

we search for single base changes that may cause a genetic defect, part of the problem is distinguishing which change(s) is responsible for the disease. The second reason is that, as argued below, the data quality from large sequencing projects also requires a change in our current concept of sequence. In fact, the concept of "the genome" as a unique entity is not quite firm, which further complicates matters. Humans differ from one another in about one nucleotide in one thousand. In addition, recombination makes it difficult to maintain genomic material in a static condition. For these reasons, genomic sequence databases must necessarily be more fluid than our current database "world view." New models of sequence are required, and some people, including database staffs, have already begun to think about these problems.

While most discussions of genomic sequencing center on volume or number of nucleotides, the real situation is much more complex. For example, a clone will be shotgun sequenced and assembled into islands of sequence. Sequencing errors will necessarily exist in these sequences. Eventually, the center will declare the clone to be sequenced. If a physical map of ordered clones exists, the clone order will allow assembly of the clone sequences into larger islands of genomic sequence. If there is no physical clone map, then island assembly will be less efficient, especially in the early stages of the project. Obviously, it is unacceptable to keep publicly funded sequence from distribution until the entire genome is sequenced. Therefore, decision as to length (in nucleotides) and quality of sequence required for its public distribution will have to be made. It will also be necessary to correct earlier sequences as more data are obtained and the sequence is revised.

In genomic sequencing, there will be new demands on data analysis, exacerbating the problems discussed earlier. Detailed laboratory analysis of sequence function will often not be performed. Consequently, computational analyses will be the only available tools with which to approach many problems. Determination of gene coding regions by computer, for example, is already a central and troublesome problem, as is locating intron-exon boundaries. Classification of genes into families and superfamilies also relies on computer analysis. It is my own view that there should not be a privileged group getting first look at the data unless it is the people actually doing the sequencing. There are many other important issues, such as relating sequence to genetic and physical maps and to available experimental materials such as clones. These relationships must be updated as more data become available. The recent concept of sequence tagged sites (STS) is likely to be very useful in this regard. STS are short sequences that promise to provide a means for correlating physical and genetic maps and reducing the need for clone banks. In general, the importance of computer analysis will increase with genomic sequencing, requiring new methods and novel hardware to meet the needs of megasequence analysis.

There is, of course, a concern that today's sequence databases, which have received criticism for both lack of timeliness and incompleteness, evolve to meet the future needs. There are some good signs and I will briefly discuss the nucleotide sequence databases, in particular GenBank, as I am most familiar with its recent progress.

An effort to reduce the backlog of all sequences from 1960 to 1987 that are not included is well along, and this effort will be complete by the end of 1990. GenBank contains 95% of the sequences published in the last 2 years in journals for which it is responsible. Today, about 80% of the published sequences are entered and annotated within 3 months, and efforts are underway to improve this percentage. An effort is made to have journals require or encourage submission of sequences to GenBank in computer-readable form. While 65% of the GenBank entries come directly from the authors, about 45% of the submissions are in computer-readable form. The program Authorin has been designed to help scientists enter and annotate their sequences. Relational database management systems are being tried as a replacement for the older, flat file system. Others are exploring object-oriented databases.

None of this is easy. Collecting and managing data that are growing so rapidly, that require constant correction, and that must be adapted to new definitions are major tasks. Cooperation between databases has obvious scientific and political difficulties, even within one country. When we factor in problems of international cooperation, the reality of a unified set of biological databases seems even more remote. These areas require policy decisions that will affect the progress of international science. Who should make these decisions? Who will actually make them? National and international databases must be coordinated. The DNA sequence databases in Japan, Europe, and the United States may serve as a model for dealing with the many unresolved issues. We seem to be moving generally in the right direction, but it is critical to accelerate our efforts. We cannot leave the future of information management in biology to chance.

HISTORICAL SKETCH

The History of the Genetic Sequence Databases

TEMPLE F. SMITH

Molecular Biology Computer Research Resource, Dana-Farber Cancer Institute, Harvard University, 44 Binney Street, Boston, Massachusetts 02115

"The (entire human) genomic sequence will be the raw material for the Science of the twenty-first century" (Walter Gilbert, 1986, Waterville Valley, New Hampshire, cited in Gruskin and Smith, 1987)

Statements such as this arise from the recognition that the wealth of sequence data becoming available will convert biology from a science primarily of data collection and exploratory experimentation to one more driven by mathe-

mathematical analyses and the testing and refinement of theoretical hypotheses. This is not to suggest that mathematical analyses or deep theoretical concepts have not played an important role in formulating our modern view of biology. Rather, we are witnessing a natural metamorphosis in which the new and, until recently, unanticipated mountain of highly syntactically structured data is opening vast new analysis and theoretical frontiers. Theories, for example, concerning the structure of the regulatory networks controlling the complex overlapping suites of genes involved in development and how they evolved will be developed and tested. As in most of biology, this will require comparative analyses, in particular sequence comparative analyses, and here, the genetic sequence databases will play a crucial role. Therefore, a review of the recent scientific, economic, and sociological (political) history of these databases is appropriate.

The most obvious events leading to the creation of the sequence databases were the development of the methods for directly determining the amino acid sequence of proteins (Sanger *et al.*, 1956; Edman and Begg, 1967) and later the base sequence of the nucleic acids (Sanger and Coulson, 1975; Sanger *et al.*, 1977; Maxam and Gilbert, 1977, 1980). Of equal importance was the early recognition by a few researchers, such as Zuckerkandl and Pauling (1965), that within these data there potentially lay a record of life's evolutionary course. Most protein chemists were also aware early on of the relationship between knowledge of a protein's amino acid sequence, its X-ray-determined structure, and its function. Thus, the 1960s began, researchers found themselves collecting sequences related to their own investigations from the literature and from colleagues.

The intertwining of the new expansive technologies of molecular biology and computer science was an early event with fortuitous timing. The collections of restriction enzymes (Nathans and Smith, 1975; Roberts *et al.*, 1977) and the new techniques (Maxam and Gilbert, 1977; Sanger *et al.*, 1977) that brought cloning and sequencing into any laboratory coincided with the dawn of the departmental mini- and bench-top computers. One should recall that it was just within the previous two decades that the structures of DNA and myoglobin were first discerned, a time during which the foundations for the revolution in computer hardware were also being laid in solid-state electronics. The VAX 780s were introduced in 1977 and the first practical microcomputer arrived in 1976 with the S100 bus and the CPM operating system. Within a few years, IBM had been attracted to the new market, introducing its personal computer in 1982. Thus, tools needed to store, search, and analyze the new data grew up alongside the tools necessary to generate the data.

The first major collection of genetic sequence information was assembled by Margaret Dayhoff and R. S. Ledley in the mid 1960s. This collection, while initially assembled for their own research, was made available to the larger research community. This was done most notably in the form of the "Atlases" of protein sequence and structure. By 1969, with the fourth volume, there were over 300 protein sequences and 16 nucleic acid sequences. Dayhoff had grouped the data into

evolutionarily related homologous functional families. This demonstrated the utility of organizing sequence data under a major theoretical construct.

It is important to note that although Dayhoff's work was supported by the National Institutes of Health, it was not as database activity but as basic research. In fact, the NIH, with two relevant exceptions, did not support database activities *per se*. The first exception was the historically autonomous National Library of Medicine, whose database activities were largely bibliographic. The second exception was a small program within the Division of Research Resources (DRR) on scientific and technology information resources. There, in the late 1960s, in a manner foreign to NIH's general policy of waiting to react to the research community's requests, an effort was initiated to create a computer system that would eventually provide an interface to a number of sequence databases. This was the PROPHET project (Castleman *et al.*, 1974), a time-sharing molecular analysis system with graphics support. It was not a project requested from within the biomedical research community, but one encouraged by a few who anticipated the computer's potential, including William Raub of NIH. There was other NIH support for the Chemical Abstracts and Structure database, for example, but this activity had been turned over to the National Science Foundation (NSF).

In 1973, the X-ray crystallographic protein atomic coordinate data collected by Helen Berman, Olga Kennard, Walter Hamilton, and Edgar Meyer, Jr., were made available. This new structural database at Brookhaven National Laboratory was under the direction of Tom Koetzle and funded by NSF. It would be over 10 years before there was joint support for this important resource that included NSF and two NIH divisions, National Institute of General Medical Sciences and DRR. The existence of the crystallographic database set an important precedent. Here data of a highly proprietary nature (years to generate and sometimes years more to interpret) were being made available to the entire research community. This database was viewed by Kennard and Sussman as a prototype for the later EMBL database.

By 1974, a second major sequence database effort, led by Elvin Kabat, began at NIH. Here again, as with Dayhoff's work, this database developed out of basic research interests. In this case, the interest was in the structure and diversity of the immunoglobulins. This database of "Proteins of Immunological Interest" was one of the first databases to become available online under the PROPHET system. Other sequence databases organized in the 1970s have continued to the present, including the RNA database of Mathias Sprinzl in Germany and the protein sequence databases started by the Protein Research Foundation (PRF) of Japan. The latter is associated with the journal *Peptide Information*, first published by the PRF in 1975.

Thus by the end of the 1970s there were three major U.S.-supported protein databases (under Dayhoff, Kabat, and Koetzle) in use worldwide. In addition, there were two nationally supported time-sharing computer analysis systems available. These were the PROPHET system at Bolt, Beranek

and Newman Inc. in Boston and the Molgen project under the Stanford University Medical Experimental Computer Resource, SUMEX, itself a DRR-supported project. The latter was included in a local attempt to make sequence analysis software readily available to Stanford molecular biologists. This was part of a larger attempt involving researchers such as Josh Lederberg to introduce Artificial Intelligence into biomedical research. The project rapidly expanded via a network "guest" account that allowed access to many scientists outside the Stanford community. The popularity of this access eventually surpassed the available support and led indirectly to its shutdown. Parts of this SUMEX-associated analysis support later became the core for the commercial venture IntelliGenetics, founded in 1980.

These events set the stage for a 1979 workshop organized by Norton Zinder and Carl Anderson under the sponsorship of the National Science Foundation. It was at this small meeting, held March 1st through 3rd at Rockefeller University, that the need for a fully supported nucleic acid sequence database was formally outlined. In attendance were some 35 scientists, including 5 from outside the United States, plus observers from NIH and NSF. The "agenda" items for this and the subsequent EMBL-sponsored meetings were not limited to the need to establish database facilities. Participants also noted the need to develop analysis tools. Recall that by 1979 there had been many applications and considerable development of the sequence comparison and evolutionary reconstruction methods (Waterman *et al.*, 1976; Waterman, 1984). These tools had evolved considerably from the heuristic methods used in the late 1960s by Fitch and Margoliash (1967), Dayhoff and Eck (1966), and others.

The discussions at Rockefeller were wide ranging. There were discussions of basic research problems in sequence generation and comparative analyses, as well as discussions on whether to include only published data, what associated information to include, and the potential need for validation. Concerns were raised as to whether a single computer facility would come to dominate the use and structure of the data in some negative manner. There was even a peek into the future when Clyde Hutchinson demonstrated what could be done on one of the new inexpensive personal computers.

A consensus emerged from the workshop on the need to establish an international computer database for nucleic acid sequence information to be correlated with as much other biological information as possible. A single database could avoid duplication of both data collection and analysis efforts. The meeting report listed "at least" six groups interested in being directly involved in the creation of such a computer resource. Among these were a group at the NBRF (led by Dayhoff), the PROPHET and SUMEX groups, a group at Los Alamos National Laboratories (including Walter Goad, George Bell, Michael Waterman, and this author), an English group (led by Olga Kennard and Fred Sanger) at the Medical Research Council in Great Britain, and finally a group at the EMBL (initially involving Ken Murray and Hans Lehrach). It is unfortunate that the report from this meeting was not available, in particular to NIH, until late November 1980. It

was never published, preventing any broadly based discussion within the research community prior to NIH's "sources sought" announcement almost 2 years later.

In August 1979, Bell and Goad organized a small meeting at which the outline of a proposal to create a DNA sequence library and analysis center at Los Alamos was discussed. While supporting the idea, those attending expressed concern about such a project being within Department of Energy rather than at an academic institution. The apparent lack of NIH commitment was also a concern. However, the strong computing facilities and the sequence analysis expertise at Los Alamos were thought important components for such a center, and a consensus formed to proceed.

The NBRF assembled a pilot nucleic acid sequence database as a logical extension to the amino acid sequence database in 1980, publishing a hard copy in 1981. This was funded in part from private commercial funds.

Ken Murray organized a meeting under EMBL sponsorship for April 24, 1980, in Schonau, Germany, entitled "EMBL Workshop on Computing and DNA Sequences." Among the events that had taken place since the Rockefeller meeting was the publication by Sutcliffe (1979) of the entire sequence of pBR322, one of the major cloning vectors. Thus, the acquiring methods had become powerful tools in the hands of others in addition to their original developers. Sanger and colleagues had completed the sequence of λ 174 2 years before and were now working on the entire lambda phage sequence.

At the Schonau meeting five attendees who had also been at the Rockefeller workshop noted that neither NIH nor NSF had publicly initiated action to establish a sequence database. This created a sense of some urgency and no doubt helped focus much of the discussion on the possible role of EMBL as a new sequence data collection and analysis center. There were four presentations at this workshop of particular interest. First, the plausibility of EMBL sequencing the entire *Escherichia coli* genome—remember this was 1980—was discussed. Second, using crystallographic databases as a model, Olga Kennard presented a detailed DNA database proposal. Third, both Joel Sussman of the Weizmann Institute and Walter Goad described existing pilot DNA databases. In particular, Goad discussed the utility of a "structured" database, an early relational database in which each field or logical subdivision is put into its own indexable table. (It was to take another 7 years before funding agencies and the community would begin to force such structures on the sequence and related databases.) Finally, Douglas Brutlag of Stanford demonstrated access to some of the pilot databases through the SUMEX computer system.

It is important to realize the degree to which this small group foresaw both the future needs and the potential of databases. Today, the utility of the computer and the databases are taken for granted. This is due in part to some of the folk history of molecular biology which grew out of the discovery by Russell Doolittle, using his personal computer, of the similarity between an oncogene and a growth hormone factor after the association had been missed by workers at Harvard and Caltech (Doolittle *et al.*, 1983). Although Zuck-

erkandl and Pauling (1965), Dayhoff and Eck (1966), Fitch and Margoliash (1967), and others foresaw the importance of computer support of databases and sequence analysis, these were not to become commonplace in molecular biology until the mid 1980s!

With prodding from a number of researchers, including Rich Roberts, John Abelson, Fred Blattner, and others, NIH (through Ruth Kirschstein and Elke Jordan) organized a "Nucleic Acid Sequence Data Bank Workshop" for July 14, 1980. This workshop was chaired by John Abelson. While other presentations were made (including Dayhoff and Goad again presenting their pilot database efforts), this was in reality an advisory board to make recommendations on the need and required nature of a U.S.-supported nucleic acid sequence database. Detailed recommendations for both the short and the long term were drawn up. Short-term recommendations included the establishment of groups to coordinate collection of both sequence data and analysis tools. For the longer term, the workshop urged NIH to establish a full Nucleic Acid Sequence Data Bank by January 1981. This data bank was to be available over a dynamic computer network and via magnetic tape distribution. It was to support subgroups working on special organisms and data subsets, such as a promoter library or the globin families. Some minimal search and analysis support should be available directly and other computer analysis programs should be collected and distributed upon request. Finally, there should be a National Advisory Committee.

Elke Jordan communicated these initial NIH recommendations to EMBL through Ken Murray at his request shortly after they were prepared. This laid the groundwork for an international collaborative database effort, which in itself would take nearly another 5½ years.

On August 11 1980, William Raub stated that NIH definitely planned to establish a Nucleic Acids database, and with EMBL continuing European planning, a number of important events followed in rapid succession. The pilot databases of Dayhoff, Goad, and Sussman all became available under the Molgen project at Stanford. Dayhoff submitted a supplemental proposal to NIH on August 13th to expand their nucleic acid sequence database. On August 28, Brutlag, Larry Kedes, and Peter Friedland submitted a proposal to convert the Molgen project into a true national analysis resource providing database access. On September 3, Los Alamos National Laboratory submitted its DNA library and analysis center proposal, and on September 8, Michael Waterman and this author submitted a grant supplement for expanding their sequence analysis development in conjunction with the Los Alamos proposal. All of these were unsolicited proposals to NIH to support work for which no program then existed.

On October 26th, Jordan convened a small subcommittee of Abelson, Roberts, Blattner, Kabat, and Greg Hamm (now involved with the EMBL efforts) to draft NIGMS project guidelines. And at a final meeting on December 7th, this ad hoc advisory committee completed definition of the project tasks with implementation in two phases. Phase I was to

establish a centralized database in collaboration with the Europeans and potentially the Japanese. The database was to be accessible electronically and distributed via magnetic media (as the protein sequence and structure databases by then were). Phase II was to establish an analysis and software library coupled to the database. For reasons that are not clear, neither the NIH Division of Research Resources nor the NSF was directly involved in these NIGMS drafts or even requested to be. On April 2, 1981, NIGMS released the "Sources Sought" for these projects. This was not a request for proposals but rather a solicitation of those able and potentially interested in carrying out specified tasks. In April 1981, just after NIH's release of the "Sources Sought," EMBL under the leadership of Murray and Hamm sponsored another workshop on organizing a European database. This workshop resulted in final EMBL plans for a database under the direction of Hamm. By the time of the European Molecular Biology Organization "Pattern Analysis in Nucleic Acid and Protein Sequences Workshop" in Saint-Agnan on October 27th, EMBL had a reasonable database in place, assembled from the existing pilots, although there were no European plans for a phase II. Many of the large number of U.S. scientists attending the Saint-Agnan workshop did not yet know of the NIH plans, or knew but were becoming impatient with the lack of official requests for proposals, with considerable discussion to this effect.

By the time proposals were requested near the end of 1981, NIH had decided to put off phase II, and under a co-funding arrangement with NSF, DOE, DOD, and a number of other different institutes within NIH, only phase I was to be supported. Three proposals were submitted: two from Los Alamos—one with Bolt, Beranek and Newman Inc., and one with IntelliGenetics—and one from Dayhoff at NBRF. There are a number of curiosities here. First, the reason for the Los Alamos group's double submission was that as a national laboratory it could not offer to collaborate with one commercial entity without being equally willing to do so with others. Second, the group at Los Alamos, associated with the Theoretical Biology and Mathematics divisions, had a very strong interest in phase II and was hoping to use either the PROPHET or the SUMEX experience in networking and distribution to minimize the phase I effort and to continue planning for phase II. The NBRF proposal would have created a nucleic acid sequence database in conjunction with the existing protein sequence database and analysis support but funded under a very different structure. With phase II on hold and the SUMEX announcement that the Molgen guest accounts would end on July 14, 1982, the future of a national analysis support center was at best unclear.

On June 30, 1982, the NIGMS announced the award for the nucleic acid sequence data bank, named GenBank, to Bolt, Beranek and Newman with a subcontract to Los Alamos National Laboratory. In general, the research community seemed pleased that a database had finally been established. There were some who were disturbed by certain particulars, no doubt for many complex reasons. Some obviously felt that given NBRF's long and successful history with the protein

database it would surely have been a better choice. This was particularly felt by Margaret Dayhoff. Others were still concerned that the database was not at an academic research center. The community showed some surprise and concern that only three proposals had been submitted. This was in part because three of the four players—Bolt, Beranek and Newman with PROPHET, IntelliGenetics as an outgrowth of the Stanford SUMEX/Molgen project, and the NBRF with the Protein Information Resource—were organizations with past links to the NIH infrastructure. No university or non-NIH-associated commercial centers applied. The question still remains whether this was only because no one else was in a position to attempt such a project or that somehow NIH and those concerned research scientists had not involved nor communicated with a wide enough community.

At a small workshop organized by Goad at the end of the summer 1982 in Aspen, Colorado, there was considerable discussion about the need for one or more phase II projects, e.g., high-speed search tools, more sophisticated pattern analysis, and increased interdisciplinary training. Two important events happened. First, David Lipman introduced the idea of developing a "hash code" method for searching the database (Wilbur and Lipman, 1983); second, William Baker of NIH/DRR suggested that monies might become available through DRR for phase II-like projects, particularly those emphasizing large-scale network access. The relationship, or more properly the lack thereof, to the NIGMS phase II task was unclear.

IntelliGenetics submitted an unsolicited proposal for BioNet in reaction to discussions with Baker. The proposal was reviewed by DRR in 1983. Some concern was raised in review about its relationship to the NIGMS-announced, but delayed, phase II and whether such a major undertaking should not be solicited from the larger community. There were also concerns that if NIH sanctioned one commercial software package, it might limit the opportunities for alternative developments. However, its funding was approved by the Division's Council, in part for "programmatic" reasons.

The BioNet proposal encompassed more than the establishment of a database access, search, and analysis center. It envisioned a major data and idea exchange network among the world's molecular biologists. Whether this was a realistic view given the highly independent and competitive nature of molecular research laboratories is, of course, open to question. The evolving computer hardware, with the proliferation of personal and microcomputers, would reduce the need for such centralized analysis. Yet while the creation of BioNet effectively prevented the full implementation of phase II as part of the new DNA sequence database effort—different NIH institutes cannot generate overlapping programs—it had many positive effects. In its 5 years of existence, BioNet provided many computer-naïve molecular biologists with their first access to the databases and a taste of the computer's utility.

The initial funding of GenBank did not prove adequate to maintain collection of the rapidly expanding data, particularly by literature extraction. There were limited funds for hard-

ware, and the computing costs at Los Alamos National Laboratory became excessive. These funding problems, along with limited computer science- and database-experienced staff, led to both the maintenance of the database in a flat file format (dropping the relational table form) on a very limited mini-computer system and the eventual introduction of incomplete or unannotated data entries. Network access through the PROPHET system, and later through BioNet, proved to be of secondary importance, as most large research laboratories and academic departments accessed the database through local installation. With more and more commercial and academic search and analysis packages becoming available on the new powerful computer workstations, this trend can only be expected to continue.

The DNA database is now being reorganized under a modern relational database management system, under pressure from many sources and with the increased funding under a new contract (now to IntelliGenetics Inc., again with a sub-contract to Los Alamos National Laboratory). The European EMBL DNA database has also recently been brought under a full relational database structure. This will make distributed collection, updating, annotation, and distribution much simpler and should improve the database's internal consistency. It is unfortunate that this has taken so long and that there still are sequence and related databases not yet using such standard computer science expertise.

One of the major problems with which the new GenBank was forced to deal was the time delay between the generation of new sequence information and its availability in the database. By working out a division of labor with the EMBL and newer Japanese database efforts, and by involving the authors and journal editors, GenBank and the EMBL databases are currently keeping pace with the literature. Today, manuscript submission to most journals requires the direct submission of relevant DNA sequence information to either GenBank or EMBL. However, one can hardly overemphasize the time and political effort this arrangement required. In addition, the databases are accepting "unpublished data." More such data can be expected as larger scale sequencing gets underway in the coming decade.

New database efforts are continuing. Following a CODATA task group recommendation, a strong international collaboration has been established between the NBRF/PIR protein database in the United States and two newer databases, the MIPS in the Federal Republic of Germany and the JIPID in Japan. The JIPID, founded in 1987, has expanded to include considerable sequence-associated biochemical information. In December 1986, the National Institute of Auto Immune Deficiency of NIH funded a new database under the leadership of Gerald Myers for human retroviral sequences as part of the nation's attack on AIDS. Here again, the growth was underestimated and major staffing and funding increases have recently taken place. The NSF and NIH have been attempting to integrate a set of old and new independent databases around the current efforts to sequence the entire *E. coli* genome. As in the earlier plans, the integration of considerable nonsequence information and analysis methods is

included. One of the major database integration efforts recently initiated is that by the Howard Hughes Medical Institute at Johns Hopkins University under Peter Pearson. While these are primarily genetic marker databases—originally organized by Victor McKusick and Frank Ruddle—they are being cross-referenced to sequence databases.

Sequence databases require a particular kind of continuous updating and cross-referencing. Note that it is the potential correlations between sequences encoded functions or their spatial and temporal expression that make sequence comparison such a powerful tool. However, much of what we learn about a sequence's function is discovered well after it has been entered into the database. Thus, updating and interdatabase cross-referencing are essential if we are to fully exploit the new sequence data. There are currently over 50 sequence related databases in existence (Lawton *et al.*, 1989), compounding cross-referencing efforts. The current surge of genome projects requires planning for expanding databases, database integration, and analysis facilities. Both the planned genome centers and the new National Center for Biotechnology Information at the National Library of Medicine have taken this as one of their prime goals. There are even new sequence data compaction methods (Smith and Smith, 1990) suggesting new data search and organization strategies.

Given these developments, perhaps we should ask what lessons if any we have learned. First, NIH has been slow to lead. Its commitment to the "human genome," for example, came only after strong pressure from a few far-sighted biologists and the competing efforts of DOE. While this may be the proper stance for this agency, it does require that the research community make conscious efforts to provide long-range planning council. Second, nearly all past databases have grown out of private collections. The conversion to international resources is often painful and always difficult. This is due to many factors, including lack of full community participation in planning and an initial reluctance to invest the needed funds or to face the political problems associated with potentially having others than the originators carry out the longer term efforts. In addition, there has been slow progress in exploiting the wealth of computer science and database management expertise available outside the biological community.

There are and will be problems in consolidating existing databases and terminating those no longer needed. How the community and/or the funding agencies deal with a constituency, albeit a dwindling one, of a canceled database or other program must be thought through. The termination of BioNet, which to some extent resulted from changing technology, may be a case in point.

An important need that has not been addressed is the training of young scientists in the interdisciplinary domains overlapped by computer science and molecular biology. The coupling of training programs with analysis development and multidatabase integration has been recognized as far back as the 1979 Rockefeller meeting and discussed at recent CO-DATA, HUGO, and other meetings (Morowitz and Smith, 1987; Baltimore, 1988; Alberts, 1988). These needs in part

lay behind the original phase II, the BioNet, the Dana-Farber Cancer Institute's MBCRR (Smith *et al.*, 1986), and other recent projects, such as the new National Center for Biotechnology Information at the National Library of Medicine (see Benson *et al.*, 1990), yet there is still no overall NIH or NSF long-term biology "informatics" strategy. This is particularly true in the area of postdoctoral interdisciplinary training so desperately needed if we are to train those who will be capable of fully exploiting the new genetic sequence data.

ACKNOWLEDGMENTS

This author is deeply indebted to the many fellow scientists who were involved in this history and who took time to recall it. In particular, I want to thank Fred Blattner—who apparently never throws anything away—for copies of all past agenda and workshop reports, and William Raub, Douglas Brutlag, Peter Friedland, Elke Jordan, Elvin Kabat, Larry Kedes, Tom Koetzle, Ruth Kirschstein, Kiyoshi Kurahashi, Robert Ledley, Maryellen Ruvolo, Wayne Rindone, and Rich Roberts for their notes and comments.

REFERENCES

1. ALBERTS, B. M. (Chairman) (1988). "Report of the Committee on Mapping and Sequencing the Human Genome," NRC, National Academic Press, Washington, DC.
2. BALTIMORE, D. (Chairman) (1988). "Recommendations and Priorities Developed by the ad hoc Program Advisory Committee on Complex Genomes," Whitehead Institute, Cambridge, MA.
3. BENSON, D., BOGUSKI, M., LIPMAN, D. J., AND OSTELL, J. (1990). The National Center for Biotechnology Information. *Genomics* **6**: 389–391.
4. CASTLEMAN, P. A., *et al.* (1974). The implementation of the PROPHET system. In "AFIPS, Conference Proceedings," Vol. 43, AFIPS Press, Montvale, NJ.
5. DAYHOFF, M. O., AND ECK, R. V. (1966). "Atlas of Protein Sequence and Structure," Vol. 2, NBRF Press, Silver Spring, MD.
6. DOOLITTLE, R. F., HUNKAPILLER, M. W., HOOD, L. E., DAVARE, S. C., ROBBINS, K. C., AARONSON, S. A., AND ANTONIADES, H. N. (1983). Simian sarcoma virus *onc* gene, *v-sis*, is derived from the gene (or genes) encoding a platelet-derived growth factor. *Science* **221**: 275–277.
7. EDMAN, P., AND BEGG, G. (1967). A protein sequenator. *Eur. J. Biochem.* **1**: 80–91.
8. FITCH, W. M., AND MARGOLIAS, E. (1967). Construction of phylogenetic trees. *Science* **155**: 279–284.
9. GRUSKIN, K. D., AND SMITH, T. F. (1987). Molecular genetics and computer analyses. *Cabios* **3**: 167–170.
10. LAWTON, J. R., MARTINEZ, F. A., AND BUVLES. (1989). Overview of the LiMB database. *Nucleic Acids Res.* **17**: 5885–5899.

11. MAXAM, A. M., AND GILBERT, W. (1977). A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA* **74**: 560-564.
12. MAXAM, A. M., AND GILBERT, W. (1980). Sequencing end-labeled DNA with base-specific chemical cleavages. In "Methods in Enzymology" (L. Grossman and K. Moldave, Eds.), Vol. 65, pp. 499-560, Academic Press, New York.
13. MOROWITZ, H. J., AND SMITH, T. F. (1987). "Report of the Matrix of Biological Knowledge Workshop," Santa Fe Institute, Santa Fe, NM.
14. NATHANS, D., AND SMITH, H. O. (1975). Restriction endonucleases in the analysis and restructuring of DNA molecules. *Annu. Rev. Biochem.* **44**: 273-293.
15. ROBERTS, R. J., *et al.* (1977). Restriction and modification enzymes and their recognition sequences. In "DNA Insertion Elements, Plasmids and Episomes" (A. I. Bukari *et al.*, Eds.), pp. 757-758, Cold Spring Harbor Laboratories, New York.
16. SANGER, F. (1956). The structure of insulin. In "Currents in Biochemical Research" (D. E. Green, Ed.), Interscience, New York.
17. SANGER, F., AND COULSON, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**: 441.
18. SANGER, F., NICKLEN, S., AND COULSON, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. USA* **74**: 5463-5467.
19. SMITH, T. F., GRUSKIN, K., TOLMAN, S., AND FAULKNER, D. (1986). The molecular biology computer research resource. *Nucleic Acids Res.* **14**: 25-29.
20. SMITH, R. F., AND SMITH, T. F. (1990). Automatic generation of primary sequence patterns from sets of related protein sequences. *Proc. Natl. Acad. Sci. USA*, in press.
21. SUTCLIFFE, J. G. (1979). Complete nucleotide sequence of the *Escherichia coli* plasmid pBR322. *Cold Spring Harbor Symp. Quant. Biol.* **43**: 77.
22. WATERMAN, M. S., SMITH, T. F., AND BEYER, W. A. (1976). Some biological sequence metrics. *Adv. Math.* **20**: 367-387.
23. WATERMAN, M. S. (1984). General methods of sequence comparison. *Bull. Math. Biol.* **46**: 473-500.
24. WILBUR, W. J., AND LIPMAN, D. (1983). Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA* **80**: 726-730.
25. ZUCKERKANDL, E., AND PAULING, L.C. (1965). Molecules as documents of evolutionary history. *J. Theoret. Biol.* **8**: 357-358.