

COMMENTARY

Genomic Sequence Databases

MICHAEL S. WATERMAN

Departments of Mathematics and Molecular Biology,
University of Southern California, Los Angeles,
California 90089-1113

Databases are increasingly important to the progress of biology. There are physical and genetic map databases, nucleotide and protein sequence databases, and structural databases for nucleic acids and proteins. The central role of this information cannot be overemphasized. Important discoveries at the molecular level are now being felt in other areas such as cell biology and medicine. The quantity and importance of these data make it essential that they be collected in easily accessible databases. At present, the sequence databases are composed of small regions of closely studied sequence. It is anticipated that soon such databases will be composed mostly of long stretches of sequence that have not been the focus of detailed experimentation. Nucleotide sequence databases currently include DDBJ (DNA Data Bank of Japan), the EMBL Data Library, and GenBank, while protein sequence databases include JIPIDS (Asian and Oceania node of the International Protein Information Database), MIPS (Martinsreid Institute for Protein Sequence Data), and PIR (National Biomedical Research Foundation Protein Identification Resource).

It is probably important to realize at the outset that these databases will never completely satisfy a very large percentage of the user community. Today, the user community is made up mostly of molecular biologists but users also include a smaller number of people from chemistry, physics, medicine, the mathematical sciences, and other fields including those who develop software. The range of interests within biology itself suggests the difficulty of constructing a database that will satisfy all the potential demands on it. Evolutionary relationships between organisms, for example, constitute an important topic in biology, and these relationships are used to organize the sequence data. Many details of classification are not agreed upon by all biologists and the consensus changes as the data increase. Even the classification of organisms into kingdoms has been revised recently. In addition, the level of detail desirable for the specialist might be irrelevant to the rest of us. A molecular biologist studying regulation of gene expression will want to know the results of deletion experiments in the promoter region of a gene, while many other molecular biologists will be interested only in the gene sequence itself. There is virtually no end to the depth and breadth of desirable information of interest and use to the biological community. Any sequence database is a

compromise between presenting only sequence data and giving all known biological information, including full text of papers, relevant to the organism.

I believe that the main purpose of central macromolecular sequence databases is information storage and retrieval. The databases should contain the sequences along with some basic information. It is now possible to find literature references from the sequence databases. In the future it should be possible to locate related information in other databases. Entering new sequences into the databases requires the database staff to analyze and interpret the sequences and the associated scientific literature. As any database user knows, organism and gene names are useful, since sequences can be classified and annotated by these names. Alternatively, many people want a database that embodies a detailed interpretation of the associated biology. Such databases should be viewed as research projects, with only a subset of the information collected into the central databases. Along these lines, interpreting biological sequence data with computers is an art and science that is flourishing these days, but the techniques are so far from settled that at this time it would be counterproductive to distribute anything other than the most standard analysis programs as an integral part of a major database. Database search techniques and other computer analyses are very useful and will become increasingly so. However, the failure of a specific search technique to find any database homology with a new gene does not mean that no such homology exists. There is no one correct computational way to view sequences and we should be diligent in keeping our focus on the primary problem, that of making available the most basic scientific data from molecular biology.

The approaching era of genomic sequencing is being discussed widely. An eight-enzyme restriction map of *Escherichia coli* has been determined and several efforts are underway to sequence its 4.7×10^6 nucleotide genome. Several other genomes, including those of yeast, *Caenorhabditis elegans*, *Drosophila*, and mouse, are being characterized. The human genome of 3×10^9 bp is being approached via genetic and physical mapping. We have found it difficult to cope with today's volume of data, but the problems of today will look elementary in just a few years. This is a strong argument for quickly solving today's problems and preparing for the future. Genomic sequencing projects will require new databases. They will also require new standards of information exchange. What I think should happen in a well-designed world is that the sequencing efforts manage their own raw data, passing sequence to central DNA databases when a given length and quality have been attained. It should not be acceptable for a publicly funded genomic sequencing project to restrict access to sequences for an extended time. Funding agencies must formulate specific policy covering such issues. Of course, sequencing centers might well maintain and study current copies of the DNA databases.

Sequences have usually been thought of as unique entities, such as "the sequence for *E. coli* lys-tRNA." With genomic sequencing this will change for two reasons. One reason is the polymorphism that is widespread in the genome. When

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-