computational challenges in the human genome project. *IEEE Engineering in Medicine and Biology Magazine.*, 11:25–34.

Robbins, R. J. 1993. Genome informatics: Requirements and challenges. In: Lim, H. A., Fickett, J. W., Cantor, C. R., and Robbins, R. J. (eds). *Bioinformatics, Supercomputing and Complex Genome Analysis.* Singapore: World Scientific Publishing Company.

Robbins, R. J. 1994a. Biological databases: A new scientific literature. *Publishing Research Quarterly*, 10:1–27.

Robbins, R. J. 1994b. Representing genomic maps in a relational database. In: Suhai, S. (ed). *Computational Methods in Genome Research*. New York: Plenum Publishing Company. (in press)

Robbins, R. J. (Ed.) 1994c. Genome informatics I: Community databases. *Journal of Computational Biology*, in press.

Robbins, R. J. 1994d. Genome Informatics: Toward a Federated Information Infrastructure (keynote address). The Third International Conference on Bioinformatics and Genome Research; Tallahassee, Florida; 1–4 June 1994.

Robbins, R. J. 1995. *Data Publishing and the Global Information Infrastructure*, manuscript in preparation. Prepublication copies available for review from the author.

Schatz, B. R., and Hardin, J. B. 1994. NCSA Mosaic and the World Wide Web: Global hypermedia protocols for the internet. *Science*, 265:895–901.

Sheth, A. P., and Larson, J. A. 1990. Federated databases systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys*, 22:183–236.

United States Department of Energy. 1990. *Understanding Our Genetic Inheritance. The U. S. Human Genome Project: The First Five Years.*

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. 1988. *Mapping and Sequencing the Human Genome*. Washington, DC: National Academy Press.

Vetter, R. J., Spell, C., and Ward, C. 1994. Mosaic and the World-Wide Web. *IEEE Computer*, 27(10):49-57.