structed networks, structures, hierarchies, and sets of movements that keep it mobile.[42]

The production of biological data entails a particular kind of working and a particular kind of knowing. Central to this point is the idea that bioinformatics requires standardization—a kind of flattening of the biological object—in order to function. Moving objects around in virtual space means making them computable, networkable, and so on. Nucleotide sequences have become such standardizable, computable, networkable objects. Making bioinformatics, then, has had much to do with constructing sequences in this way—as just such standard objects. The reducibility of sequence to data, to objects that can flow through the computer, has played a major role in establishing its importance and ubiquity in contemporary biological work. Sequence permits precisely the kind of abstraction or stripping down that is required for samples to be transformed into data. Bioinformatics has emerged "out of sequence" because it is sequence that has made it possible to move biology around virtual space.

5 Ordering Objects

Of all the structures that computers impose, databases are the most important. If we wish to understand classification and its consequences in the late twentieth and early twenty-first century, we need to understand databases and their role in the making of scientific knowledge. In biology, the influence of databases on practice and knowledge is profound: they play a role in data analysis, transmission, and communication, as well as in the verification and authentication of knowledge. In their day-to-day work, many biologists use databases for checking the results of their experiments, for storing and managing their data, or for performing simulations and experiments. How are such databases built, and by whom? How do they work? What kinds of structures do they contain? How do these structures influence the knowledge that can be made with them? This chapter explores the role of databases in scientific knowledge making using one prominent example: GenBank.

Biological databases, organized with computers, cannot be thought of as just *collections*.[1] Instead, biological databases are *orderings* of biological materials. They provide ways of dividing up the biological world; they are tools that biologists use and interact with. Computer databases store information within carefully crafted digital structures. Such tabulations can have profound social,

cultural, political, and economic consequences.[2] But as well as ordering society, databases construct orderings of scientific knowledge: they are powerful classification schemes that make some information accessible and some relationships obvious, while making other orderings and relationships less natural and familiar.[3] Organizing and linking sequence elements in databases can be understood as a way of representing the connections between those elements in real organisms. Like a billiard-ball model of a gas in physics, databases do not aim to be a straightforward representation of a biological system; rather, they aim to capture only some of its important features. The database becomes a digital idealization of a living system, emphasizing particular relationships between particular objects.

As I worked with biological databases in my fieldwork, I started to ask why the information in them was arranged the way it was. Indeed, how did databases become the preeminent way of storing biological data? Answering these questions required an interrogation of the history of databases. By examining a database diachronically, we can discover how changes in structure correspond to changes in the kind of work being performed (and in the knowledge being produced) through databases.

The different database structures that GenBank has used represent different ways of understanding and ordering biological knowledge. Early "flat-file" databases, such as those constructed by Margaret Dayhoff and the first iterations of GenBank, instantiated a protein-centered view of life in which single sequence elements were placed at the center of biological understanding. The "relational" databases that gradually replaced the flat files in the 1980s and 1990s emphasized the interconnections between sequence elements—biological function was produced by interaction between different elements, and the connections were reflected in the database. Finally, the "federated" databases of the postgenomic era, while still placing sequences at the center, allowed much wider integration of other (extra-sequence) data types. This gave structural expression to the notion that biological function could be best understood by modeling the relationships between genes, proteins, transcription factors, RNA, small molecules, and so on. By following data into databases, we see how the rigid structures of information technologies impose constraints on how data can move and be shaped into knowledge.

The activities of data storing and knowledge making are not separate and are not separable. Biological databases are not like archives and museums—they are oriented toward the future more than the past.

Treating them as part of "natural history" can cause us to overlook their importance in structuring the way biologists make knowledge about life. The computer is more than a mere organizing tool or memory device—it provides ways of representing, modeling, and testing biological systems. The sort of biological databases that arose in association with computers in the 1960s marked a new kind of object and a new, "theoretical" way of doing biology.[4]

A Brief History of Databases

Databases have a history independent of their use in biology. Like the computer, they were built for specific purposes for the military and for business data management. How do they work? What were they designed to do? The first databases—or data banks, as they were often called—were built, like so many other tools of the information age, for military purposes. Thomas Haigh argues that the Semi-Automatic Ground Environment (SAGE) was the first "data base." SAGE needed to keep track, in real time, of the status of bombers, fighters, and bases in order to serve as an automated early warning and coordinated response system in the event of aerial attack on the United States.[5] In the early 1960s, SAGE's creators at Systems Development Corporation were actively promoting "computer-centered data base systems" to business. The corporate world soon took up the idea of a management information system (MIS), which many hoped would provide an executive with instant access to all the pertinent information about his organization. Early MISs were essentially file management systems—pieces of software that contained generalized subroutines to open, close, and retrieve data from files. This technology was limited, however, by the fact that data records were stored one after another along a tape, and that to find a particular record, it was often necessary to scroll through large portions of tape. The introduction of disks in the early 1960s meant that data could be accessed at "random," and new possibilities arose for organizing data access and storage.

Beginning in 1963, Charles W. Bachman of IBM developed the Integrated Data Store (IDS), which became one of the most effective and influential file management systems. The IDS, designed for use with disks rather than tapes, allowed linkages between records in what came to be known as the "network data model." To find particular records in such a system, the user had to navigate from record to record using the various links.[6] A version of the IDS was used to run computers for the Apollo program in the late 1960s.

In 1970, Edgar F. Codd, working for IBM in San Jose, California, wrote a paper describing a new system for organizing data records. "A Relational Model of Data for Large Shared Data Banks" set out a scheme by which the representation of data could be completely divorced from their physical organization on a disk or tape: "Activities of users at terminals and most applications programs should remain unaffected when the internal representation of the data is changed. . . . Changes in data representation will often be needed as a result of changes in query, update, and report traffic and natural growth in the types of stored information."[7] Codd's idea was to organize the data into a set of tables that were related to one another by "keys" that linked data across tables.[8] For instance, a library database might contain information about all the books in its collections. Such information could be spread over multiple tables, as in this example (a library system with just four books):

BOOK_ID	AUTHOR_ID	BOOKNAME	CALL_NO	LIBRARY_NO
1	1	Making Sense of Life	XYZ	1
2	2	Simians, Cyborgs, Women	ZYX	2
3	2	Primate Visions	YXY	1
4	3	Nature and Empire	YYZ	2

AUTHOR_ID	FIRST_NAME	LAST_NAME
1	Evelyn	Keller
2	Donna	Haraway
3	Londa	Schiebinger

LIBRARY_NO	LIBRARY_NAME	LIBRARY_ADDRESS
1	Library of Congress	Washington
2	New York Public Library	New York

Columns with identical names in different tables ("AUTHOR_ID" and "LIBRARY_NO") are linked together in the database. To find the author of *Primate Visions*, for example, the database must first look up that title in the first table, retrieve the AUTHOR_ID (the "key" for the table of authors), and then look up the corresponding author in the second table. The query itself creates a link or "join" between two tables in the database. Codd's paper also suggested a "universal data sublanguage" that could be used to query and update the database. Such a language would refer only to the names of the tables and the names

of the columns within them, meaning that a command or query would still work even if data were reorganized; as can be seen in the library example, the rows and the columns could be rearranged without affecting the outcome of a query.

The relational model had two advantages over its "network" rivals. First, it did not require relationships between data to be specified during the design of the database; second, the abstraction of the structure from the physical storage of the data greatly simplified the language that could be used to manipulate the database. "Because the relational model shifted the responsibility of specifying relationships between tables from the person designing them to the person querying them," Haigh argues, "it permitted tables to be joined in different ways for different purposes."[9] Relational databases present an open-ended, flexible, and adaptable means to store large amounts of data. Despite IBM's initial support for the "network" model, the development of Codd's ideas through the 1970s led to the development of SQL (Structured Query Language) and the commercialization of the relational model through firms such as Oracle and Sybase.

Even from this brief history, it is clear that different types of database structures are appropriate for different types of data and for different types of uses. Moreover, this history suggests that databases act as more or less rigid structures for containing information—that the proximity and accessibility of particular kinds of data are determined by the form of the database itself.

Dayhoff and a New Kind of Biology

The first biological databases—that is, the first groupings of biological information ordered on a computer—were produced by Margaret Oakley Dayhoff. Dayhoff, born in 1925, was trained in quantum chemistry under George E. Kimball at Columbia University, receiving her PhD in 1948. Her thesis work involved calculating the molecular resonance energies of several polycyclic organic molecules—a computationally intensive problem that involved finding the principal eigenvalues of large matrices.[10] In approaching this problem, Dayhoff devised a way to use punched-card business machines for the calculations. After her graduate studies, Dayhoff pursued her research at the Rockefeller Institute (1948–1951) and at the University of Maryland (1951–1959). In 1960, she joined Robert Ledley at the National Biomedical Research Foundation (NBRF, based at Georgetown University Medical Center, where she also became a professor of physiology and biophysics), and it was here

that she turned her attention to problems of proteins and evolution. Ledley himself was a pioneer in bringing computers to bear on biomedical problems, trained as a dentist, but also as a physicist and a mathematician, during the early 1950s, Ledley worked with the Standards Eastern Automatic Computer at the National Bureau of Standards (NBS) in Maryland. The knowledge of digital computing architecture that Ledley attained at the NBS led him first to problems in operations research (OR) and then to the application of computers to biomedicine. In particular, Ledley was interested in using computers to create a mathematized biology that would allow, for example, computerized medical diagnosis.[11]

Ledley had a very specific vision of how computers would be useful to biology. In his OR work, Ledley had emphasized the translation of messy situational data into logical problems that computers could understand and solve. In his work with George Gamow on the genetic code, Ledley devised a way for biologists to translate their protein-coding schemes into matrices and symbolic logic that could be easily dealt with on a computer.[12] Likewise, in biology and medicine, computers would be tools that could be used for statistics, accounting, and data management.[13] In his lengthy survey of the field (published in 1964, although much of it was written some years earlier), Ledley outlined his justification for the computer management of biomedical information:

> The feasibility of such a system from a computer-technology point of view is unquestioned; there are already computers that carry out such closely related processes as making nation-wide airline and hotel reservations, recording, updating, and tallying bank accounts and other financial records, controlling large-scale defense installations, and so forth.[14]

At the NBRF, Dayhoff and Ledley began to apply computers to the solution of problems involving large quantities of experimental data. In 1962, the pair developed a computer program to aid in the experimental determination of protein sequences. Previously, the only way to experimentally determine a protein's complete sequence was to find the sequences of short fragments of the chain and then "try to reconstruct the entire protein chain by a logical and combinatorial examination of overlapping fragments."[15] For larger protein chains, it quickly became an almost impossible task to assemble fragments by hand. Dayhoff and Ledley's program not only rapidly checked possible arrangements, but

also suggested the best approach for further experiments where results were inconclusive. This program was followed by another, more sophisticated program that allowed for experimental errors and assessed the reliability of the finished sequence.[16] Dayhoff's choice of journal for publishing this work—the *Journal of Theoretical Biology*—suggests that she saw it as making a contribution to the organization and systematization of biological knowledge.

At about this time, in the early 1960s, Dayhoff began to collect complete protein sequences. Her reasons were twofold. First, protein sequences were important in their own right, since they contained the key information about how biology worked. Second, and perhaps more importantly for Dayhoff, proteins contained information about evolutionary history. At the same time that Dayhoff was beginning her collection efforts, Linus Pauling and Emile Zuckerkandl were developing a new method of studying evolution, and the relationships between organisms, using protein sequences as "documents of evolutionary history."[17] Dayhoff and others saw that such work would require both collections of proteins and computer programs to perform the computationally intensive tasks of sequence comparison and phylogenetic tree construction. Dayhoff and her colleagues at the NBRF scoured the published literature for experimentally determined protein sequences and entered those sequences on punched cards for computer processing. Although the collection itself was a nontrivial task, it was never Dayhoff's ultimate aim to be a botanist of sequences: "There is a tremendous amount of information regarding evolutionary history and biochemical function implicit in each sequence," she wrote to a colleague, "and the number of known sequences is growing explosively. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it."[18] Collection was a means to an end.

The first edition of the *Atlas of Protein Sequence and Structure*, published in 1965, listed some seventy sequences. Subsequent editions contained not only the protein sequences themselves but also extensive analyses performed by computer. These analyses included studies of the evolution of specific protein families, the development of a model of evolutionary change in proteins, an analysis of patterns in amino acid alleles, simulation of protein evolution, and studies of abnormal human hemoglobins, ribosomal RNA, enzyme activity sites, and transfer RNA.[19] The *Atlas* also provided phylogenetic trees and protein secondary (three-dimensional) structures. In the preface to the third edition of the *Atlas*, Dayhoff and her collaborator Richard Eck outlined their approach to the sequence collection problem:

> The mechanical aspects of the data presentation have been automated. The title, references, comments, and protein sequences in one-letter notation are kept on punched cards. The alignments, the three-letter notation sequences, the amino-acid compositions, the page layouts and numbering, and the author and subject index entries from the data section are produced automatically by computer.[20]

Although sequences had to be collected and entered from the published literature by hand, the aim was a computer-ready set of sequence information that could be rapidly subjected to analysis.

Two of Dayhoff's analytical concepts are particularly significant. First, Dayhoff realized that a careful study of the evolution of proteins would require a model of how proteins could mutate, and in particular, which amino acids could be swapped with one another in a sequence (called a "point mutation"). A naïve approach would treat all such swaps as equally likely—asparagine could just as easily be swapped for lysine as for valine, despite the chemical differences between the two amino acids. If biologists wanted a better account of the evolutionary distance between sequences, however, a more sophisticated approach was required. To provide this approach, Dayhoff invented the notion of a PAM (point accepted mutation) matrix. The idea was to use the protein sequence data she had collected to create a table (or matrix) showing the number of times each amino acid was observed to mutate to each other amino acid. Dayhoff then computed a "relative mutability" for each amino acid by dividing the total number of observed changes in an amino acid by the number of total occurrences of that amino acid in all the proteins examined. By using the relative mutability to normalize the mutation data, Dayhoff arrived at a matrix that "gives the probability that the amino acid in column *j* will be replaced by the amino acid in row *i* after a given evolutionary interval."[21] The non-diagonal elements of the PAM have the values

$$M_{ij} = \frac{\lambda m_j A_{ij}}{\sum_i A_{ij}},$$

where A is the matrix containing the point mutation values, m is the relative mutability of each amino acid, and λ is a proportionality constant. The elegance of Dayhoff's scheme is that it is possible to simulate different periods of evolutionary time by multiplying the matrix by itself—a single PAM matrix corresponds to the amount of time in which

each amino acid has a 1% chance of mutation. For instance, multiplying PAM by itself 20 times—often the result is called PAM20—yields a matrix in which each amino acid has a 20% chance of mutating. As Dayhoff was aware, using matrices such as PAM250 can be extremely helpful in detecting distant evolutionary relationships between proteins.

Dayhoff's larger aim was to use such models of mutation to explore the evolutionary relationships among all proteins. From the fifth edition onward (1973), the *Atlas* was organized using the concept of protein "superfamilies," Dayhoff's second major analytical contribution. Families of proteins were already well recognized and easily determined through simple measurements of sequence similarity. The sensitivity of Dayhoff's methods of comparison (using the PAMs), however, allowed her to sort proteins into larger groups, organized according to common lines of descent.[22] Such classifications were not merely an organizational convenience—they provided theoretical insight into the process of evolution. The ultimate aim of the NBRF's sequence collection work was this kind of conclusion:

> In examining superfamilies, one is struck by the highly conservative nature of the evolutionary process at the molecular level. Protein structures persist through species divergences and through gene duplications within organisms. There is a gradual accumulation of change, including deletions and insertions as well as point mutations, until the similarity of two protein sequences may no longer be detectable, even though they may be connected by a continuum of small changes.[23]

The superfamily concept was both a tool of classification and a biological theory. It was a way of conceptualizing the relationships among the entities that made up living things and of making sense of their history. In an article published in *Scientific American* in 1969, Dayhoff outlined some of the conclusions of her work on the classification and history of life: "The body of data available in protein sequences," she argued, "is something fundamentally new in biology and biochemistry, unprecedented in quantity and in concentrated information content and in conceptual simplicity . . . because of our interest in the theoretical aspects of protein structure our group at the National Biomedical Research Foundation has long maintained a collection of known sequences. . . . In addition to the sequences, we include in the *Atlas* theoretical inferences and the results of computer-aided analyses that illuminate such inferences."[24]

Understanding Dayhoff's databasing and collection efforts requires understanding of the computational-theoretical practices in which they were embedded. Although Dayhoff's database was not distributed electronically (it was available on magnetic tape from 1972, but only a handful of tapes were sold[25]), it was stored in computer-readable form, and all the data processing was performed digitally. The *Atlas* was something fundamentally new because it was not just a collection, but provided a system and a means for ordering, classifying, and investigating the living world without doing bench-top experiments. Producing PAMs, defining superfamilies, and generating phylogenetic trees from the sequences were integral parts of the process of producing the *Atlas*. These activities, which were woven into the production and structure of the *Atlas* itself, made it more than a means of collecting and redistributing data; rather, it was a way of organizing, systematizing, and creating biological knowledge.

Bruno J. Strasser argues that Dayhoff's collection efforts (much like botanical gardens of the eighteenth and nineteenth centuries) relied on creating a "network of exchange" or a "Maussian system of gift and counter-gift," but that this system conflicted with "ideas about credit, authorship, and the property of knowledge in the experimental sciences."[26] In particular, Dayhoff's collection and use of other researchers' experimental work (some of it unpublished) conflicted with the dominant norms in biochemistry and molecular biology, in which one's own work was one's own property (particularly if it was unpublished). This conflict manifested itself in several ways. First, it meant that researchers were, by and large, uncooperative—experimenters were reluctant to share their unpublished sequences with the NBRF. Second, Dayhoff had trouble receiving scientific credit for her work. John T. Edsall commented on Dayhoff's prospects for election to the American Society of Biological Chemists:

> Personally I believe you are the kind of person who should become a member of the American Society of Biological Chemists . . . but knowing the general policies that guide the work of the Membership Committee I must add that I can not feel at all sure about your prospects for election. Election is almost invariably based on the research contributions of the candidate in the field of biochemistry, and the nomination papers must include . . . recent work published by the candidate, to demonstrate that he or she has done research which is clearly his own. The compilation of the *Atlas of Protein Sequence and Structure* scarcely fits into this pattern.[27]

Dayhoff's *Atlas* was considered by some to be nothing more than a mere aggregation of others' work.

No doubt some of Dayhoff's problems stemmed from researchers' reluctance to share unpublished data. At a deeper level, though, this reluctance stemmed from a misunderstanding of Dayhoff's project. As Edsall's attitude suggests, Dayhoff's work was understood as the unoriginal work of collection and compilation, rather than as an attempt to systematize biological knowledge. Indeed, Dayhoff complained about the "great hostility of journal reviewers" when she tried to present her work as a theoretical contribution to biology.[28] No doubt this had to do with the generally marginal status of theory within biology, and with the prevalent notion that any such theory should look like a mathematical theory in physics, rather than a system of categorization or a database.

Ultimately, after struggling to maintain funding for her *Atlas*, in 1981, Dayhoff and the NBRF failed to win the contract from the NIH to build and maintain a national sequence database (as described in chapter 1, the contract was awarded to Walter Goad at Los Alamos). This failure was a harsh blow for Dayhoff, who had struggled for over a decade to gain support and recognition for her work. The lack of adequate funding had forced the NBRF to charge research biologists a fee for the *Atlas*. This, in turn, embittered the biological community, who saw the NBRF as taking their own work (for free) and selling it for a profit.

The NIH's decision was based on the conclusion that Dayhoff did not have the technical expertise to build and run a modern database.[29] It was Dayhoff, however, who had pioneered the idea of organizing biological data into computerized databases. Although GenBank, as we shall see in the next section, placed a far greater emphasis on using electronic means to collect and communicate data, the notion of using a structured digital space to order biological knowledge and create models of the biological world was Dayhoff's.

Dayhoff created a model for studying evolution. The use of sequence data in conjunction with the PAM matrices and mathematics developed by Dayhoff and her collaborators made it possible to apply evolutionary theory to make specific predictions about the relatedness of species and hence about the history of life. In other words, it was a way of making biological knowledge—without the laboratory or the field—through the structuring and ordering of data.

Dayhoff's principal innovation was not the collection of the sequences, but the use of this collection to investigate biology without doing lab experiments. Because the *Atlas* was largely distributed on paper, this type of investigation was at first mostly limited to the NBRF. As GenBank developed mechanisms for electronic distribution (via magnetic tape and over telephone-based networks), such practices spread.

GenBank

Like Dayhoff's work, the early history of GenBank must be embedded within a culture of practice—databases were developed not just as collections or repositories of data, but as tools for performing specific kinds of biological work. In other words, they were active sites for the development of biological knowledge. An account of the events that led to the creation of GenBank has been given by Temple Smith, who was closely involved with the events he describes.[30] Smith ascribes the advent of sequence databases to the coincidentally simultaneous invention of techniques for sequencing DNA and of mini- and bench-top computers. Although he describes some of the problems encountered by the early databases, he emphasizes that the founders "foresaw both the future needs and the potential of databases."[31]

The advocates of GenBank certainly saw the value of creating a repository for nucleotide sequences in order to manage the output of large-scale sequencing efforts, but they had to do much work to convince potential funders and other biologists of its value. Those actively managing the databases had to make the case that they were far more than collections; they argued that databases should be dynamic structures and tools through which a new kind of biology could be practiced. To most biologists, a database meant little more than an archive, not an important tool for basic research. The caution with which the NIH approached databases led to the construction of a "flat-file" structure for early versions of GenBank. Even this flat-file database, however, had important consequences for how biologists were able to construe and construct the relationships between biological entities.

In addition to Dayhoff's efforts at the NBRF, several other biological database efforts had been inaugurated by the late 1970s. In 1973, protein X-ray crystallographic data collected by Helen Berman, Olga Kennard, Walter Hamilton, and Edgar Meyer had been made available through Brookhaven National Laboratory under the direction of Thomas Koetzle.[32] The following year, Elvin Kabat, an immunologist at Columbia University, made available a collection of "proteins of immunological interest" (largely immunoglobulins) via the PROPHET computer system.[33] By the end of the 1970s, there was sufficient interest in biological databases to attract about thirty-five scientists to a meeting on the subject organized by Norton Zinder, Robert Pollack, and Carl W. Anderson at Rockefeller University in March 1979. A summary of this meeting circulated the following year within the NIH, listing the reasons why a nucleic acid sequence database was needed:

> 1) the rapidly increasing rate at which nucleic acid sequence information is becoming available (approaching 10^6 nucleotides per year); 2) the wide range of biological questions that can be asked using a sequence data base; 3) the fact that only a computer can efficiently compare and transform the data base to ask questions of interest; 4) the desirability of avoiding a duplication of effort in both adding to the data base and analyzing it; 5) the desirability of correlating a nucleic acid sequence data base with other features of biological importance including mutations, natural species variation, control signals, protein sequence and structure, nucleic acid secondary and tertiary structure.[34]

For the workshop participants, the main point of the database was to "ask questions of interest" and to "correlate" the sequence data with other sorts of biological information. It was not supposed to be an archive or a stand-alone repository. But the applicability of computers, and particularly computer databases, for asking and answering biological questions was not universally acknowledged. The NIH had not up to this time funded biological databases, and it had to be convinced that the effort was worthwhile. The fact that a report of the Rockefeller meeting took over eighteen months to reach the NIH is perhaps indicative of the priority that it was accorded.

Moves toward a database continued to proceed slowly. Dayhoff, Goad, Frederick Blattner (from the University of Wisconsin), Laurence Kedes (from Stanford University Medical Center), Richard J. Roberts (from Cold Spring Harbor Laboratories), and a few others were pushing for the NIH to fund a database effort. In July 1980, Elke Jordan and Marvin Cassman from the National Institute of General Medical Sciences (NIGMS) convened a further workshop to discuss prospects for a database. In contrast to the report of the Rockefeller meeting, the official report stated only that "an organized effort must be initiated to store, catalog, and disperse" nucleotide sequence information.[35] Around the middle of 1980, there was considerable uncertainly as to whether

any federally funded database effort would proceed. Jordan and Cassman received numerous letters supporting the proposed database from molecular biologists around the country. The correspondence argued for the database on the grounds that it would act as a powerful organizing resource for biology as well as a repository:

> There appears to be some question as to the utility of a national DNA sequence analysis and databank facility. We wish to express our strong support in this matter. . . . In our laboratory, we have used Seq [a sequence analysis program available at Stanford], for example, to locate transcripts from an *in vitro* transcription system when we could not find them ourselves. . . . Such a system for DNA sequence analysis would open a new way of thinking about sequence analysis for researchers who do not now have access to a computing center or staff available to maintain a local facility.[36]

The database would not be just a library or an information-sharing scheme, but provide a "new way of thinking" about sequences for molecular geneticists.

By mid-1980, in order to encourage the NIH to act, both Dayhoff and Goad had begun pilot nucleotide sequence banks (no doubt they both also hoped to improve their own chances of winning any NIH contract that might be tendered). As described in chapter 1, Goad was a theoretical physicist by training, and after working on nuclear weapons, he became interested in molecular biology in the mid-1960s. At Los Alamos, he assembled a small group of mathematicians, physicists, and biologists to work on problems of protein and nucleotide sequence analysis. Already by December 1979, Goad and his team had written a proposal for a "national center for collection and computer storage and analysis of nucleic acid sequences" based on their pilot project. The aims of such a facility were clearly set out:

> The discovery of patterns inherent in base sequences can be aided by computer manipulation to an even greater extent than for either numerical relationships (where there is a natural ordering) or natural language text (where we are habituated to certain patterns). . . . The library would be invaluable for relating sequences across many biological systems, testing hypotheses, and designing experiments for elucidating both general and particular biological questions. . . . The development of methods capable of answering the most penetrating questions will result from dedicated, ongoing research combining mathematics, computer science and molecular biology at a high level of expertise and sophistication.[37]

By this time the pilot project contained about 100,000 bases. More importantly, though, its sequences were not only embedded within a sophisticated set of programs for performing analysis, but also "arranged in a number of tables for access and manipulation."[38] The team at Los Alamos had adapted a system called FRAMIS, developed by Stephen E. Jones at Lawrence Livermore National Laboratories, that allowed sequence data and the associated biological information to be linked together by sophisticated logical and set theoretic operations.[39] Although this system was difficult to implement (compared with just listing the sequences one after another in a file), the advantage of storing sequence data and other biological information in such a way was that it allowed relationships to be rearranged or information added at a later point without having to alter each individual database entry.

During the late summer and fall of 1980, several unsolicited proposals were made to the NIH. On August 13, Dayhoff requested funds to expand her pilot project; on August 28, Douglas Brutlag, Peter Friedland, and Laurence Kedes submitted a proposal that would turn their MOLGEN project into a national computer center for sequence analysis; on September 3, Los Alamos submitted a revised proposal based on its pilot DNA data bank; and on September 8, Michael Waterman and Temple Smith submitted a supplementary proposal for sequence analysis.[40] The NIH, however, continued to hesitate. Jordan convened a follow-up to the July meeting on October 26. Notes from this meeting made by Frederick Blattner indicate that the decision had been made to segregate the databasing efforts into two separate projects: the first would focus on collection and distribution of the sequence data, and the second on software and more sophisticated analysis tools with which to manage and use these data. Blattner's early sketch of John Abelson's proposed "planning structure" divided the database project between "data collection groups" and "programming": the data collection was to be "annotated, but not sophisticated."[41] The agenda for the third meeting, held in early December, already included a detailed breakdown of the tasks to be performed under the two separate contracts. The scope of work for the first project was to "acquire, check and organize in machine-readable form the published data concerning base sequences in polynucleotides," while the efforts to develop a "database

management system . . . for the sequence data that allows sophisticated search capabilities comparable to relational data base" were reserved for the second.[42] Although the NIH intended the two contracts to be contemporaneous and closely connected, by the time the request for proposals was finally made (near the end of 1981), only the first was to be funded.

Dayhoff, Goad, and a small group of other computer-savvy biologists realized that a nucleotide sequence database had to be a sophisticated theoretical apparatus for approaching biological problems. The majority of their colleagues, however, while realizing the importance of a repository, believed that making a database was essentially the trivial process of reading old journal articles and typing in the sequences. The NIH, reflecting this latter view, attempted to create a database with this simple model in mind. For many, the data bank was a "service" and therefore dubiously worthy of federal support under the aegis of basic research. Those at the NIGMS who supported the project had to work hard to generate financial support by stressing the wide range of researchers, including academic, industrial, and medical, who would use the database for basic research.[43] Moreover, Jordan and her co-workers promised that the intention of the funding was only to effect a "start-up" and that it was anticipated that the database would ultimately be supported by user charges.[44] Like lab apparatus or journal subscriptions, the biological database was understood to be something that researchers could pay for out of their own budgets. While providing support for basic researchers, it was not an activity that would contribute fundamentally to biological understanding.

The NIH issued a request for proposals for a nucleic acid sequence database on December 1, 1981. Three proposals were forthcoming: one from Dayhoff and the NBRF, one based on a collaboration between Los Alamos and IntelliGenetics (a company based in Palo Alto, California, and run by Stanford biologists and computer scientists), and a further joint proposal between Los Alamos and Bolt, Beranek and Newman (BBN) of Cambridge, Massachusetts.[45] On June 30, 1982, the NIGMS announced that a contract of $3.2 million (over five years) had been awarded to BBN and Los Alamos. Los Alamos was to be responsible for collecting sequences from the published record, while BBN was to use its expertise in computation to translate the data into a format suitable for distribution by magnetic tape and over dial-up connections to the PROPHET computer (an NIH-funded machine based at BBN). The NBRF was especially disappointed by this decision; others in the community, too, were concerned about the choice of a nonacademic institution to manage the distribution efforts.[46]

Despite the fact that Goad's pilot project had used a sophisticated database structure, the NIH insisted that the new data bank—which would become GenBank—be built as a "flat file." A flat file is a text-based computer file that simply lists information about nucleotide sequences line by line. Each line begins with a two-letter code specifying the information to be found on that line—"ID" gives identifying information about the sequence, "DT" gives the date of its publication, "KW" provides keywords, "FT" lists features in the sequence, and the sequence itself corresponds to lines beginning with "SQ." Different sequences could be listed one after another in a long text file separated by the delimiter "//" (figure 5.1).

The NIH held the view that the GenBank format should be readable both by computers and by humans. By using the two-letter line identifiers, a simple program could extract information from the flat-file entries. A major disadvantage of this format, however, was the difficulty involved in updating it. If, for instance, it was decided that it was important to add a further line including information about the type of sequencing experiment used to generate the sequence, the database curators would have to modify each sequence entry one by one. Moreover, a flat file does not lend itself to the representation of relationships between different entries—the list format makes it impossible to group entries in more than one way or to link information across more than one entry.

The flat-file format was suited to the NIH's notion that a nucleotide database should be no more than a simple collection, a laundry list of sequences. However, it also embodied a particular way of understanding biology and the function of genes. George Beadle and Edward Tatum's "one gene–one enzyme" hypothesis is considered one of the founding dogmas of molecular biology. Although the idea (and its successor, "one gene–one polypeptide") had been shown to be an oversimplification even by the 1950s, the notion that it is possible to understand life by considering the actions of individual genes exerted a profound influence on at least forty years of biological research.[47]

In the late 1970s, as a result of the sequencing methods invented by Allan Maxam, Walter Gilbert, and Frederick Sanger, the possibility of discovering the mechanism of action of particular genes seemed within reach. Some molecular geneticists began to focus their efforts on finding and sequencing the genes responsible for particular diseases, such as

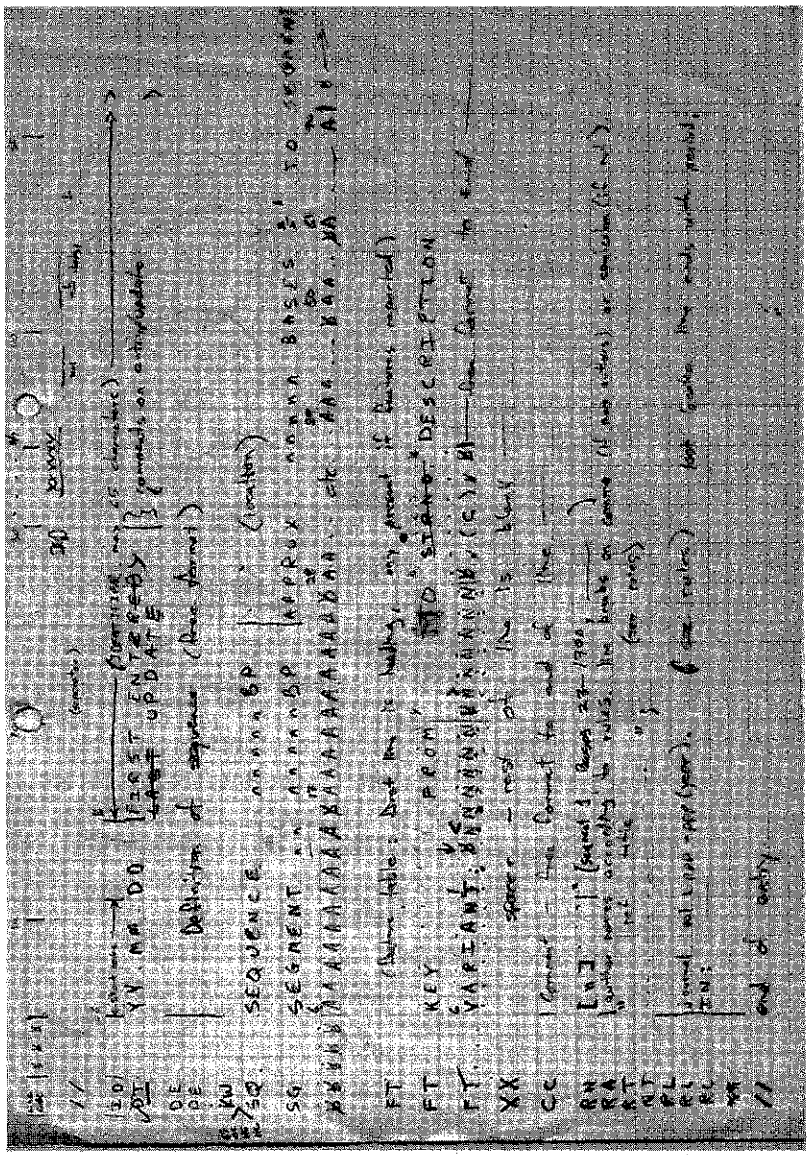


FIGURE 5.1 Sketch of flat-file structure for EMBL-Bank (European Molecular Biology Laboratory Data Bank). The EMBL-Bank flat-file structure was very similar to the one used for GenBank. Note the capital letters on the left ("SQ," "ST," "FT") that define the content of the

cystic fibrosis or Duchenne muscular dystrophy.[48] For such an activity, a flat-file database was an appropriate resource: each sequence entry in the database could be roughly equated with one gene, and that gene had a specific, definite, and singular effect on the biology of the organism in which it dwelt. The NIH imagined that researchers would use the database primarily as a central repository or archive into which sequences would be deposited once and for all; for the most part, sequences would only need to be retrieved one at a time in order to make comparisons with experimental work.[49] Similar sequences usually resulted in similar protein structures with similar functions; hence, matching an unknown sequence to a known sequence in the database could provide invaluable information. For such activities, the gene of interest could be simply compared with the long list of entries in the database one by one. The flat-file database structure was ideal for this kind of search operation; it entailed and represented a theory of how sequence elements acted to produce biological effects.

Once GenBank began operations in July 1982, it became clear to those doing the work of collection and distribution at Los Alamos and BBN that the database was attracting a far wider scope of use. As well as revealing the sequences of genes, the new sequencing technologies had an unexpected consequence: they allowed biologists to sequence not only individual genes, but also regulatory regions, structural RNA-coding regions, regions of unknown function, and even whole genomes (at first limited to small genomes such as those of viruses or cloning vectors). This meant that a sequence in the database did not necessarily correspond neatly to a single gene. Molecular geneticists began to realize that not all the information necessary to understand gene action was contained within the gene sequence—how the gene was spliced, where it was expressed, and how it was phosphorylated were also crucially important.[50] In the flat-file format, such information was contained within the "Features" table for each entry. The Features table consisted of a single line for each feature, as in this example:

FT	firstexon	EXON	273–286
FT	tatabox	TATA	577–595

The three columns identified the name of the feature (such as "firstexon"), the feature type (here an exon or a TATA box), and the coordinates in the sequence at which that feature was to be found.[51] Entering this in-

formation from the published literature (it was often necessary to read an entire article or even several articles) and keeping it up to date was a gigantic task. But biologists often wanted to use the database to retrieve and aggregate information located across many entries. For instance, a biologist might want to find all the protein-coding sequences in the database that contained exons with a size greater than 100 kilobases. An excerpt from a long list of criticisms of GenBank reads:

> The BB&N [GenBank] retrieval system is not suited to this scientific area. Modern systems permit the user to construct current lists of entries retrieved on various criteria and to perform manipulations on these sequences. The organization of the BB&N system is archaic, because it does not readily permit these manipulations.[52]

The flat file and features table were not well adapted to sophisticated cross-entry queries. Moreover, as biologists produced more and more sequence, it was inevitable that sequences began to overlap; in order for this work to be useful, the database had to identify such overlaps and organize the data in a way that represented these fragments. Another user wrote to Los Alamos complaining that the flat-file data format was not always consistent enough to be computer readable and suggesting "a language for reliably referring to sections of other entries in the database. If this language is sufficiently powerful, many of the synthetic sequences could be expressed in this form."[53] In other words, the user wanted the database to be organized so as to allow the linkages between different entries and different sequences to be made manifest.

The result of these demands was that GenBank was unable to keep pace with the publication of sequences, and particularly with the kinds of annotations that were supposed to appear in the Features table. By 1985, it took an average of ten months for a published sequence to appear in the database. This was not only an unacceptably long delay from the point of view of researchers, but also stood in breach of GenBank's contract with the NIH (which required sequences to be available within three months). A progress report from early 1985 explained the problem:

> Since the inception of GenBank . . . there has been a rapid increase in both the rate at which sequence data is reported and in the complexity of related information that needs to be annotated. As should be expected, many reported sequences repeat, correct, extend, or otherwise relate to previous work, and as a result a substantial number—in fact, a majority—of database entries have to be updated each year; thus GenBank is not an archival operation such that an entry, once made, just stays in place.[54]

By this time, some members of GenBank's scientific advisory panel considered the growing backlog an "emergency."[55] Los Alamos responded by requesting more money to employ more "curators" to enter data from the published literature. Goad and his co-workers, however, realized that the root of the problem was that the structure of the database was increasingly inadequate for the needs of biological research. James Fickett and Christian Burks, leading contributors to the Los Alamos effort, argued that "the scope and interconnectedness of the data will grow at a pace hard to keep abreast of," and that consequently, the greatest challenge would be to "organize the data in a connected way."[56]

Because the NIH saw the nucleotide sequence database as a mere archiving activity, they attempted to create an atheoretical database. This was impossible: even the minimalist flat file encoded a particular structure, a particular way of doing biology, and a particular idea about how sequences related to organismic function. The flat-file structure instantiated an ordering of biological elements based on the one gene–one enzyme hypothesis. During the early 1980s, that hypothesis was in the process of being displaced and superseded by other ideas about how biology worked.

Biological Relations

The original GenBank contract ran for five years, expiring in September 1987. As that date approached, two concerns were paramount. First, GenBank continued to struggle to remain up to date in entering sequence information from journals.[57] Second, it was clear that the structure of the database required a significant overhaul. As such, NIH's new request for proposals specified that the contractor for the next five-year period would develop a new system whereby authors would be able to submit their sequence data directly in electronic form (preferably over a dial-up telephone network). In addition, the contractor would be obligated to find ways to increase the cross-referencing of the data and to make sure that "new data items which become important can be added to the data base without restructuring."[58] The NIH received three "competitive" proposals for the new contract: one from BBN, one

from DNAStar (a company based in Madison, Wisconsin), and one from IntelliGenetics. Each of the contractors would subcontract with Los Alamos. Of singular importance in the eventual decision to award the contract to IntelliGenetics was the perception that it, more than BBN, was in touch with the needs of the biological research community. IntelliGenetics had close ties to the molecular biologists at Stanford—particularly Douglas Brutlag—and had successfully run BIONET, a network resource for providing software tools for biologists, since 1983.[59] No doubt the NIH hoped that a greater awareness of the research needs of molecular biologists would translate into a more usable and flexible database system.

At around this time, many biologists were beginning to think about biology in new ways. The first plans for determining the sequence of the entire human genome were made at a meeting in Santa Fe, New Mexico, in 1986.[60] Even at this early stage, the planners of what came to be called the Human Genome Project (HGP) realized the need for "computational technology" capable of "acquiring, storing, retrieving, and analyzing" the sequence data.[61] Since both Los Alamos and the early stages of the HGP were funded and organized by the Department of Energy, GenBank personnel were well aware of the plans for a massive scaling up of sequencing efforts and the effect that it could have on their already strained ability to get data into the database in a timely fashion. Those advocating the HGP were soon talking to Goad and other GenBank staff about the demands that their project would place on GenBank. By 1988, James Watson, in his capacity as director of the National Center for Human Genome Research (NCHGR), was well aware of the importance of GenBank for the HGP:

> Primary products of the human genome project will be information—genetic linkage maps, cytological maps, physical maps, DNA sequences. This information will be collected and stored in databases, from which it will be made available to scientists and clinicians. In this sense, the raison d'etre of the genome project is the production of databases.[62]

Los Alamos and IntelliGenetics too realized that data coming from the HGP would not only strain the capacity of their staff, but also require thoroughgoing structural changes. In 1985, the complete sequence of the Epstein-Barr virus (about 170,000 bases) had already caused trouble for BBN's computers.[63] In 1988, a "technical overview" of GenBank reported that the addition of human genomic data would require the

database to "store entirely new types of data that could not be easily integrated into the original structure."[64] As plans for the HGP (and other smaller genome projects) were developed, the concept of what a sequence database was, and what it could be used for, had to be rethought.

The flat-file database, much like the early file management systems, created a rigid ordering of entries with no explicit cross-linking possible. A relational model would impose different kinds of orderings on the data. The 1988 technical overview of GenBank justified the change to a relational model on the following bases:

> One, because the domain of knowledge we are dealing with is extremely dynamic at this point in history, we had to expect our understanding of the data to change radically during the lifetime of the database. The relational model is well suited to such applications. Two, even if our view of the inherent structure of the data did not change, the ways in which the data could be used almost certainly would change. This makes the ease of performing ad hoc queries extremely important.[65]

By the end of 1986, GenBank staff at Los Alamos had worked out a structure to implement GenBank in relational form. Their plan was set out in a document titled "A Relational Architecture for a Nucleotide Sequence Database," written by Michael Cinkosky and James Fickett.[66] The schema included thirty-three tables that described the sequence itself, its physical context (for instance, its taxonomy or the type of molecule it represented), its logical context (features such as exons, genes, promoters), its citations, and pertinent operational data (tables of synonyms). Tables could be modified or added to (or extra tables could even be added) without disrupting the overall structure or having to amend each entry individually.

The descriptions of the "sequences" and "alignments" tables are reproduced here. Each sequence is given an accession number that acts as the primary key for the table. The "publication_#" and "reference_#" keys link to a table of publications, and "entered_by" and "revised_by" keys link to tables of people (curators or authors). As is noted in the description, such sequences may not correspond to actual physical fragments—that is, they may not represent a particular gene or a particular sequence produced in a sequencing reaction. Rather, the relationship between sequences and physical fragments is "many-to-many": a fragment may be made up of many sequences, and any given sequence may be a part of multiple fragments. In other words, there is no straight-

forward relationship between DNA sequences as they appear in the database and objects such as "genes" or "exons" or "BAC clones."

```
TABLE sequences
UNIQUE KEY (sequence_#)
INDEX KEYS (publication_#, reference_#), (entered_by, entered_date)
sequence_#      REQ     /* accession number for the sequence */
sequence        REQ     /* the sequence itself */
length          REQ     /* redundant, but convenient */
topology        OPT     /* circular, linear, tandem, NULL-unknown */
publication_#   OPT     /* next two give bibliographic source */
reference_#     OPT
entered_date    OPT     /* next two give history of initial entry */
entered_by      OPT
revised_date    OPT     /* next two give history of revision */
revised_by      OPT
```

DESCRIPTION. The reported sequences. There can be at most one citation, so it is given here. But the relationship to physical fragments can be many-many, so that is given in a separate table.

```
TABLE alignments
UNIQUE KEY (alignment_#, sequence_1, left_end_1, sequence_2)
alignment_#     REQ     /* accession number for alignment */
sequence_1      REQ     /* next three specify first interval to align */
left_end_1      REQ
right_end_1     REQ
sequence_2      REQ     /* next three specify second interval to align */
left_end_2      REQ
right_end_2     REQ
preference      OPT     /* 1 or 2; which one to prefer */
type            OPT     /* conflict, revision, allele, etc. */
```

DESCRIPTION. Give an alignment of any number of sequences by specifying pairs of intervals for the line-up. One record of this table gives a pair of intervals, one from each of two sequences. The set of all records with a given alignment number gives a complete alignment.

This structure for storing sequence data allows objects of interest to be reconstructed from the sequences in multiple ways as needed. The second table shown here—"alignments"—allows different entries in the "sequences" table to be stitched together in multiple ways by referring to their sequence accession numbers and coordinates. For example, it would be possible to create an alignment that spliced sequence A to sequence B, or the first 252 base pairs of sequence A to the last 1,095 base

pairs of sequence B. With sufficiently sophisticated queries, it would be possible to join not only sequences, but also any features described in the tables (for example, to join all the exons from given sequences to reproduce a protein-coding region). Sequence data could be linked together dynamically by the user in a flexible manner. But within this flexibility, this relational structure emphasizes the rearrangement of sequence elements. If the flat-file structure was gene-centric, the relational database was alignment-centric. It was designed to make visible the multiple possible orderings, combinations, and contexts of sequence elements.

By 1989, over 80% of GenBank's data had been imported into the relational database.[67] The HGP and the relational sequence database could not have existed without each other—they came into being together. GenBank and the HGP became mutually constitutive projects, making each other thinkable and doable enterprises. Moreover, just as flat files had, both genome projects and relational database systems embodied a particular notion of biological action: namely, one centered on the genome as a densely networked and highly interconnected object. In 1991, when Walter Gilbert wrote of a "paradigm shift" in biology, he argued that soon, "all the 'genes' will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical."[68] This "theory" was built into the structure of the database: phenotype or function does not depend on a single sequence, but rather depends in complicated ways on arrangements of sets of different sequences. The relational database was designed to represent such arrangements.

During the 1990s, biologists investigated the "added value that is provided by completely sequenced genomes in function prediction."[69] As the complete genomes of bacterial organisms, including *Haemophilus influenzae*, *Mycoplasma genitalium*, *Methanococcus jannaschii*, and *Mycoplasma pneumoniae*, became available in GenBank, biologists attempted to learn about biological function through comparative analysis. The existence of orthologs, the relative placement of genes in the genome, and the absence of genes provided important insights into the relationship between genotype and phenotype.[70] The important differences among the bacteria and how they worked were not dependent on individual genes, but on their arrangements and combinations within their whole genomes. But this was exactly what the relational structure of GenBank was designed to expose—not the details of any particular sequence, but the ways in which sequences could be arranged and combined into different "alignments."

GenBank as a relational database provided a structure for thinking

about biology through the genome. It made possible orderings and reorderings of biological elements and reinforced biologists' notion that function depends on multiple sequence elements acting together in interconnected ways.

NCBI And Biological Databases in the Genomic Age

In his opening remarks at the celebratory conference marking the twenty-fifth anniversary of GenBank in 2008, Donald Lindberg remembered the transformative effect of a paper published in *Science* by Renato Dulbecco. Dulbecco argued that sequencing the human genome would be a national effort comparable to the "conquest of space." This argument convinced Lindberg, who was the director of the National Library of Medicine (NLM) at the NIH, that the genome project had to be undertaken and that the NLM should play a key role. This commitment was reflected in the NLM's "Long Range Plan" for 1987:

> Currently no organization is taking the leadership to promote keys and standards by which the information from the related research data bases can be systematically interlinked or retrieved by investigators. The full potential of the rapidly expanding information base of molecular biology will be realized only if an organization with a public mandate such as the Library's takes the lead to coordinate and link related research data bases.[71]

During 1986 and 1987, Lindberg worked to convince Congress of the importance of this mission. The campaign was taken up first by Representative Claude Pepper (D-Florida), who introduced the National Center for Biotechnology Information Act of 1986. This bill would give the NLM the responsibility to "develop new communications tools and serve as a repository and as a center for the distribution of molecular biology information" (H.R. 99–5271). The NLM circulated a document on Capitol Hill, titled "Talking One Genetic Language: The Need for a National Biotechnology Information Center," that made the case for the new center.[72] Pepper reintroduced the bill with minor modifications in the next session of Congress (H.R. 100–393), while Senator Lawton Chiles introduced similar legislation into the Senate on June 11, 1987 (S. 100–1354).[73] The bill entered Congress at the same time the debates about the HGP were taking place (Senator Pete Dominici [R-New Mexico] introduced legislation to fund the HGP on July 21). The bill was amended once more and introduced a third time by Senators Chiles, Dominici,

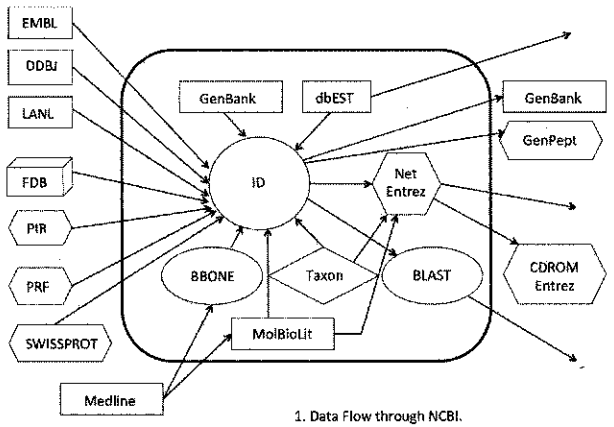
Ted Kennedy, and others in December 1987 (S. 100–1966). Hearings were held on February 22, 1988, at which Victor McKusick, James Wyngaarden (director of the NIH), and Lindberg testified. Supporters of the bill had closely connected it to the HGP, portraying the need for biotechnology information coordination as central to the project and important for American competitiveness in biotechnology. As support for the HGP grew, the bill's passage became more likely; it was signed into law by President Reagan on November 4, 1988. It provided for the creation of a National Center for Biotechnology Information (NCBI), under the auspices of the NLM, with twelve full-time employees and a budget of $10 million per year for fiscal years 1988 through 1992.[74]

Lindberg conceived the role of the NCBI not as a replacement or supplement for GenBank, but as a way to bring order to the different kinds of biological information and databases that had begun to proliferate. In his testimony in support of the legislation, Donald Fredrickson, president of the Howard Hughes Medical Institute, argued that the NCBI was necessitated by the fact that "not only are the databases being flooded with information they cannot manage, but each database uses a different information system or computer language. We have created a sort of Tower of Babel."[75] "Talking one genetic language" characterizes how the NCBI sought to coordinate diverse sorts of biological information from many sources and at many levels, from cell types to pharmaceuticals. By the time funds for the NCBI were appropriated, Lindberg had already recruited David Lipman to direct the new center. Lipman had been working in Bethesda since 1983 and was already widely respected in the small community of computational biologists for his contribution to sequence-matching algorithms. In the existing biological databases, Lipman saw a tangled mess of overlapping systems and overly complicated schemas; he brought a youthful energy to the task of integrating databases and restoring sense and simplicity to GenBank and other biological information resources.[76] Under Lipman's direction, the NCBI moved quickly to take over GenBank, arguing that its mission to integrate and link databases required close control.[77] By October 1989, it had been agreed that after the end of the current GenBank contract, control of the database would be passed from NIGMS to NCBI—it would be managed in-house rather than under contract to a third party.[78] NCBI took over the task of collecting nucleotide sequence data as Los Alamos' role was phased out.

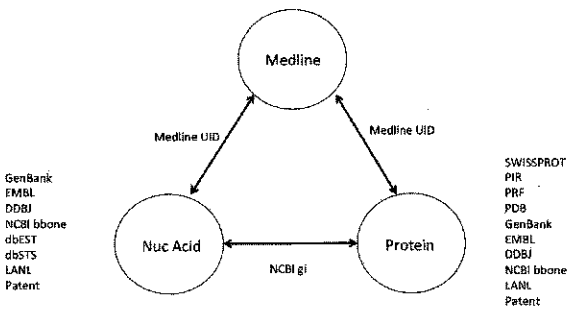
Before GenBank formally arrived at NCBI in 1992, efforts were already under way to fundamentally change its structure. Commensurate with the overall mission of the NCBI, the aim was the make GenBank

data more amenable to integration and federation with other data types. The NCBI data model was the work of James Ostell, who had been asked by Lipman to join NCBI as its chief of information engineering in 1988.[79] Ostell needed to solve two problems. The first was how to make data available to the widest possible number of biological users by ensuring that they could be shared across different computer platforms. Ostell's solution was to adopt an international standard (ISo8824 and ISo8825) called ASN.1 (Abstract Syntax Notation 1). Like the hypertext transfer protocol (HTTP) used on the Internet, ASN is a way for computers to communicate with one another—it specifies rules for describing data objects and the relationships between them. Unlike HTTP, however, it is not text-based, but renders data into binary code. ASN.1 was developed in 1984 for the purpose of structuring email messages; it describes in bits and bytes the layout of messages as they are transmitted between programs or between different computers. ASN.1 acts as a universal grammar that is completely independent of any particular machine architecture or programming language.[80] Ostell chose ASN.1 because "we did not want to tie our data to a particular database technology or a particular programming language."[81] Using ASN.1 meant that biologists using any programming language or computer system could use the GenBank database.

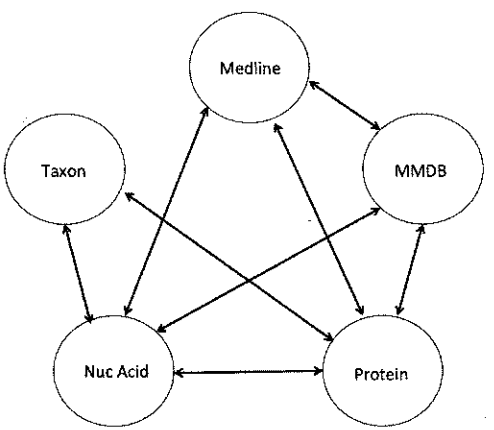
The second problem was to find a way of storing various kinds of data in a form that was suited to the needs of biologists who wanted not just DNA sequence information, but also data about protein sequence, protein structure, and expression, as well as information contained in the published literature. The scale of this problem of "heterogeneous sources" had become such that relational databases were no longer appropriate for such linking. "It is clear that the cost of having to stay current on the details of a large number of relational schemas makes this approach impractical," Ostell argued. "It requires a many-to-many mapping among databases, with all the frailties of that approach."[82] In other words, keeping the structure of each database consistent with the structure of a large number of others would quickly prove an impossible task. The alternative was to find a way to link the databases using ASN.1 via what Ostell called a "loose federation." The first such application, which became known as Entrez, used ASN.1 to link nucleic acid databases, protein sequence databases, and a large database of biomedical literature (MEDLINE). Wherever an article was cited in a sequence database (for instance, the publication from which the sequence was taken), the NCBI created a link to the relevant article in MEDLINE using the MEDLINE ID (figure 5.2). Likewise, NCBI created links



1. Data Flow through NCBI.



2. Entrez is three linked data spaces.



3. Five linked information spaces

FIGURE 5.2 Integration of data at NCBI. Diagrams show the emphasis on connections between different databases such as sequence, protein, MEDLINE (publications), dbEST (expressed sequence tags), and so on. (Ostell, "Integrated access." Reproduced by permission of IEEE.)

between nucleotide sequences and their translated protein equivalents. "Once the work has been done to get this kind of cross-reference into the databases," Ostell wrote, "it becomes relatively straightforward to tie them into a unified system using hypertext tools such as Mosaic."[83] The new structures for organizing biological information were closely connected to the new tools of the Internet.

But even building these kinds of hard links between databases was too difficult. For Ostell, the ultimate solution was to create a "data model." Like animal models, which biologists can use for the study of human diseases, or mathematical models, which they can use for describing forces inside cells, the NCBI data model provides a structure for sequence information that allows "meaningful predictions to be made and tested about the obviously much more complex biological system under consideration."[84] Ostell reasoned that the basic elements of the database should, as closely as possible, resemble the basic "facts" collected by biologists in the laboratory—that is, sequence elements. In the data model, sequences are represented as objects called "Bioseqs," which constitute a "linear, integer coordinate system." Importantly, the sequence information itself (the As, Gs, Ts and Cs) is not contained in the Bioseqs. Rather, a particular Bioseq contains coordinate-based instructions on how to build a sequence from fragmentary pieces of sequenced DNA or RNA. Such instructions could be: "Take *sequence1*, then a gap of unknown length, then *sequence3*, then a gap of 30 base pairs, then *sequence2*." The Bioseq may consist of full-length sequences, partial sequences, gaps, overlapping sequences, or genetic or physical maps, since all of these can be constructed by placing different sorts of objects along the coordinates of the Bioseq. This structure allows for a very different representation of biological data within the system:

> The GenBank flatfile format . . . is simply a particular style of report, one that is more "human-readable" and that ultimately flattens the connected collection of sequences back into the familiar one-sequence, DNA-centered view. [The NCBI data model] much more directly reflects the underlying structure of such data.[85]

Indeed, the aim of the data model was a "natural mapping of how biologists think of sequence relationships and how they annotate these sequences. . . . The model concentrates on fundamental data elements that can be measured in the laboratory, such as the sequence of an isolated molecule."[86] This system not only allowed the expression of very

complex relationships between sequences and pieces of sequences based on maps or alignments,[87] but also provided sophisticated and robust links to published scientific articles. Citations in various databases were mapped to MEDLINE via unique integer identification numbers. Appropriate software could rapidly search (across multiple databases) for objects that cited the same article and link those objects together, or it could go even further and make links based on keywords from the abstracts contained in MEDLINE. By rendering the model in ASN.1, the NCBI created a system that combined objects (DNA sequences, protein sequences, references, sequence features) from a variety of databases and manipulated them all with a common set of software tools.

DNA-centered relational databases provided more flexible ways to recombine and reorder sequences. ASN.1 and the data model permitted no static biological objects. Rather, it was assumed that the process of doing biology would involve recombination and reordering of different biological objects across a wide range of databases. Relational databases were a framework within which to investigate the properties of dynamically rearrangeable sequence elements. The data model was a framework within which to investigate genomes using a wide variety of other data and data types.

The data model has provided a framework for exemplary experiments of the postgenomic era. Although it was developed in 1990, it remains a powerful tool for moving biological investigation beyond the genome. As biologists began to realize the limitations of studying the genome in isolation, the data model demonstrated ways in which to integrate more and more kinds of biological data.

In 2005, the bioinformatician Hans P. Fischer called for "inventorizing biology"—capturing the entirety of information about an organism in databases. Genomes, transcriptomes, proteomes, metabolomes, interactomes, and phenomes should be characterized, entered into databases, and integrated. This new "quantitative biology" would transform drug discovery and allow us to understand human disease pathways. This vision of "tightly integrated biological data" would allow an engineering-like approach to biological questions—drug design or even understanding a disease would become more like building an aircraft wing.[88] In the postgenomic era, the organization and integration of biological information provides a structure or blueprint from which biologists can work. At the beginning of each year, *Nucleic Acids Research* publishes a "database issue" that provides an inventory of biological databases. In 2009, that list included 1,170 databases, including about 100 new entries.[89] The ways in which the information in those databases is con-

nected provide theories of biological action. It is now clear that the sequence of the genome alone does not determine phenotypic traits. Databases provide ways of linking genomic information to the other vast amounts of experimental data that deal with transcripts, proteins, epigenetics, interactions, microRNAs, and so on; each of those links constitutes a representation of how sequences act and are acted on in vivo to make life work.

Conclusions

Biological databases impose particular limitations on how biological objects can be related to one another. In other words, the structure of a database predetermines the sorts of biological relationships that can be "discovered." To use the language of Bowker and Star, the database "torques," or twists, objects into particular conformations with respect to one another.[90] The creation of a database generates a particular and rigid structure of relationships between biological objects, and these relationships guide biologists in thinking about how living systems work. The evolution of GenBank from flat-file to relational to federated database paralleled biologists' moves from gene-centric to alignment-centric to multielement views of biological action. Of course, it was always possible to use a flat-file database to link sequence elements or to join protein interaction data to a relational database, but the specific structures and orderings of these database types emphasized particular kinds of relationships, made them visible and tractable.

One corollary of this argument is that biological databases are a form of theoretical biology. Theoretical biology has had a fraught history. In the twentieth and twenty-first centuries, several attempts have been made to reinvent biology as a theoretical, and in particular a mathematical, science. The work of D'Arcy Thompson, C. H. Waddington, Nicolas Rashevsky, and Ludwig von Bertalanffy has stood out for historians.[91] The work of all these authors could be understood as an attempt to discover some general or underlying biological principles from which the facts of biology (or the conditions of life) might be derived and deduced. In the twentieth century, such efforts were almost completely overshadowed by the successes of experimental biology, and molecular biology in particular. As Evelyn Fox Keller recognizes, however, the increasing use of computers in biological research has relied on modes of practice that might be called theoretical. "In molecular analyses of molecular genetics," Keller argues, "observed effects are given meaning through the construction of provisional (and often quite elaborate) models formulated to integrate new data with previous observations from related experiments. As the observations become more complex, so too do the models the biologists must construct to make sense of their data. And as the models become more complex, the computer becomes an increasingly indispensable partner in their representation, analysis, and interpretation."[92] This description might apply equally well to biological databases as to the type of computer models that Keller describes. The structures and categories that databases impose are models for integrating and making sense of large sets of data. As categorizations of organisms, sequences, genes, transposable elements, exon junctions, and so forth, databases are built on sets of structures or principles of biological organization that are then tested in experiments. Far from being lists or collections of information, biological databases entail testable theories of how biological entities function and fit together.

This understanding of biological databases as models also demonstrates that the flow and ordering of data are central to the constitution of biological objects and knowledge in bioinformatics. Here we have once again followed the data into the structures and spaces inside computers. Databases, which summarize, integrate, and synthesize vast amounts of heterogeneous information, are the key tools that allow biologists to ask questions that pertain to large numbers of sequences, genes, organisms, species, and so on. Databases allow these objects to be constituted "out of sequence"—that is, brought into new orderings or relationships with one another that do not necessarily reflect their order in cells or on chromosomes. The form of such relationships is constrained, however—flat files and relational databases were not designed for biology, but rather have their own particular histories. The ways in which biological objects are related to one another have been conditioned by the structural possibilities and limitations of existing database models—that is, by the histories of databases themselves.