

MOLECULAR BIOLOGICAL DATABASES: THE CHALLENGE OF THE GENOME ERA

RAINER FUCHS and GRAHAM N. CAMERON

The EMBL Data Library, European Molecular Biology Laboratory, Postfach 10.2209, D-6900 Heidelberg, Germany

CONTENTS

I. INTRODUCTION	215
II. SEQUENCE DATABANKS AND GENOME PROJECTS	216
1. <i>Sequence Databanks—A Historical Overview</i>	216
2. <i>Today's Sequence Databanks</i>	217
3. <i>Genome Analysis Projects</i>	220
III. GENOME ANALYSIS AND LARGE-SCALE SEQUENCING PROJECTS—IMPLICATIONS FOR THE NUCLEOTIDE SEQUENCE DATABASES	222
1. <i>Increasing Data Rate</i>	222
2. <i>Data Publication and Data Acquisition</i>	223
3. <i>A New Quality of Data: Less Annotation and Continuous Updating</i>	226
4. <i>Changing Requirements for Data Access and Data Distribution</i>	227
5. <i>Integration of Databases—Linking Related Data Sets</i>	229
6. <i>Customization of the Databases—Different Views of the Data Set</i>	230
IV. COPING WITH THE NEW REQUIREMENTS—STRATEGIES FOR THE SEQUENCE DATABANKS	232
1. <i>Data Acquisition</i>	232
2. <i>Data Distribution</i>	234
3. <i>Data Handling and Storage</i>	237
4. <i>Annotation</i>	238
V. A MODEL FOR THE NEXT GENERATION OF SEQUENCE DATABASES	239
1. <i>Principles of the Model</i>	239
2. <i>Some Details of the Model</i>	241
ACKNOWLEDGEMENTS	243
REFERENCES	243

I. INTRODUCTION

Soon after publication of the first few sequences of biological macromolecules, scientists began to organize this information into databases (Dayhoff, 1966). Since then the sequence databases have evolved from mere by-products of research projects into a major international and collaborative investment which aims to collect and redistribute all available sequence data. Today, sequence databases have become invaluable and indispensable research tools in many domains of modern molecular biology.

In the last few years cloning and DNA sequencing technology has improved considerably, with new techniques such as pulsed-field gel electrophoresis (Smith *et al.*, 1986a), yeast artificial chromosome vectors (Burke *et al.*, 1987), multiplex sequencing (Church and Kieffer-Higgins, 1988) and the introduction of devices for the automatic extraction and sequencing of DNA (Connell *et al.*, 1987; Edwards *et al.*, 1990; Knobeloch *et al.*, 1987) having a major impact. Probably the most important innovation in molecular biology in the last decade has been the revolutionizing introduction of the polymerase chain reaction (PCR) (Innis *et al.*, 1988; Saiki, 1985), by which many of the traditional methods of cloning and sequencing can be complemented or even circumvented (White *et al.*, 1989). All these achievements have greatly affected the rate and the costs at which new sequence data can now be obtained.

As a result of this progress the elucidation of the complete genomic information of the cell

now seems feasible. In fact, the era of genome sequencing has already begun: projects to determine the complete nucleotide sequences of several prokaryotic and eukaryotic genomes are well under way. These initiatives will have profound effects on the operation of the sequence databanks (Waterman, 1990). The expected amount of data arising from genome analysis projects will make the databanks' recent problems look rather trivial.

In this article we discuss the implications which the advances in sequencing technology and the genome analysis projects will have for the existing sequence databanks and how they can react to the challenges of the future. The focus is on nucleotide sequence databases, and the database maintained at the European Molecular Biology Laboratory (EMBL), Heidelberg, is frequently used as an example. However, many of the issues discussed here are of equal importance for protein sequence and other kinds of molecular biological databases. The first section provides some basic information on sequence databases and genome projects in order to improve the understanding of the problems which the databanks will have to face in the coming years. Then, the consequences of large-scale sequencing and genome analysis projects are explained in detail and it is shown that they require fundamental changes to the work of the sequence databanks. Next, different approaches and strategies for coping with the forthcoming problems are outlined, and finally we present a model for a next generation of sequence and other biological databases which requires a conceptional reorganization of these databases, but which offers good chances for successfully mastering the challenges of the future.

II. SEQUENCE DATABANKS AND GENOME PROJECTS

1. *Sequence Databanks—A Historical Overview*

In order to understand how today's nucleotide and protein sequence databanks operate it is helpful to have a brief look back on the history of these databanks and to see how they evolved in the context of the developments in sequencing technology. The history of the nucleotide sequence databases was recently reviewed in more detail by Smith (1990).

The first report of the complete sequence of a biological macromolecule goes back to 1956 when Sanger described the amino acid sequence of bovine insulin, consisting of 51 residues (Sanger, 1956). Almost 10 years later, in 1965, Holley *et al.* published the first nucleic acid sequence, the sequence of yeast alanine tRNA with 77 bases (Holley *et al.*, 1965). Initially, the number of sequences published was very low, however the value of sequence information was realized very early. Primarily intended as a tool for her own research interests, Dayhoff (1966) assembled and published the first major collection of protein sequences about 25 years ago. In 1967, the introduction of the automated protein sequenator (Edman and Begg, 1967) greatly facilitated the determination of protein sequences, but it took another 10 years before the advent of recombinant cloning techniques (Maniatis *et al.*, 1982) and the development of methods for the direct sequencing of DNA (Maxam and Gilbert, 1977; Sanger *et al.*, 1977) brought sequencing into any biochemical laboratory. The simplicity of these new techniques quickly resulted in a dramatic increase in the number of reported nucleic acid sequences in the following years.

In 1980, the EMBL Data Library was established (Hamm and Cameron, 1986) with the explicit goal to collect, organize and distribute a database of all nucleotide sequences and related descriptive information extracted from publications in scientific journals. Since 1982, this work has been done in international collaboration with the American GenBank group (Bilofsky *et al.*, 1986), and more recently the DNA Databank of Japan (DDBJ) joined the collaboration. Data collection is now being shared between these databanks, and newly created database entries are exchanged on a daily basis.

Dayhoff's protein sequence database has evolved into a similar international tripartite cooperation of databanks called PIR-International (Barker *et al.*, 1990), which was established in 1987 and now consists of the Protein Identification Resource (PIR) in the U.S.A., the Martinsried Institute for Protein Sequences (MIPS) in Germany, and the Japanese International Protein Sequence Database (JIPID). Another important protein database these days is Swiss-Prot (Bairoch and Boeckmann, 1991), a collaboration between EMBL and A. Bairoch, Geneva. The simplicity of DNA cloning and sequencing compared to

protein isolation and sequence determination has resulted in more nucleotide than protein sequences being published. Most protein sequences now included in the protein databases are inferred from protein-coding nucleotide sequences. There is a close and effective collaboration between the nucleotide and protein databanks; EMBL and GenBank promptly transmit new protein-coding sequences to PIR, and preliminary Swiss-Prot entries are created daily by automatic translations of new EMBL entries.

DDBJ/EMBL/GenBank on the one hand and PIR and Swiss-Prot on the other hand are fundamental databanks which centrally collect all available nucleotide and protein sequence information, thus making a comprehensive compendium of sequence data available for general usage. In addition to these main databases a variety of smaller and specialized data collections have been established through the years, many of them only short-lived, but some of them of great importance to specific groups of scientists. These specialized databases concentrate on certain molecule types, e.g. the tRNA database (Erdmann and Wolters, 1987), on properties of the sequences such as the Transcription Factor Database (Ghosh, 1990) or Prosite (Bairoch, 1991), or they try to collect data from certain species, e.g. *E. coli* (Kröger *et al.*, 1990; Rudd *et al.*, 1990), thus integrating pure sequence data with other information.

In an attempt to focus the American efforts in the field of biocomputing and biological databases the new National Center for Biotechnology Information (NCBI) has been established recently as part of the U.S. National Library of Medicine (NLM), excellently funded with a budget of \$10 million per year (Benson *et al.*, 1990). One of the declared goals of NCBI is the creation of a new sequence database, the GenInfo Backbone Database (NCBI, 1990) which will concentrate on collecting all available nucleotide and protein sequence information from the scientific literature. NCBI will also take over responsibility from the National Institute of General Medical Sciences (NIGMS) for the GenBank database when the present GenBank contract runs out in 1992. Although the next few years will probably bring some major reorganization in the area of sequence databases in the United States, at the moment, it can only be speculated on where these new developments will lead and how they will affect the international collaboration of the major sequence databanks.

2. Today's Sequence Databanks

In order to fully appreciate the effects which genome projects and large-scale sequencing efforts will have on the sequence databases it is obviously important to understand how the current sequence databanks operate. Their work is exemplified here by the EMBL nucleotide sequence database.

The database work can be divided in three different aspects: data collection, data handling, and data distribution. Figure 1 shows a schematic flow of nucleotide sequence data from the scientists to the databanks and back to the scientific community.

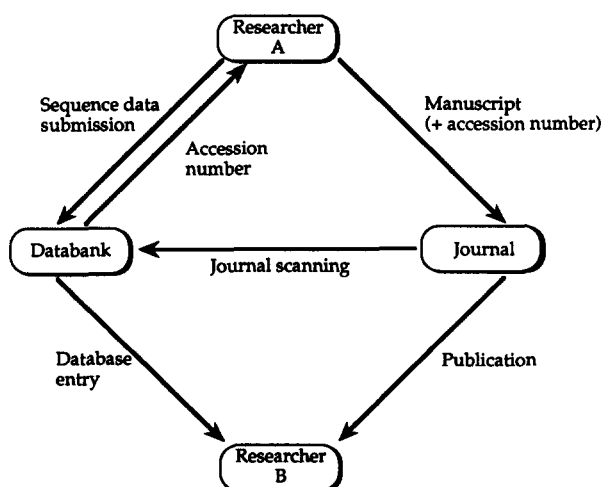


FIG. 1. Data flow between the nucleotide sequence databases and the scientific community.

Until 1988, more than 70% of all sequence information stored in the EMBL database came from scientific publications. The process of scanning the literature and typing in sequences is obviously a time-consuming task, resulting in a severe lag period of several months or more between the publication of a sequence and its availability in the database. Printed sequences are not amenable to computer analysis, the only realistic way of further interpretation. An important principle of science is that experiments should be repeatable and that results can be verified independently by other researchers. Sequence data is clearly the outcome of scientific experiments, and as such they must, therefore, be accessible to checking and verification. However, for the normal biologist typing in a sequence manually from a publication in order to work with it is error-prone and almost impossible for sequences longer than a few thousand base pairs, thus making the availability of a computer-readable copy necessary. The growing lag between a sequence publication and the time it was incorporated in the databases led the databanks to develop new strategies for data acquisition, and in 1988 a direct submission scheme was introduced as a collaboration between the nucleotide sequence databanks and some important journals, which strongly encouraged or insisted on the submission of new sequences to the databases prior to publication.

There has been some controversy in the past about the possible drawbacks of a mandatory direct sequence submission (Cameron *et al.*, 1989; Maddox, 1989a,b; Roberts, 1989), but nowadays almost all leading journals accept sequence-containing manuscripts only when the sequence information has previously been deposited in the public databases. The databanks return an accession number to the submitter after they have received a sequence submission. The accession number is a unique, unchanging identifier for the sequence and proof of the deposition of the sequence in the database. As a result of the direct submission system, some 80% of all sequences entering the nucleotide databases now come directly from the authors instead of being picked up by journal scanning. Consequently, it was possible to reduce the average turn-around time for a new database entry to a few weeks, and the databanks try hard to make new sequences available at the same time they appear in print.

Although direct data submission to the databanks is now common practice there is still a considerable number of sequences published which have not been communicated to the databanks beforehand. The sequence databanks, therefore, have to scan the relevant literature regularly for sequence-containing articles. If a sequence publication is detected, the database is checked for whether the sequence has been submitted previously by the authors, and the sequence is entered into the database otherwise.

The third route of data acquisition is not indicated in Fig. 1. Nucleotide and protein sequence data are collected in international collaborations of independent groups. EMBL, GenBank and DDBJ have divided up the tasks of scanning the literature for nucleotide sequences and handling direct submissions, and in order to guarantee a unified database newly created entries are exchanged on a daily basis. A similar mode of sharing the workload has been adopted by the members of the PIR-International collaboration. An important difference between the nucleotide and protein sequence databanks is the fact that the PIR-International group distributes only one version of their database whereas EMBL, GenBank and DDBJ all produce independent releases of the common data set, in different formats. Striving for unification, the nucleotide sequence databanks are now in the process of establishing a new format-independent data exchange protocol which should eliminate the existing differences.

Data acquisition by the protein databanks is somewhat different, since most sequences are deduced from nucleotide sequences. The main route for their data is thus the forwarding of protein-coding DNA sequences from the nucleotide sequence databanks.

Every new direct sequence submission and every sequence picked up from the literature is turned into a new database entry after acquisition. Figure 2 shows an example of a typical EMBL nucleotide sequence database entry. Although the formats of the EMBL, GenBank and PIR databases are different, they all have in common that an entry consists of different line types, identified by some code (such as ID, AC, etc.) and presenting some well-defined sort of information. A complete entry contains not only the sequence data but a lot of other relevant information attached to it, called annotation. Articles and direct data submissions

are scrutinized by a team of graduate biologists who extract all the important pieces of information. Annotation includes source information, reference information, keywords and pointers to other databases, but most importantly information about the biological function and properties of a sequence in the form of the feature table. The feature table shown in Fig. 2—identified by the FT line type code—tells the reader, for instance, that the sequence in that entry represents an incomplete mRNA which codes for a part of the Xl-pou protein. It specifies the protein-coding region and contains some reading-frame information so that it is possible to automatically translate this nucleotide sequence into the corresponding protein sequence.

```

ID   XLNRL20      standard; RNA; VRT; 317 BP.
XX
AC   X54681;
XX
DT   05-APR-1991 (Rel. 28, Last updated, Version 5)
DT   31-OCT-1990 (Rel. 25, Created)
XX
DE   Xenopus laevis mRNA for nrl-20 POU-homeobox protein
XX
KW   homeo box; transcription factor.
XX
OS   Xenopus laevis (clawed frog)
OC   Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Amphibia;
OC   Lissamphibia; Anura; Archeobatrachia; Pipidae; Pipidae.
XX
RN   [1]
RP   1-317
RA   Stiegler P.;
RT   ;
RL   Submitted (06-SEP-1990) on tape to the EMBL Data Library by:
RL   Stiegler P., Institut de Biologie Moleculaire et Cellulaire du
RL   CNRS, 15 rue Rene Descartes, 67084 Strasbourg Cedex, France.
XX
RN   [2]
RP   1-317
RA   Baltzinger M., Stiegler P., Remy P.;
RT   "Cloning and sequencing of POU-boxes expressed in Xenopus laevis
RT   neurula embryos";
RL   Nucleic Acids Res. 18:6131-6131(1990).
XX
DR   SWISS-PROT; P20914; HM20$XENLA.
XX
CC   *source: developmental stage=neurula;
CC   See X54677 - <X54685 for analysed POU-box mRNAs.
XX
FH   Key          Location/Qualifiers
FH
FT   CDS          <1..>317
FT                /product="Xl-pou protein" /codon_start=2
XX
SQ   Sequence 317 BP; 84 A; 95 C; 86 G; 52 T; 0 other;
      tcaggcagat gtgggcctgg ccctgggcac cctctatggc aatgtcttct cccagaccac
      catctgcagg ttcgaggcgc tccagctcag ctttaagaac atgtgcaagc tcaagcctct
      gctcaacaag tggctggagg aggccgactc ctccactggc agccccacca gcatcgacaa
      aatcgcagcg cagggcagga agagaaagaa gaggacttca atagaggtag gcgtaaaaag
      ggcatggag agccacttcc tcaagtgcc taaaccagcg gctcaggaaa tcaccacact
      ggcggacagc ctccaac
  
```

FIG. 2. A sample entry of the EMBL nucleotide sequence database.

Data storage and management is handled very differently by the major databanks. At EMBL and GenBank, entries are actually not stored in the form shown in Fig. 2 but as data in a commercial relational database management system (RDBMS). Only for distribution purposes are entries such as the one shown built by extracting the necessary information from the RDBMS. The protein databanks, in contrast, work directly with files similar to that shown in Fig. 2 and use self-written software for data management.

The collected sequence information plus attached annotation is made available to the scientific community as regular releases of the database. In most cases a new release is distributed every 3 months. Releases of the major nucleotide and protein sequence databases

are only supplied on electronic media and simply consist of one or more flat files containing all entries appended and sorted by some criteria such as taxonomy. The traditional distribution medium has been magnetic tape, but the databanks have recently started to use alternative means of data distribution such as CD-ROM and computer networks (see below).

3. *Genome Analysis Projects*

The first plans for sequencing complete genomes were already formulated in the mid-80s (Bitensky, 1986). For obvious reasons the human genome has attracted most interest as the main target for genome research since then. Although there has been some vehement controversy about the reasonableness and usefulness of this project, there is now a strong world-wide initiative to elucidate the structure of the human genome by determining the complete nucleotide sequence of all its chromosomes. The strongest player, by financial resources, in this game is certainly the United States. After extensive discussions (U.S. Congress, Office of Technology Assessment, 1988) the United States Government eventually launched the U.S. Human Genome Project in 1988, mainly funded by the U.S. Department of Energy (DoE) (Barnhart, 1989) and the U.S. Department of Health and Human Services (Watson and Jordan, 1989).

The scientific plan for the first 5 years of this project was formulated recently (U.S. Department of Health and Human Services and U.S. Department of Energy, 1990), and it is of interest to have a closer look at the goals formulated by this plan since it is representative for most other genome initiatives which are being considered or planned. It is estimated that \$200 million per year are necessary to ensure the success of this effort within the next 15 years. Today's DNA sequencing technology is not seen to be appropriate for the task of sequencing the 3×10^9 million base pairs of human DNA, thus systematic sequencing of large stretches of DNA is deferred to a later stage when technology is improved. The current cost of sequencing a base of DNA is between \$2 and \$5, but has to drop to less than \$0.50 per base before large-scale sequencing will be initiated. The emphasis in the first few years, besides improvements of technology, will be on the construction of complete detailed genetic and physical maps of the human genome, and the construction of ordered clone libraries. Sequencing of larger regions of the genome will only be performed in the course of technology improvement.

Deciphering the nucleotide sequence of the genome is obviously not the final goal of genome researchers; eventually they want to *understand* the genome. That implies that an indispensable part of any genome project is the thorough analysis and interpretation of the collected data. In fact, considerable resources have been allocated in the U.S. Human Genome Project for the improvement of data management and analysis, aiming at \$30 million per year. To supervise and coordinate the efforts in this area, a Joint Informatics Task Force has been established, made up of experts chosen by the two main funding agencies DoE and NIH (National Institutes of Health).

In addition to the United States, several other countries have already joined the global human genome initiative, including several European countries and the European Community. Interestingly, in contrast to the American approach most European human genome research programs do not intend to sequence genomic DNA, but concentrate on the analysis of cDNA libraries instead (Alwin, 1990; Jordan, 1991). In late 1990, the Commission of the European Communities (CEC) launched the European Human Genome Analysis Program with a budget of 15 million ECU for the time period 1990 to 1992. As in the U.S. Human Genome Project, the program does not promote large-scale sequencing, but concentrates on genetic and physical mapping and the development of technology instead, including new and improved methods of data handling. Fifteen percent of the total budget, i.e. 2.2 million ECU has been allocated to database activities and the production and improvement of software and algorithms.

Although the public interest clearly focuses on the exploration of the human genome, several projects to examine the genomes of other organisms are being planned or are even well under way. Besides the genuine interest in the biology of these organisms, the analysis of their genomes may also serve as a model for the analysis of the human genome. The first

5-year-plan of the U.S. Human Genome Project indeed foresees the sequencing of parts or even complete genomes of model organisms as an intermediate step towards the sequencing of the human genome. Table 1 summarizes the European efforts supported by the European Community.

TABLE 1. GENOME RESEARCH PROJECTS SUPPORTED BY THE EUROPEAN COMMUNITY (GOFFEAU AND VAN HOECK, 1990, MODIFIED)

Title	Programme	No. of labs.	Period of execution	EC contribution (in ECU)
Sequencing of the chromosome III from yeast	BAP	35	89-90	2,635,000
Sequencing of the yeast genome	BRIDGE	31	91-93	5,060,000
Molecular identification of new plant genes (focused on the <i>Arabidopsis</i> genome)	BRIDGE	27	91-92	3,000,000
Establishment of a complete physical map and strategic approach to the sequencing of the <i>Bacillus subtilis</i> genome	SCIENCE	5	89-91	609,000
A complete physical map of the <i>Drosophila melanogaster</i> genome	SCIENCE	3	88-93	718,000
Functional and structural analysis of the mouse genome	SCIENCE	3	89-92	996,000
Development of a genetic and physical map of the porcine genome	BRIDGE	11	91-93	1,200,000
Eukaryote genome organization: repeated DNA elements and evolution in the genome of <i>Caenorhabditis</i>	SCIENCE	?	91-93	Under negotiation
Physical map of the human genome	HGAP	?	91-92	15,000,000

BAP=Biotechnology Action Programme; BRIDGE=Biotechnology Research for Innovation, Development and Growth in Europe RTD Programme, HGAP=Human Genome Analysis Programme.

Most progress so far has been achieved in the yeast genome project (Mewes and Sgouros, personal communication). The strong industrial interest in *Saccharomyces cerevisiae* and the vast amount of knowledge already collected about this microorganism made this species a prime candidate for the analysis of its genome. Good maps and an ordered clone library have been available, and in early 1989 systematic sequencing of the yeast genome began under the Biotechnology Action Programme of the European Community. Thirty-five laboratories in 10 European countries are sequencing the complete chromosome III of about 370,000 base pairs, and work on three other chromosomes of yeast is already scheduled. The total sequence of chromosome III should have been determined by early 1991, and the data will be made available during 1991 by deposition in the EMBL sequence database.

Sequencing has also been initiated in the nematode sequencing project which aims at determining the complete genomic sequence of the worm *Caenorhabditis elegans* (Coulson *et al.*, 1986). This invertebrate is of particular interest because it consists of only a few thousand cells, and the fate of each of these cells during differentiation is well-known, thus making *C. elegans* an ideal object for studies on gene regulation and development.

Most of the other genome projects are much less advanced, with research mostly focusing initially on the construction of genetic and physical maps. Concerted efforts to systematically sequence genomes of higher eukaryotes are not expected to be initiated within the next 5 years, but the targeted sequence determination of selected genome regions will certainly start sooner.

It should be stressed that neither the European nor the American efforts in genome analysis are isolated. In fact, most genome initiatives are, like the human genome project, international activities with collaborators from all over the world. European projects such as the nematode, the *Drosophila* or the *Arabidopsis* project have similar counterparts in the United States, and scientists from different continents collaborate closely. Co-ordination is essential, and in 1988 the Human Genome Organisation (HUGO) was founded to

coordinate international genome research not only in the human but also other genome projects (McKusick, 1989).

One initiative which is of a particular interest because it is fundamentally different from the previously mentioned genome projects is the analysis of the *Escherichia coli* genome. *E. coli* is probably the best-characterized organism on this planet; the accumulated knowledge about the biochemistry and genetics of this bacterium is overwhelming. It is one of the few organisms of which detailed genetic and physical maps have already been established (Bachmann, 1990; Kohara *et al.*, 1987; Smith *et al.*, 1987). Although several groups have announced the systematic sequencing of the *E. coli* genome (Anderson, 1989; Church and Kieffer-Higgins, 1988; Daniels and Blattner, 1987), no results of these initiatives have yet been published, and essentially all sequences currently available in the public databases were collected in an uncoordinated effort. Despite this, more than 30% of its genome (Kröger *et al.*, 1990) has been sequenced by now, and by looking at the rate at which new *E. coli* sequences are deposited in the databases it can be expected that the complete sequence will be available in a few years time.

III. GENOME ANALYSIS AND LARGE-SCALE SEQUENCING PROJECTS— IMPLICATIONS FOR THE NUCLEOTIDE SEQUENCE DATABASES

The rapid improvement of sequencing techniques and their application in large-scale sequencing and genome research projects will severely affect the current operation of the nucleotide sequence databanks. Their task will change dramatically with concerted, international attempts to sequence genomes of hitherto unapproachable magnitude and increasing automation of the sequencing process.

The most obvious effects on the databases will result from the sheer amount of data arising from these projects. But apart from the increase of the workload on the databanks, the genome projects will also drastically accelerate other developments which are already beginning to become visible. These challenging developments will create a new set of requirements for the sequence databases.

1. Increasing Data Rate

Figure 3 shows the growth of the EMBL nucleotide sequence database and the Swiss-Prot protein sequence database for the last few years. The first release of the EMBL database was published in 1982 containing 568 entries and about 600,000 bases, whereas Rel. 26 (February 1991) now contains 43,745 sequence entries comprising more than 55 million nucleotides. The graph shows a stable growth of the database with a doubling time of less than 2 years. Extrapolating this curve, assuming no dramatic changes to the current data rates, lets us assume that the nucleotide sequence database will be more than one hundred times larger than at present within the next 10 years. But two factors will certainly increase the pace of the database growth. Firstly, the general progress in sequencing technology will result in the determination of more bases in less time. This development will affect the database in its entirety and will clearly accelerate the overall growth rate. Secondly, the database is biased; certain species are overrepresented due to an increased research interest in these organisms. The EMBL database currently contains data from about 3000 different organisms. Table 2 shows that almost half of the database consists of data from just 10 different species, most notably sequences of human origin. Not surprisingly, most of the genome analysis projects being planned concentrate on the organisms represented in this list. As a result, the data for these organisms will accumulate even faster than they would anyway due to constant improvement of technology, which will eventually accelerate the growth of the databases even more.

The largest complete genome in the database so far is that of the human cytomegalovirus with about 200,000 bases (Chee *et al.*, 1990), only less than one ten-thousandth the size of the human genome. Table 3 compares the genome sizes of some of the organisms whose genomes will probably be determined in the near future. Note that the figures are not cleaned for overlapping or identical sequences, such as sequences determined from both cDNA and genomic DNA. A recent compilation of cloned *E. coli* sequences (Kröger *et al.*, 1990) showed

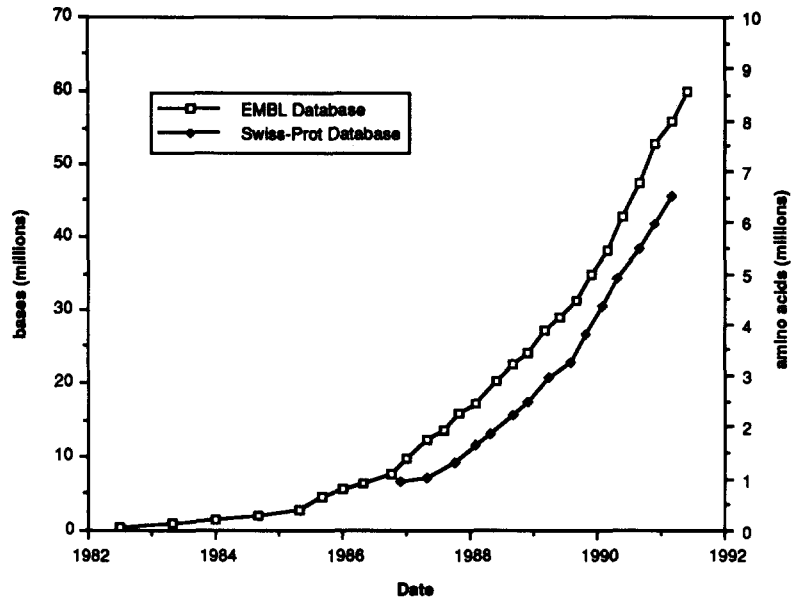


FIG. 3. Growth of the EMBL nucleotide sequence database and the Swiss-Prot protein sequence database.

TABLE 2. THE TEN MOST HEAVILY REPRESENTED SPECIES IN THE EMBL NUCLEOTIDE SEQUENCE DATABASE. (TOTAL SIZE OF THE DATABASE AS OF MARCH 1991: 66×10^6 BP AND 51,974 ENTRIES; FIGURES NOT CLEANED FOR OVERLAPS)

Species	Bases in database	% of database
<i>Homo sapiens</i> (Man)	1.1×10^7	17
<i>Mus musculus</i> (Mouse)	5.2×10^6	8
<i>Rattus norvegicus</i> (Rat)	3.3×10^6	5
<i>Escherichia coli</i>	2.5×10^6	4
<i>Saccharomyces cerevisiae</i> (Yeast)	2.4×10^6	4
<i>Drosophila melanogaster</i> (Fruit fly)	1.9×10^6	3
<i>Gallus gallus</i> (Chicken)	1.2×10^6	2
<i>Bos taurus</i> (Cattle)	9.3×10^5	1.5
<i>Xenopus laevis</i> (Clawed frog)	7×10^5	1
<i>Oryctolagus cuniculus</i> (Rabbit)	6.6×10^5	1

TABLE 3. GENOME SIZES AND SEQUENCE COVERAGE BY THE EMBL SEQUENCE DATABASE OF SPECIES WITH INTEREST TO GENOME RESEARCHERS. (DATABASE INFORMATION SEE TABLE 2)

Species	Approx. genome size ($\times 10^6$ bp)	% of genome in database
<i>Homo sapiens</i>	3000	0.4
<i>Mus musculus</i>	3000	0.2
<i>Drosophila melanogaster</i>	165	1.2
<i>Arabidopsis thaliana</i>	100	0.3
<i>Caenorhabditis elegans</i>	80	0.4
<i>Saccharomyces cerevisiae</i>	15	16
<i>Escherichia coli</i>	4.5	55

that almost 30% of the *E. coli* data in the database is not unique, due to overlapping sequences. Although the redundancy will certainly be much lower for most other species listed in Table 3, it shows that simply calculating the number of base pairs in the database for a given species overestimates the amount of information already known. Assuming that only between 0.3% and 0.4% of the human genome is now in the database, and assuming that the international human genome initiative will achieve its goal of sequencing the whole genome within the next 15 years, then this alone will result in a sequence database 300 times larger than the current one.

2. Data Publication and Data Acquisition

Neglecting the fact that "the" nucleotide sequence database is actually the mutual collaboration of DDBJ, EMBL and GenBank which exchange collected information, there are currently two main routes of data acquisition: direct submissions from the scientific community, and the scientific literature. In the past the primary way of reporting the results of sequencing experiments has been to publish them in a scientific journal. It seems likely that the future will show much fewer publications of sequence data in this traditional form; an increasing amount of sequence information will instead be published by depositing the data directly, and exclusively, in the public databases, reserving journal publications predominantly for scientific discussion and conclusion.

The successful implementation of the direct submission scheme which was outlined above has resulted in the availability of new sequences soon after or simultaneously with the appearance of the corresponding journal article. The growing number of publications per year makes it necessary for journals to save precious space, and an obvious target for savings are the printed sequences. The reader can easily obtain this information in more convenient form from the databanks. The past also showed an increase in the number of manuscripts which mainly consisted of sequence data and which contained little or only marginal additional biological information. In order to save space and to further improve the quality of publications some journals are now going to restrict the publication of pure "sequence papers" with questionable specific relevance, and, in fact, one can observe a growing reluctance to publish sequence data at all. Instead, researchers are encouraged to publish their sequences by submitting them to the databanks and to refer to them in their papers by citing the database entries (Walker, 1990). This tendency will also affect the dissemination of the results of genome sequencing, because sequence data from these projects will initially have little additional biological information attached to them. It seems unlikely and undesirable that sequence data will be published in the traditional manner in scientific journals. Direct deposition of sequences from genome projects into the public databases will become the main route for publication instead. Nevertheless, scientists need to get recognition for their work and, therefore, it is very important that standards be developed between publishers and the databanks which allow one to cite database entries in the same way as any journal publication.

In the past, direct submissions of nucleotide sequence data have almost always been directly received from the scientists who did the sequencing work. Large-scale sequencing and genome analysis projects will instead establish project-specific informatics resource centres which gather the primary sequence data from the collaborating laboratories and forward the collected information to the central public databanks after some period of data checking and evaluation. The interaction between the European yeast chromosome III project and the EMBL Data Library is representative of this new route of sequence data acquisition. Sequence information from all labs participating in this project is collected centrally by the Martinsried Institute for Protein Sequences (MIPS) in Germany, whose scientists check the data, assemble contigs, analyze the sequences and finally forward the information to the EMBL Data Library for public distribution.

Although project-specific databases will play an important role for rapid and convenient dissemination of new data to all collaborators of a project, the interests of researchers not directly involved in these efforts will still be served best by the public databases. Thus, the sequence data collected in a specialized genome project database will eventually have to be

included in the general sequence databases in order to allow other scientists to access and work with this information. The importance of appropriate data exchange mechanisms between the project and the public databanks which guarantee the rapid integration of sequencing project data into the public databases can hardly be overestimated. Data will be electronically exchanged between these groups using global computer networks, and thus good links between the computer systems of the central public databanks and the project databanks have to be established. In view of the great amount of data to be expected from these project databanks it is necessary to develop procedures which allow the databanks to automate the process of data submission and data integration as much as possible. It is clear that these problems must be tackled before the project databases have accumulated large amounts of sequence information, and thus close collaborations between the public databanks and project informatics resource centres have to be initiated at the earliest possible stage.

Up to now, the nucleotide and protein sequence data collection has been a successful and close collaboration between groups from different countries and continents. A particular aspect of the genome projects may threaten this collaboration and the free availability of sequence information in general: sequence data has potential commercial value. Pharmaceutical, biotechnological or agricultural applications of genome sequence information have been evident for some time. In fact, it has been suggested already to restrict the dissemination of genome data to certain groups or countries in order to reserve any commercial benefit (Marshall, 1990). These proposals illustrate a potential problem which the databanks might have to face in the future. Such restrictions would be detrimental to the work of the public databanks which currently strongly benefit from the mutual, free exchange of sequence data. It is necessary, as formulated by the U.S. Human Genome Project Joint Informatics Task Force, "... to make the information and analysis tools from this project freely available to the widest possible range of scientists and physicians in the most useful, timely and cost-effective fashion" (U.S. Department of Health and Human Services and U.S. Department of Energy, 1990.)

The previously discussed topics were all related to the routes sequence information has to take to enter the databases. But other, hitherto unknown, problems of data collection will be independent of the actual route, but will arise from the fact that the switch from highly targeted sequencing experiments to systematic sequencing efforts will result in a continuous flow of sequence data and continuous updating, thus severely undermining the current concept of a database "entry".

At present, we can regard an entry in the database as being static, despite the fact that a few percent of all entries in every release are updated in some way. These updates mainly affect the correction of spelling errors or factual errors in annotation, although in some cases authors supply us with corrected or additional sequence information. But the majority of entries, and in particular the sequence data in the entries, are stable. Additionally, most of the sequences in the database do not overlap. Therefore, one entry represents one defined and independent DNA sequence. This picture will change with data from genome analysis projects. Sequences submitted to the databanks will no longer be independent but will share long regions of identity. While sequencing large regions of DNA by analyzing individual clones of a library it will be inevitable that, firstly, these clones will overlap to a certain degree, and, secondly, there will be sequence gaps in the published data which can only be filled after some time, unless the data is withheld until all gaps are closed.

The problem is illustrated in Fig. 4, assuming that sequence data is forwarded from some genome project to the public databank as sequences of individual clones. In the beginning these sequences will probably be independent and the databanks can simply create one entry for every new submission (clone 1 and clone 2). However, at some stage newly submitted sequences (clone 3) will overlap with others already in the database leading to the construction of sequence contigs (Staden, 1980). Figure 4 depicts three possible strategies for the databanks:

- (A) The databanks will simply continue to create new entries for every incoming clone sequence. As a consequence, large redundancy will be introduced into the database, due

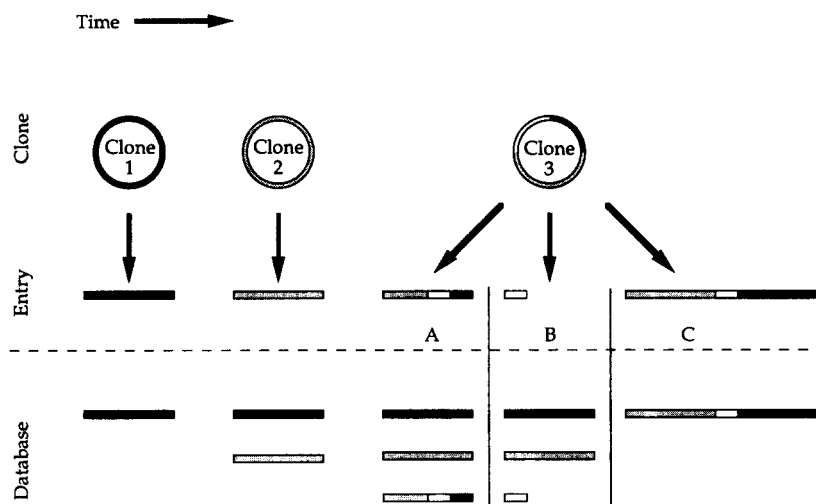


FIG. 4. Submission of overlapping sequence data from genome projects to the databanks. The top row represents the individual clone sequences submitted to the databanks. The second row are the entries produced from these sequences. The linear sequences are not drawn to scale with the circular clone sequences. The bottom row indicates the status of the database after the addition of the newly submitted sequence. The dithering reflects overlapping sequences.

to the duplication of overlapping sequences and its extent will eventually depend on the composition of the gene library. Additionally, the fact that entries overlap is not immediately visible.

- (B) An entry is built from that part of the new submission which is not already in the database. Redundancy is avoided, but the scientific report—the submission—is not properly reflected. Again, overlaps are not explicitly represented.
- (C) If overlaps are detected, a new entry is constructed whose sequence is the contig formed by merging the individual sequences. The redundancy problem is avoided again, and the sequence overlap is immediately recognized by the database user. Nevertheless, the individual reports disappear from the database, and the independently reported sequences cannot easily be reconstructed.

Clearly, none of these possibilities is perfect. The problem obviously exists already but it emerges only sporadically, so databanks can afford to handle these few cases pragmatically. The present nucleotide sequence databanks do not have a clear-cut policy for merging entries if they detect overlaps. Sometimes model A is adopted, i.e. the databanks keep overlapping sequences as individual entries, but they note the overlap somewhere in the annotation; in other cases they join overlapping entries according to model C. Nonetheless, it is evident that in the future this situation will occur more often, and a clear strategy for handling these cases is required. As discussed below, finding a satisfying solution is not trivial, because disagreements in the region of overlap will be found quite frequently which have to be represented in an appropriate manner.

3. A New Quality of Data: Less Annotation and Continuous Updating

It is well-established that the genomes of higher organisms such as man or mouse mainly consist of non-protein coding DNA regions with hitherto no obvious function. The estimated 100,000 human genes which code for proteins or RNA probably make up for only 5 to 10% of the human genome. The rest of the genome is often called “junk DNA”, however it is highly unlikely that entirely superfluous material has persisted for so long during evolution. It is one of the great challenges of the genome projects to elucidate the biological role of these parts of the genome.

At present most research projects which involve cloning and sequencing of genes are highly targeted at solving particular biological problems. Most often cloning and sequencing is initiated to find a specific piece of DNA with a known or putative function. The analysis of

the new sequence is therefore guided and facilitated by the additional information available on the biological importance and role of this gene. This fact is normally reflected by the quality and amount of information supplied if a sequence is submitted to the databanks or if the sequencing and cloning results are reported in a journal publication. It enables the databanks to attach a great richness of biological information (annotation) to almost every entry in the database.

In contrast, systematic sequencing of a whole genome will inevitably yield lots of sequence data for which the only information initially available will be the source information, a clone number and a map position. In some cases it will be possible to deduce putative functions for a newly determined sequence by applying computer algorithms for the prediction of coding regions, regulatory elements, etc., or by sequence comparison. In fact, the importance of methods for deducing potential functions of unknown DNA has been realized clearly, and resources are allocated for research in this area (U.S. Department of Health and Human Services and U.S. Department of Energy, 1990).

On the other hand, current progress in sequencing technology makes it likely that sequence information will be obtained much faster than it can be analyzed. This generates a specific problem for the genome projects in regard to data submission to the databanks. Presently, a researcher submitting sequences to the databanks prior to publication can request that his sequences be treated as confidential, and the databanks will make his data publicly available only when they appear in print. With the shift towards "electronic publishing" by direct and exclusive deposition of sequence data in the databases this option will no longer make sense. There is an obvious contradiction between the public interest in free and immediate access to new data and the interest of the individual scientist to withhold his or her data for some period of time in order to analyze and interpret them, and possibly to prepare a publication. Indeed, there has been intensive discussion about the acceptable delay for the publication of new sequences. It now seems that a period between 6 months and 1 year is considered to be adequate for an initial analysis without reducing the currency of sequence data too much. Longer periods are unacceptable in view of the public interest, however it is questionable whether in the end even 1 year will be sufficient for data analysis. It is not unreasonable to assume that the time-limiting factor in genome research will soon become the data analysis and not the sequencing itself. This argues for early release of data so that all interested scientists can carry out analyses. Additionally, achievement of sequencing costs of \$0.50 per base or less will require technological advance which will render sequencing routine enough to be of little interest in the careers of research scientists. Commercial sequencing companies may be a solution to these problems. They could deliver data quickly and under contract to public databanks allowing the maximum resources to be brought to bear on interpretation of the data. Effective approaches must be discussed and policies formulated now, before significant quantities of data are generated.

Even if quality sequence data can be made rapidly available to researchers for interpretation the analysis task will be formidable. The computer will be as important as the lab bench in elucidating features and functions of sequences. But computer deduced annotation included in the databases will be of varying quality, often not confirmed experimentally and sometimes simply wrong due to inherent limitations of any algorithm. Today's databanks distinguish only between annotated and (a few) unannotated entries, but in the future the databanks will have to develop systems for representing the multiple levels of confidence and reliability of sequence annotation.

However, the analysis of genome data clearly does not end after the submission of the primary sequence data to the public databanks. Many researchers will eagerly wait for the data to appear in the databases to make them subject to their own analyses, and they might want to communicate their results to the databanks in order to attach their new findings to bare entries or to improve and correct existing annotation. The analysis of all the data obtained by genome research projects is expected to keep biologists occupied for many years. In fact, it has been suggested that biology will soon turn into a "theoretical" science where experiments will only be needed occasionally to test some hypothesis (Gilbert, 1991). Since the relevance of a particular part of the genome might only become evident long after it has

been sequenced, we will probably see a process of continuous updating of the data from genome projects for an extended period of time. Sequence databanks have to react by developing improved mechanisms which make it possible for them to attach annotation and to update annotation and sequences much later than usually done at present.

4. *Changing Requirements for Data Access and Data Distribution*

The first important protein sequence collection, Dayhoff's *Atlas of Protein Sequences and Structure* (Dayhoff, 1966), was published in the 60s and 70s in printed form as a book, and new releases from the database were only available every few years. Today's requirements for database distribution are totally different: databases have to be available in computer-readable form to make the information susceptible to computer analysis, and they do not only have to be complete but also as recent as possible. Ideally, a scientist wants to be able to access a machine-readable copy of a sequence at the same time as the sequence is published in a journal. That means for the databanks: fast data collection, small data processing time, and fast redistribution of data. As a consequence, nowadays all major databanks rely on modern computer technology to maintain and distribute their data collections. Almost all of the existing databases are available in electronically readable form, and, furthermore, most of them had to abandon the production of printed copies, simply due to the unmanageable size of the product and its uselessness to the researcher.

The traditional main mode of data distribution for the current sequence databases has been quarterly releases on magnetic tape. A quick look at the growth rate of the databases shows that there is always a large increment between two releases, and the concept of quarterly releases inevitably introduces some delay in data distribution which is no longer acceptable for many researchers. The future will call for rapid, perhaps daily, distribution of new sequence data and this requires improved communication channels between the database producers and the database users. This data transfer must be based on direct computer communication and modern computer networks, which is the only means that allows rapid distribution of data around the world in reasonable time. Computer networks will also become increasingly important for the success of genome projects which bring together researchers from all over the world. In order to achieve a maximum of co-ordination and to reduce redundant work as much as possible it is essential that there is excellent communication between the collaborators and in particular along the axis of sequencing labs, project databanks and public databanks, allowing them to exchange the latest data quickly and conveniently. Several projects are currently under way to explore new network-based channels for sequence data distribution and to improve the network infrastructure in the biological research community in general. They are discussed in detail in Section IV.

Although the future will certainly bring a continuous, daily distribution of data from the databanks to the scientific community, the need for regularly appearing compendiums of data will not disappear. Access to the latest data is crucial to many scientists, but others need to use sequence databases only occasionally, and they are perhaps not willing or simply cannot afford to invest resources in the constant updating of their local copy of the database. These users have to be served as well, and the databanks must therefore continue the production of new releases on a regular schedule. However, even for such distribution, magnetic tape is no longer the preferred medium. Since 1989 EMBL has distributed data on CD-ROM, a medium now also used by GenBank, and the use of other future media will no doubt be necessary.

The problems of data distribution mentioned before have in common the notion that sequence data has to be transmitted somehow to the individual researcher who, in turn, works on a local copy of the database. Although this is the usual procedure at present, it can be argued whether the steady increase of data will not call for some fundamental changes. The effort required to maintain local copies of a database should not be underestimated. Even now, many research groups simply do not want to spend time and resources maintaining and updating local copies of all the databases they use. This trend will continue with the ever-growing amount of data and the creation of new sorts of databases. Easy-to-use media like CD-ROM will certainly reduce some of the technical problems of database

maintenance, but do not help if access to latest data is essential or if several databases have to be maintained simultaneously. Similar maintenance problems arise with analysis software. For the average laboratory scientist it is almost impossible to keep an overview of the existing software in molecular biology and its usefulness, and even buying a comprehensive commercial software package will not solve the problem completely. The complexity of these packages necessitates considerable maintenance efforts, and, nonetheless, there will be specific problems which cannot be solved by a particular package.

The increasing demands of database and software maintenance make it important to explore alternative models where the data collections and analysis programs are accessible at one or more central places. Scientists can then work directly on these remote copies instead of using local copies, and the efforts to maintain these databases and software can be handed over to specialists. Larger centres will more easily solve problems of database and software maintenance, but for the average biologist a model where he or she can remotely access centrally stored data and choose from a variety of different programs will become increasingly attractive. The success of such a model is of course strongly dependent on the effectiveness and convenience of the connection between the researcher and the central database server, and on the services and the user support offered.

5. Integration of Databases—Linking Related Data Sets

In the past, sequence databases, map databases, structure databases, literature databases, and many others have existed as islands of information unconnected to each other (for a list of databases relevant to molecular biology see Lawton *et al.*, 1989). Current databases concentrate on small, limited areas of the biological knowledge, neglecting the complex network of interactions in living systems. This piecemeal approach, added to an endless number of different database formats, will become increasingly unsatisfactory. The future will see an increasing demand for the proper representation of the relationships between different kinds of biological data. The desire for “higher-order” or “second-generation” databases has been formulated previously (Pabo, 1987; Pongor, 1988), but it is far from clear what they should look like.

Obviously, it is desirable to integrate all available biological information; on the other hand, it appears unlikely that, for instance, all available information about a certain gene can be properly represented by a single database. Mapping data is a simple example. The current workplans for most genome analysis projects envisage as the first steps towards the determination of the genome sequence the establishment of refined physical and genetic maps in order to provide some guideline for sequencing. At present, there is no clearly defined standard for representing these maps, and comparison and integration of existing maps is not trivial. Recent proposals (Grausz, 1991; Olson *et al.*, 1989) based on the idea of “tagging” genetic and physical marker sites by short sequences open up the possibility to reconcile these maps and integrate them with the sequence databases. Nevertheless, information represented in these maps is conceptually different from sequence data, and it is not obvious how the relationships between linkage data, for example, and sequences can be properly indicated.

Genome analysis projects will also increasingly yield new data which do not fit into any of the existing databases at all. Whereas the determination of the nucleotide sequence is clearly the major goal of these initiatives, systematic genome research will, of course, also include research on other topics, guided by the sequence information available. We will see a wealth of diverse information coming from these projects, such as information about the regulatory network of the cell, structural organization of the genome, methylation patterns or spatial conformation of the chromosome.

It is both impractical and undesirable to merge all available information into one database, and thus the design of future databases should concentrate on individual data collections, best suited for the representation of the information contained in them, and on the creation of appropriate links between these data collections. It has been proposed to base the next generation of databases on the “most important relationships” (Pabo, 1987) or on a “systematic model of modern biology” (Rawlings, 1988). However, the complexity of

biological systems and the diversity of interests within the scientific community imply that there is a virtually endless number of possibilities for representing the relationships between different kinds of biological data and for creating links between them, thus limiting the chances for successfully choosing a single conceptual model for a higher-order database which would satisfy all database users (Waterman, 1990). In fact, this approach would introduce the same inflexibility which we currently have in the existing sequence databases, where the user is bound to a specific view of the data, dictated by the databanks' judgement of scientific relevance.

Living systems are not static; in contrast, they are characterized by a high degree of flexibility. We believe that in order to come closer to a "matrix of biological knowledge" (Morowitz and Smith, 1987), future databases must in some sense try to reflect this flexibility. Instead of limiting the database users to a certain model of biological interactions by building one comprehensive database from all the knowledge available or by defining explicit relationships between different data collections a new generation of databases must instead allow the user to build different links in many varying ways. This approach allows for independent data sets, and it would concentrate on defining *how* links between these databases can be created, and not *which* should be established, leaving the latter to the users or the developers of the necessary software to navigate between these data collections along a variety of modifiable links. A flexible approach is consistent with the experiences of the last decade of database design where the relational model (Codd, 1970) allowing flexible links between different sets of information (tables) has replaced systems where the links are explicitly designed into the database.

The difficulties which software developers will have to face with the next generation of databases will be formidable. At present, most database access software is limited to one database format or even to one specific database. Some recently developed programs (Etzold, 1990; PIR, 1990) are very flexible in handling different database formats and allow the user to query different databases at the same time, but they are still far from allowing free movement between databases. Although there is an obvious need for software which is able to visualize the relationships between different kinds of biological data, it is difficult to define the exact details of how this should be done. Even for a simple example such as the link between a protein-coding gene and the corresponding protein sequence there are numerous possibilities. One user might simply want to see the translation beneath the DNA sequence, another wants to see the sequence annotation of the protein database entry as well, and the next is interested in seeing even more complicated information like the relationship between the exon structure and the protein domain structure.

As a first step towards better links between independent data collections some databanks have already introduced pointers to other databases. Figure 5 shows the network of cross-references centred around the EMBL and Swiss-Prot databases. This cross-referencing mechanism is rather crude. It only connects complete entries, but the syntax of the new common DDBJ/EMBL/GenBank feature table (EMBL Data Library and GenBank, 1990) will allow the databanks to refine the details of cross-references to the sub-entry level, and cross-references from individual features to some external databases have indeed already been introduced into some GenBank entries. Nonetheless, we still have to wait for the first programs to be developed which make full use of these cross-referencing systems.

6. Customization of the Databases—Different Views of the Data Set

Similar problems to those just discussed for the integration of different data collections apply to entries from individual databases as well. The currently existing sequence databases are organized as series of entries, each one describing one sequence plus attached annotation. New releases of the databases are built by grouping these entries according to their level of annotation (PIR) or taxonomic criteria (EMBL, GenBank) or simply by appending all entries into one file (Swiss-Prot). This organization was felt appropriate when the databanks were established, but it only presents one specific "view" of the data set.

The growing size of the databases and the growing diversity of applications will soon render this approach inappropriate. It will become necessary to provide customized subsets

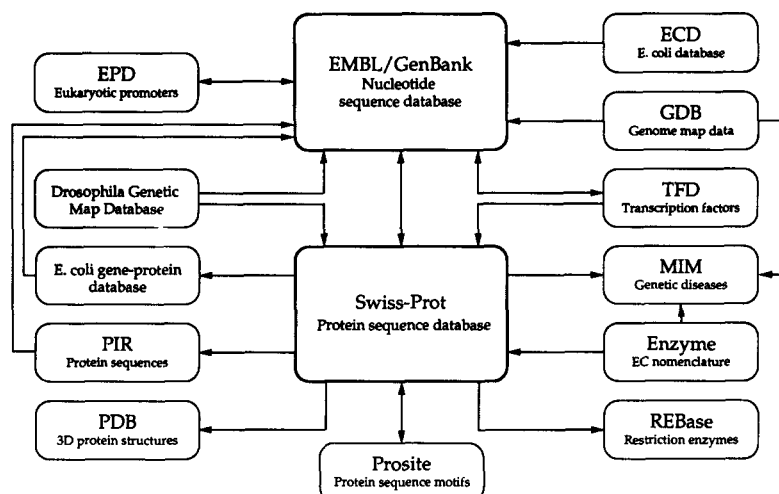


FIG. 5. Cross-references between the EMBL nucleotide sequence database, the Swiss-Prot protein sequence database and other data collections.

of information adjusted to the particular research requirements of different researchers. A simple step in this direction was the splitting of the DNA databases into several taxonomic divisions which allows a researcher who is interested in specific species to work with only a subset of the database.

A more complicated example of different views of the databases is the elimination of database bias due to the over-representation of some sequence families. A typical application of sequence databases is the comparison of a new protein or DNA sequence to all the sequences already in the databases. If no other information is available about this sequence, the identification of and the comparison to similar sequences might give important clues to the function and biological role of the new sequence. However, if, for example, a sequence is similar to a globin, an immunoglobulin or some other member of a heavily represented family in the databases, then a database query will find hundreds of hits with all members of that family. In fact, one hit with a characteristic representative of a family would give enough information to indicate the similarity to this gene or protein family. Once a hit is found with a member of a sequence family, the researcher should then be able to retrieve all members of this family from the database for closer scrutiny. A database of sequence family prototypes would, therefore, be very useful and work in this field is in progress (Bishop and Parsons, personal communication).

Another kind of customization simply affects the depth of information supplied with the database. Many scientists will only be interested in the sequence itself, while others might want to get as much additional detail as possible. A taxonomist will probably have little interest in information which might be of great relevance for someone else working on gene regulation. Different views of the databases are becoming increasingly important. The current implementation of the nucleotide sequence databases under relational database management systems (RDBMS) (Burks *et al.*, 1990; Kahn and Cameron, 1990) makes it possible to provide such customization. In addition to the traditional form of database releases the databanks could distribute their RDBMS tables or dumps of them, and customized collections of tables could be prepared to satisfy the requirements of different user groups. The EMBL Data Library is currently investigating these options. The fact that the internal database design of the sequence databases is optimized for data storage and management makes it necessary to transform the existing tables into a form which is more appropriate for applications such as database queries and data retrieval.

Customization by organism, sequence family or depth of annotation are all in principle supportable under existing relational schemata. Genome projects, however, will strain these models. The entry concept on which the sequence databases are built implies that an entry is

a well-defined unique entity, such as “the sequence for *E. coli* lys-tRNA” (Waterman, 1990). The limitations of this notion are already apparent, but genome initiatives will cause its complete failure. There is simply no such thing as “the genome”. It is impossible to find two genomes from one species which are identical. Mutation and recombination, the motors of evolution, guarantee that there is always some polymorphism in most genes. The genome projects will elucidate such information and so greatly intensify this problem. The current design of databases makes it difficult to properly represent highly polymorphic regions, different alleles of a gene, repetitive and jumping elements, and so on. The expected increase of this kind of sequence information will make it necessary to reconsider how we represent sequence data.

This problem is perhaps most evident when one wants to assemble sequences. A common criticism of the current nucleotide sequence databases is that of redundancy due to the existence of overlapping and identical sequences in the database. Consequently, data collections have been established, e.g. for *E. coli* (Rudd *et al.*, 1990), which contain species-specific sequence data, but which remove overlaps by joining adjacent sequences. In genome research the ultimate goal is the complete sequence of the genome. It is a matter of debate whether this should also be reflected in the sequence database, i.e., whether in the end there should be only one entry representing a whole chromosome or the complete genome. Although superficially reasonable, this approach quickly runs into major problems. Figure 6 shows how contiguous or overlapping sequences from the database could be merged into one consensus sequence. Differences in overlapping regions, such as ‘x’ and ‘y’ in Fig. 6, require some “polishing” in order to remove these “small” discrepancies. However, overlapping

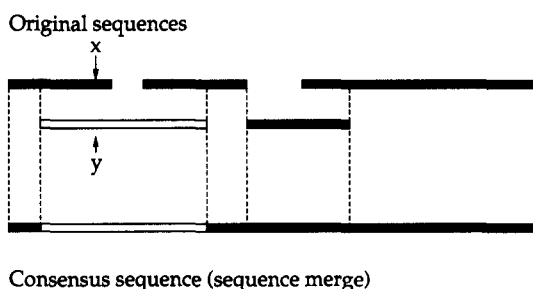


FIG. 6. Merging individual entries of the nucleotide sequence databases. x and y symbolize small differences between two overlapping sequences.

sequences may come from different strains or from different tissue of an organism, and, depending on their research interests, some scientists would merge them while others would not. A researcher interested in genome organization might not care at all about these “small” differences; however, for somebody working on polymorphisms such sequence merges would be catastrophic. It is therefore evident that the sequence databanks must be extremely cautious in assembling sequences, and that a design for future databases should allow this possibility while still retaining the original data accessible.

Although the complete continuous sequence of the genome is the conceptual goal, it poses practical problems as well as the more fundamental objections discussed above. Some of them are simply due to current hardware and software technology which does not allow the convenient handling of sequences which are more than a few hundred thousand bases long, some orders of magnitude smaller than a typical chromosome sequence, but it is not only the sequence data, but also the attached information which will add to the size and complexity of the database. In the current database design all this information is attached to the sequence in one entry, and it is hard to see how a “mega”-entry including all the information about thousands of kilobases could be handled in a convenient fashion. While the rapid progress in hardware and software technology may certainly help to overcome the technical problems of storing and manipulating huge entries, future sequence databases should support both the

inspection of specific regions of the genome in detail and the analysis of the overall genome organization.

IV. COPING WITH THE NEW REQUIREMENTS—STRATEGIES FOR THE SEQUENCE DATABANKS

The challenges which the concerted, international attempts to sequence complete genomes will produce for the sequence databanks are daunting, but there are a number of novel approaches in the areas of data acquisition, management and distribution which can be explored in an attempt to solve some of these problems.

1. Data Acquisition

The two most important requirements for any sequence database are completeness and timeliness. The databanks will therefore most severely be hit by the expected increase of data in these two areas.

Although it is pleasing to see that some 80% of new sequence data is nowadays submitted directly from the authors on diskette or via electronic mail networks (see above), it can be seen from Fig. 3 that the remaining 20% of data is as much as all the data a few years ago. It is of great importance for the nucleotide sequence databanks that they continue and intensify their efforts to raise this percentage as much as possible; obtaining the highest possible rate of direct submissions is crucial for the success of their operation. The importance of direct submissions of their data to the databanks can be hardly over-emphasized:

- only directly submitted data can appear in the database quickly;
- only direct submissions guarantee that a sequence is not missed;
- directly submitted data is more accurate;
- better annotation can be built on the expert information supplied by the submitter.

The obvious incentive for the observed increase of direct submissions of sequence data to the databanks has been that most leading journals have recently made data deposition in the sequence databases mandatory for the publication of a manuscript containing new sequence data. Although this system has proven to be very effective—resulting in a decrease of the average processing time for new sequences from many months to a few weeks—there is still a long way to go to educate the scientific community to accept direct sequence data submission as an integral part of the scientific publication process.

Although most scientists who sequence DNA frequently use computers for data handling, it remains surprisingly difficult and time-consuming to prepare a sequence submission, which obviously negatively affects the motivation to do so. In an attempt to simplify this process, GenBank has developed a computer program which guides researchers in preparing a submission (Burks *et al.*, 1990; Moore, 1988). This program, called Authorin, is available for IBM-compatible and Macintosh computers and formats the information so as to allow automatic incorporation in the database. It would be desirable to provide this support at even earlier stages of the sequence determination and analysis process, for instance as an integral part of sequence analysis packages or the software supplied with automated DNA sequencers. Closer collaboration between software companies and databanks is clearly necessary in this field to encourage the production of sequence submission modules for sequence analysis programs.

The importance of the databanks as a prime source for sequence information has already been appreciated by many scientists. The nucleotide sequence databanks observe a steady increase in submitted sequence information not intended for publication in printed form, and a growing number of researchers also provide the databanks with corrections of previously submitted sequences and additional, newly discovered information. This trend supports the assumption that in the future sequences will be published in journals less frequently than today. Nevertheless, there will always be a certain amount of sequence information which has to be gleaned from the literature. Scanning the literature is a time-consuming task which, given the number of scientific journals, is bound to be less than completely successful.

In order to improve the efficiency of data acquisition by journal scanning, the nucleotide sequence databanks are investigating possible collaborations with the producers of literature

databases. Scientific literature databanks like MEDLINE or EMBase routinely scan thousands of journals and extract relevant information from the articles published therein. The forthcoming NCBI GenInfo Backbone Database (NCBI, 1990) is the result of probably the closest collaboration between a sequence databank and a literature databank to date. A new set of indexing terms has been introduced in MEDLINE to identify all articles reporting nucleotide or protein sequence data. This information is utilized by NCBI to produce a sequence database of all sequences published in the scientific literature. Although this close collaboration is strongly facilitated by the fact that NCBI is a part of the National Library of Medicine which is also responsible for the production of MEDLINE, it may nevertheless serve as a model for other sequence databanks to overcome the problem of scanning the literature. In particular, possible approaches for a closer collaboration between European publishers, European literature databanks and the EMBL sequence databank are currently being evaluated.

Besides direct submissions and journal scanning, the third main route of new sequence data into the nucleotide sequence databases is the exchange of new database entries between the collaborating groups. At present, the databanks exchange their data on a daily basis by sending new entries to the other sites by electronic mail in their flat format. In the past, mail transfer problems and format conversion caused the loss of some information, resulting in some inconsistency of the common DDBJ/EMBL/GenBank data collection. The installation of the nucleotide sequence databases in a relational database management system at the collaborating sites opened up ways of improved data exchange, and in 1990 the nucleotide sequence databanks agreed on a new data exchange protocol to be implemented during 1991. Instead of shipping flat files, the data exchange protocol allows one to send transactions which directly modify the remote databases, keeping them synchronized. This will be an important step towards the unification of the nucleotide sequence databases. A similar data exchange protocol has also been developed by PIR-International (Mewes, personal communication).

Important as these data acquisition streams are, the volume of data they generate will be overshadowed by that generated by genome sequencing projects. Such information will be transferred directly from local, project-specific databanks into the central databases. The internal structure of local project databases may vary considerably and will probably be different from the format used by the main sequence databases. Thus it is necessary to concentrate on proper data exchange mechanisms which transcend the internal data representations and which allow the highest possible automation of the data acquisition process. At EMBL, we have developed and are currently testing procedures for the automated integration of sequence data submissions from the European yeast chromosome III and the *Caenorhabditis elegans* projects.

The sequence databanks may become involved in the genome projects at very different levels. They may simply pick up the sequencing results from project-specific databanks, they may act as project databanks themselves, or may even provide the sequence analysis of new sequences. In any case, early and close collaboration with genome initiatives and their informatics resource centres is essential to guarantee a smooth transition for new sequence data from the local working databases to the main public repositories.

2. Data Distribution

The rapid increase of sequence information has made the traditional distribution of new quarterly database releases on magnetic tape increasingly unmanageable, and forced the sequence databanks to investigate alternative, state-of-the-art technologies.

EMBL, and more recently GenBank, have begun to encourage CD-ROM as the preferred medium for their database distribution (Cameron, 1989). The most obvious advantage of the CD-ROM is its storage capacity; a single EMBL CD-ROM contains not only the EMBL nucleotide sequence database, but also the Swiss-Prot sequence database plus a variety of other important molecular biological data collections. In addition to high storage capacity, CD-ROM offers the advantage of robustness, low cost of production and distribution, and, perhaps more importantly, the existence of an ISO standard (ISO 9660) renders the same

CD-ROM usable on a wide range of machines. EMBL's subscription figures show an increasing preference for CD-ROM over magnetic tape, especially at the low end in the PC environment. Such users had previously little chance to use magnetic tapes but can inexpensively buy CD-ROM drives for their laboratory PC or workstation. Although the transfer of the information to other storage media such as high-capacity hard disks which allow faster access is certainly possible, many users will prefer to analyze data directly on the CD-ROM. Important for the success of CD-ROM as the standard medium for database distribution will, therefore, be the availability of software which enables one to work with this device. To encourage users of small computers data retrieval and database searching software for IBM-compatible personal computers is supplied on the EMBL CD-ROM (Higgins and Stoehr, submitted), and several academic groups have communicated that they are working on similar software for the Macintosh. Additionally, well-known programs such as FASTA (Pearson and Lipman, 1988) have been modified to work directly with the databases on the CD-ROM.

Despite all advantages, CD-ROM suffers one disadvantage. Most of the cost is in mastering and preparing a given CD thus rendering it only suitable for periodic releases of the databases rather than continuous updates. This disadvantage, which applies to magnetic tape releases as well, can be overcome making intermediate data available via computer networks.

As a simple but effective means for direct access to the latest sequence data, the EMBL Network File Server was established in late 1987 (Stoehr and Omond, 1989). The File Server is a facility available on the EMBL computing system enabling external users to retrieve files via electronic mail. Any scientist who has access to a wide-area computer network such as Internet or Bitnet/EARN can retrieve data from the file server by sending commands in a simple, well-defined language to the EMBL computers, which will then automatically return the requested information. The File Server not only offers access to the most recent release of the EMBL and Swiss-Prot databases, but also to the newest entries in these databases as soon as they are created at EMBL. Because sequence data is exchanged between EMBL, GenBank and DDBJ on a daily basis, the latest GenBank and DDBJ entries are available as well. The success of this service has encouraged EMBL to extend the initial range of data collections offered on the server and a variety of different molecular biological databases can now be accessed. Currently, about 3000 requests are processed each month. In the meantime, similar services have been established at other sites as well, some of them exploiting the advantages of the Internet file transfer protocol (ftp) instead of using standard mail for access (Davison and Chapple, 1990; Yudin, 1990). Recent developments also include the introduction of new functionality for these servers such as database queries and sequence comparisons over the network (Fuchs *et al.*, 1990) which seem to be particularly attractive for those scientists who do not want to maintain local copies of the database but nevertheless want access to the latest sequences. With all these file servers or ftp servers being of differing size, content, functionality and timeliness the molecular biologist can now choose from a variety of services according to the kind of information he is looking for and the network connectivity available (Gribskov, 1990).

Although e-mail and ftp servers have been very successful in the past, their usefulness for the wide-spread distribution of recent data is limited. If a scientist is simply looking for an entry whose accession number was given in a publication, the task of retrieving the database entry using a file server is trivial. Maintaining a complete and up-to-date copy of the sequence database by this means, however, is cumbersome and inconvenient. Although the sequence databases offer some help by providing daily updated index files and weekly batches of new entries, it is the responsibility of the individual biologist to guarantee that his local copy is complete by explicitly requesting all of the new entries from the server.

In early 1990, in an attempt to overcome this problem, GenBank introduced a new mechanism for data distribution (Smith *et al.*, 1991), based on the Usenet logical computer network (Horton and Adams, 1987) using the NNTP Network News Transfer Protocol (Kantor and Lapsley, 1986). Every new database entry is simply treated as one new message sent to an electronic bulletin board (newsgroup). New messages, i.e. entries, are

automatically spread over the network and forwarded to any site which has subscribed to this newsgroup. The distribution of new GenBank entries is part of the international BIOSCI newsgroup system, which has more than 1000 subscribers all over the world. The data exchange protocol has proven to be very efficient, and software is available for managing the received new entries and updating local copies of the sequence database. Data distribution via Usenet alleviates the task of maintaining a local copy of the database. NNTP will take care of the updating by automatically determining which new entries are missing in the local data collection. Therefore, the efforts for the biologist who wants to receive the latest data is minimal. It also reduces network traffic, thus saving bandwidth, because entries are not independently transmitted from the databank to each recipient, but distributed in a tree-like fashion.

Whereas the Usenet model is certainly an elegant solution at present, it is arguable whether this model will cope with the drastic increase of data in the future and its implications for data distribution. It is built on the concept of distributed local copies of the database, and, as argued above, we may see a future preference for remote access to centralized databases for the majority of users. Additionally, Usenet/NNTP data transfer is limited in its functionality just as is electronic mail- or ftp-based data exchange. Many applications such as complex database queries simply call for more flexibility and, perhaps more importantly, interactive access to the data collections.

An early attempt to satisfy these requirements was the BIONET network which was initiated in 1984. BIONET was as a non-profit resource for molecular biological computing funded by the NIH and run by IntelliGenetics, Inc. (Roode *et al.*, 1988; Smith *et al.*, 1986b). For a small annual charge, access was provided to a range of important databases and to a comprehensive set of analysis software. BIONET funding was discontinued in late 1989 and the service was superseded by the GenBank On-line Service (GOS), also run by IntelliGenetics (Benton, 1990). Different levels of services are now offered at different fees, including database queries and searches, sequence analysis and access to electronic mail networks and bulletin boards. The usage of GOS is not restricted to American users, however, telecommunication costs may become prohibitively expensive from other countries. Although GOS is perhaps the most prominent example of a molecular biology on-line service, there are nonetheless several other, mostly regional, resources of this type (Smith *et al.*, 1986).

American computer networking capabilities make a service such as GOS an appropriate solution for the United States. The situation in Europe is significantly different. Present academic and commercial telecommunication networks in Europe suffer from cross-border delays and charges, and academic and commercial networks are not well integrated. Costs for data connections between partners from different countries are often extremely high, and line speeds are often rather low. These factors greatly reduce the potential effectiveness of electronic communication in Europe. The need for efficient communication networks to receive and distribute biological data within Europe has been realized and acknowledged in two recent studies performed on behalf of the E.C. and the European chemical industries (CEFIC, 1990a,b). It was recommended that the E.C. support the development of European research networks and improve the necessary infrastructure, and it was estimated that the EC should spend at least 10 million ECU per year on bioinformatics.

In 1988, EMBL initiated the European Molecular Biology Network (EMBnet) project in order to take the first step towards this goal. This approach is also based on the provision of on-line services, but, in contrast to the GenBank On-line Service, the EMBL model envisages a network of nodes each acting independently but in a co-ordinated manner. Europe is extremely heterogeneous in terms of science, politics, culture and language, thus favouring the establishment of a decentralized network of nodes which serve individual countries in contrast to one centralized service. The EMBnet strategy is built on the idea of having national nodes in each collaborating country, selected by governments or research councils, which provide comprehensive biocomputing services to their national academic and commercial user communities. The operation of these national nodes is independent of each other and in fact rather heterogeneous, but all nodes are linked via DECnet and TCP/IP

networks. In 1991, national centres were established in 12 countries, listed in Table 4. The national nodes are supplied with the latest sequence data every night by EMBL, enabling them to provide a complete and up-to-date sequence collection to their users. In addition to the on-line services they offer, many nodes also redistribute the sequence data within their countries to other academic and commercial institutions, thus updating approximately 40 remote copies of the EMBL database in Europe at present. The involvement of commercial partners is considered to be a vital element of EMBnet, in contrast to most other projects in this area, which severely neglect the requirements of commercial biotechnological and pharmaceutical companies, and thus Hoffmann-LaRoche has been involved in the EMBnet project from the beginning.

TABLE 4. THE EUROPEAN MOLECULAR BIOLOGY NETWORK (EMBnet)

<i>National EMBnet nodes</i>	
Denmark	Biobase, Aarhus
France	CITI2, Paris
Germany	DKFZ, Heidelberg
Greece	IMBB, Crete
Israel	Weizmann Institute, Rehovot
Italy	University of Bari
The Netherlands	CAOS/CAMM Centre, Nijmegen
Norway	Institute of Biotechnology, Oslo
Spain	CNB, Madrid
Sweden	Biomedical Centre, Uppsala
Switzerland	Biozentrum, Basel
U.K.	SERC Daresbury Laboratory, Warrington
<i>Other nodes</i>	
EMBL	Heidelberg (co-ordinator, database provider)
Hoffman-LaRoche	Basel, Switzerland (industrial node)

While the EMBnet project currently mainly concentrates on establishing the necessary connectivity within Europe and improving the mechanisms of sequence data distribution, its scope is much broader. Efforts to install a network-wide conferencing system are well under way, and other network services such as remote access to specialized facilities are being investigated.

3. Data Handling and Storage

In the area of data management it is foreseeable that the flow of data from genome research initiatives will soon push the currently available hardware and software to their limits. But CPU, disk and memory prices are plummeting, hardware is constantly being improved and new computers are introduced every few months which are faster, better, and cheaper. Progress in this sector is so rapid that it is likely that future hardware technology will guarantee that the operation of the databanks will not be severely affected by purely technical problems.

Software limitations, on the other hand, may pose more of a problem. Some major databanks still use "home-brew" software for maintaining their data collections, while others have recently moved a step forward and installed their data under commercially available relational database management systems (RDBMS) (Burks *et al.*, 1990; Kahn and Cameron, 1990). However, relational database management systems were developed with business applications in mind, and therefore they have some limitations which restrict their effectiveness for molecular biological applications, particularly sequence handling and manipulation.

Object-oriented databases are often the subject of current discussion as an alternative to RDBMS, and it is often claimed that they are more suitable for the management of biological data (Gray *et al.*, 1990). Unfortunately, object-oriented database management systems are very diverse, and, in contrast to relational systems where vendors have agreed on SQL as a

common language, no such standard is in sight for object-oriented systems. While object-oriented systems promise much for the future, the relational model, despite all its restrictions, is well tried and probably more appropriate for the operation of a production database. In order to ensure the continuity of their operation, the major databanks are very cautious about any changes to their existing data handling systems, but the importance of keeping abreast of the current developments in computer science and database technology is nevertheless well-recognized and progress in this field is carefully observed.

4. Annotation

An analysis at EMBL of the time spent on different aspects of handling an entry—from data acquisition to data distribution—showed that most of the workload is not in acquiring sequences but in attaching detailed biological information. If a new sequence is entered from a journal publication, the annotator has to carefully read and understand the article in order to extract the important information which applies to the new sequence. This process of extracting information from an article is a time-consuming task. Annotating a new sequence is greatly simplified if the sequence and the relevant biological information is directly supplied by the author in a suitable form. As already discussed, the databanks strongly encourage direct submissions, and tools like Authorin (Moore, 1988) will help the scientist to prepare sequence submissions. These tools also diminish the need for cross-checking of submitted data by the databanks, because much of the necessary checks have been done locally by the programs when the sequence was submitted. At present, the annotation staff routinely check the information supplied by a submitter or extracted from the literature for internal consistency and against the sequence data, for instance by translating a protein-coding region and checking for frameshifts and stop codons. The number of problematic cases is surprisingly high. About 30% of submitted data shows some inconsistency which makes it necessary to go back to the submitter to clarify these issues, introducing an additional delay between data submission and publication.

All of the present databanks aim for the highest possible quality of annotation which necessarily requires a high degree of specialized knowledge. Although all annotators are at least graduate biologists the wide range of different areas in molecular biology simply makes it impossible to have specialists in all fields. It is important to get as much expertise as possible from the scientific community itself. A scientist who submits a sequence is in a much better position to provide relevant information than an annotator who has to extract it from an article, and, therefore, soliciting information from the submitter is very important. However, the best annotation would contain information which no one submitter could supply because it would depend on a detailed study of the relationship between entries, and, even where data were annotated to a high standard, evolving scientific understanding and terminology would require constant updates. In order to improve the quality of annotation not only on the level of individual entries but also for families of related sequences, and to introduce a consistent usage of standardized nomenclature, the nucleotide sequence databanks have recently initiated efforts to encourage scientists with extensive knowledge in specific areas to share this knowledge by contributing to the annotation of database entries.

The approaches taken by EMBL and GenBank are distinct. GenBank's "curator program" (Burks *et al.*, 1990) envisages that experts in particular domains of molecular biology are equipped with the necessary tools to update the GenBank database remotely over computer networks by accessing the database from their laboratories and working on the existing annotation. GenBank expects to bring 15 curators on-line over the next 3 years; the first of them have already started, working on *E. coli* nomenclature and vector contamination in the database (Gilna, 1991).

The EMBL efforts are different because they are based on the concept of special annotation databases, thus separating the sequence information from the annotation. Biological knowledge is built up in specialist databases remote from the central sequence database in a way that it is designed to be used with the sequence database. The information supplied by the experts is kept separately from the sequence database and organized in a way which allows the best possible representation of this kind of data. The term "affiliated data

unit" (ADU) has been coined to describe such databases. The first two prototypes of these ADUs are the Eukaryotic Promotor Database (EPD), maintained by P. Bucher at Stanford (Bucher and Trifonov, 1986), and the *Escherichia coli* Database (ECD), compiled by Kröger *et al.* (1990). The formats and contents of these databases are very different, but both of them contain no sequence data, the information contained in them being linked to the EMBL database via pointers to EMBL entries. EPD provides important information about eukaryotic promoters in the EMBL database, whereas ECD is a compilation of *E. coli* sequences, containing additional information about genetic map locations and the construction of contigs from these sequences. Other collaborations of this kind are being planned.

These affiliated data units can be seen as a first step towards the next generation of biological databases as envisaged by the model presented below. The experiences gained from these projects are extremely important in order to further refine the presented model.

V. A MODEL FOR THE NEXT GENERATION OF SEQUENCE DATABASES

1. Principles of the Model

Although there is plenty of scope for further improvements of the operation of the current sequence databanks, it is questionable whether their basic design will allow them to cope with the volume and complexity of genome data, providing the necessary flexibility of data representation. Instead, a new, conceptually different, generation of sequence databases must be built to meet the challenges of the future.

The model presented here distinguishes between three different conceptualizations of the information in the sequence databases:

Exact representations of scientific reports.

Interpretations of these data to reflect our hypothesis or "best bets" as to what the information in the cells of organisms is.

The data as they are in nature—the real information in which scientists are actually interested.

Current databases confuse these concepts, in particular the scientific reports and their interpretation. We believe that it is crucial for the success of future databases that the tasks of data collection and data interpretation be clearly separated; they are both crucial, but distinct. Hypotheses about how the information in the scientific reports should be interpreted and assembled are important, but even the most minor interpretation of the underlying data is a subject of judgement. In our view the first task of the sequence databanks should be to ensure that hard information from scientific reports is presented correctly and consistently. Interpretive work on this information is necessary, but the task of collecting it is significantly different from that of collecting the basic sequence data. At present, the centralized databanks attempt both tasks.

In fact, the PIR protein databank has deliberately developed a conceptual model of a sequence database, exemplifying a "second-generation database" (Pablo, 1987), which explicitly aims at adding an additional level of interpretation or inference to the raw sequence data, according to a schema which represents a particular view of the scientific literature (George and Barker, 1990). In this approach sequence data is analyzed and reviewed in depth by scientific staff, and emphasis is on reporting scientific knowledge rather than on accurately representing the results of individual reports (Barker *et al.*, 1990). Although the attractiveness of this model is unquestionable we believe that it is not appropriate for coping with the future requirements. The user of such a "scientific database" (George and Barker, 1990) is inevitably restricted to a certain view of the biological data, and the model does not provide the necessary flexibility of data representation. Additionally, the scientific analysis work which is necessary to organize such a database is extremely demanding. It is unlikely that it will be possible to raise the necessary resources in accordance with the increase of data. Indeed, this discrepancy was acknowledged when the delay between the publication of a new sequence and the appearance of a fully-annotated database entry which fulfilled the self-imposed high quality standards became unacceptably long. The PIR database is now divided

into three different partitions with varying degrees of attached biological data, from fully annotated to unverified data (Barker *et al.*, 1991). This fact clearly illustrates the difficulties which arise from a model that tries to reconcile timeliness and completeness of data representation with intensive interpretive work. These difficulties can only increase with genome sequencing where the rate of acquisition of sequence data will rise dramatically with biological understanding lagging far behind.

In the model proposed here the central databank concentrates on building a database of scientific reports, while other groups, which could be located anywhere in the world, produce data collections containing the interpretations of the underlying raw data. Clear standards have to be defined in order to link the interpretive information to the sequence data and to make the complete knowledge available in a suitable manner.

It should be pointed out that the importance of separating data and interpretation is consistent with the model adopted by the GenInfo Backbone Database (GBD) being established at NCBI (Benson *et al.*, 1990; NCBI, 1990). GenInfo will concentrate on collecting all available sequence information from the literature, but will only contain minimal annotation such as information about protein-coding regions. Detailed annotation is expected to be carried out by remote groups, linking their information to GBD via pointers.

The core of our design (Fig. 7) is a stable, citable *backbone* database, representing a collection of all kinds of scientific reports, including publications, direct submissions, patent applications, etc. These sequences are referred to as *real entries*, and are identified by unique, meaningless and unchanging entry codes, equivalent to today's accession numbers. Each real

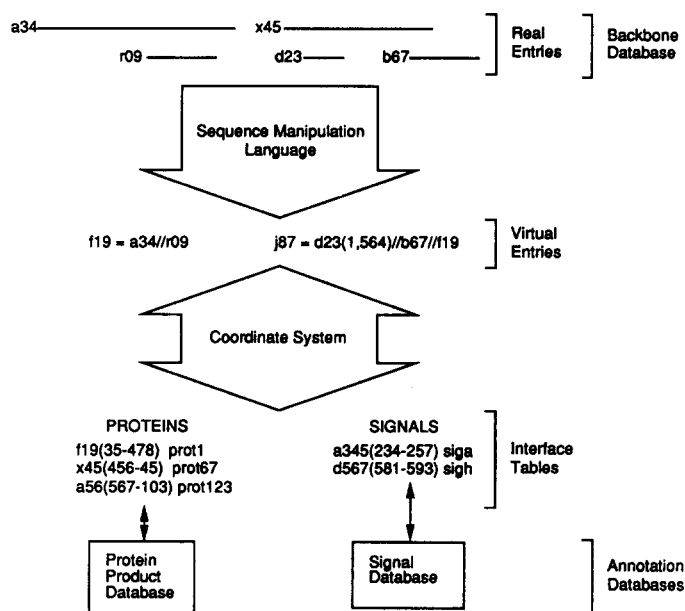


FIG. 7. The EMBL model for the next generation of sequence databases.

entry will represent a single sequence as reported. It will never be modified and will always be accessible in the database under its entry code, clearly resulting in an underlying database with many errors and overlaps. This is unlikely to pose problems for tomorrow's storage media, but is certainly not what the database user wants to see. The end user will, therefore, typically not see the real entries of the sequence database themselves, but only cleaned views of them. Such views will mainly consist of *virtual entries* which are generated by applying corrections and updates to real entries or by assembling real entries into larger contigs. Unlike real entries the sequences of virtual entries are not stored as such, but as instructions on how to assemble or modify real entries in the database. The instructions are based on a

sequence manipulation language (SML). Virtual entries are identified by unchanging entry codes, exactly like real entries.

Together, real and virtual entries constitute the sequence database. The entries in this database will only contain sequence data and minimum information to identify the sequences, such as literature references.

In this model detailed biological annotation of sequences will be prepared by groups independent of, but co-ordinated with, the central database. Such *annotation databanks* will retain sufficient design autonomy to organize the information from their specialist area into a representation most suited to their perceived needs. The annotation databases and the sequence database will be connected via standard *interface tables* which link objects in the annotation database to the relevant parts of sequences. The sequence manipulation language which is used for the construction of virtual entries can be applied to this purpose as well.

Although we used the term “annotation databank” here to explain the principle of independent data units linked to the main backbone sequence database, it is obvious that this concept is not restricted to attaching additional biological information to sequences. The same mechanism can be used as well for producing specific views of the database by extracting and grouping sequences from the backbone database, for example to produce databases of non-redundant information for studies of genomic organization.

The advantages of separating the task of data collection and data interpretation as proposed by this model are:

- Data collection and data interpretation can be performed by experts in these fields, concentrating on their part of the job.

- Different hypothesis and user views can coexist. The database user is not restricted to a certain conceptual model used for the representation of the data. Different and even contradictory interpretations of the same sequence data are possible. Annotation databases can be added or removed, thus interpretive work can keep up with the development of the science.

- Annotation databases can adopt the structure best suited to the representation of a particular specialist area.

- Any changes to the annotation databases do not affect or interfere with the underlying raw data. The backbone is stable.

- The maintenance of the sequence database and the maintenance of the annotation databases are uncoupled and only dependent upon each other in a limited way.

- Funding for the backbone database is separate from funding for the annotation databases.

- Where interpretation is more a research than a service activity it can compete for research funds rather than service funds.

As a consequence of our model, the traditional concepts of database “entries” as the combination of a sequence and all related biological information and “flat-file databases” as collections of these entries do no longer apply. Instead of combining all information related to a particular sequence into one entry as done currently by the sequence databases, resulting in all kinds of difficulties when there is no clear one-to-one relationship between sequence and biological feature, our future model keeps sequence data and related biological information clearly separated, thus allowing one to link one sequence to several features and *vice versa* without being restricted by the limitations of the entry concept.

2. Some Details of the Model

Real and virtual entries in the sequence database will never change. Once a sequence has been assigned an entry code it will always be accessible under that code. If a sequence has to be corrected, a new entry will be created. This will typically be done by building a virtual entry which specifies the modifications necessary to the original entry. Virtual entries can be built from both real and virtual entries, thus there is no limit to the number of corrections which may be applied consecutively. Thus, in order to modify a virtual entry a new virtual entry is created pointing to the underlying virtual entry and specifying the necessary modifications.

The permanence of entries and entry codes is crucial to the success of the independent

annotation databases. The current ways in which databanks constantly modify and merge entries make it difficult to build related databases which refer to sequences or parts of sequences. Indeed the entire model is completely dependent on the consistent use of unique, unchanging identifiers for objects to be referenced in all participating databases. This ensures that updates can never render links between databases invalid. In the past the unfortunate practise of referring to database entries by their mnemonic names (for instance, EMBL entry names) has created problems whenever names were changed.

A consequence of the requirement that identifiers must be unchanging is that they must also be meaningless. If any biological meaning (say the organism from which the sequence originates) is coded into the identifiers, then, as soon as an error is made, an inconsistency is created which either must persist or be corrected by changing the (unchanging!) identifier. From our experience, errors of this kind, for instance resulting from experimental errors, from misinterpretation of experimental results, or simply from confusing sequences from two figures in an article, are inevitable.

Virtual entries in the backbone database will be described by a sequence manipulation language. Elements of the same language could also be used by the annotation databanks to refer to specific DNA sequences in the backbone database. This language will provide the necessary operators and a co-ordinate system for manipulating real entries without directly affecting the raw data. In order to allow automatic processing, a formal description of the SML has to be developed. The SML only requires a small set of operators necessary for the manipulations to build virtual entries in the backbone database. These operations most notably include the merging of entries and the modification and addition of residues. A language, DNA*, which provides most of these capabilities has been described previously (Schroeder and Blattner, 1982), however, linking annotation to sequences would require extensions to this language. In some cases, for instance, it might not be possible to clearly define the ends of a DNA segment, and the SML must allow for some ambiguity. An improved and better defined version of the language used by the common DDBJ/EMBL/GenBank feature table might be a good starting point for the development of a suitable SML.

Crucial for the success of this model is the design of the links between the backbone database and the annotation databases and probably between different annotation databases. Our model does not enforce any specific internal structure on the annotation databases, but leaves it up to their developers to create a schema which is best suited for representing their data. It is important that the exportable objects in these databases and the interfaces to these objects are exactly specified and formally described. The fact that the formats of almost all current databases are not defined in this manner is a major problem for all software developers at present. For example, it is very difficult to refer to any object except a complete entry. Formal descriptions of annotation databases, however, will allow arbitrary connections to be built between objects on a sub-entry level. Amongst the several data specification languages which have been proposed in the past, Abstract Syntax Notation 1 (ASN.1) is probably the most attractive one because it has been defined as an international ANSI/ISO standard (ISO 8824, 1987; ISO 8825, 1987). Several tools are available for working with ASN.1, and the commitment of the National Center for Biotechnology Information to using ASN.1 as the basic means for the description and exchange of information in the GenInfo database (Ostell and Wooton, personal communication) will certainly give an additional impetus to the application of this language in the area of biological databases.

Linkage between objects of different databases can then be created by means of interface tables describing relationships between the objects. In contrast to other models of future databases (Pabo, 1987; Rawlings, 1988), our approach does not try to define the links between the databases, assuming some commonly agreed conceptual model of biology, but it concentrates on standardizing the interface, thus allowing one to create links as desired. These links have to be defined outside entries in the databases in order to facilitate adding and deleting links without affecting the underlying data. Although we used the term "interface tables", this does not imply that our model is requiring a relational database scheme. While a relational model seems to be appropriate for some purposes, annotation

databases can nevertheless be built on other principles, for instance object-oriented approaches, as long as the interface is clearly defined.

As attractive as the concept of the centralized sequence database with a surrounding network of value-added annotation databases is, it can only succeed if the composite of information is accessible to users in a convenient form.

Firstly, researchers need not have to explicitly assemble the information from various annotation databases. A more suitable system will be one whereby annotation databanks transmit information back to the central databank which would collect these data sets and distribute them, as local copies, to the users. Future models may involve the processing of user queries by accessing distributed annotation database servers on the fly, but today's European networks will not yet support this.

Secondly, a new generation of database query software will be required for the appropriate presentation and utilization of the information organized according to our model. Software to explore such information will be very different to that used today and its development will be challenging. The current view where a sequence and all associated information are packaged into entries which are concatenated into flat files is already strained and will become completely unworkable in the future. Programs must be developed which will provide the database users with a hitherto unknown flexibility in data manipulation, allowing them to freely navigate between the sequence database and the annotation databases and to build their own views of the underlying information. Object-oriented methodologies seem very promising for this task. There is much room for independent software developers and commercial companies to engage themselves in this challenging field, and in fact it is likely and desirable that, in analogy to the separation of tasks between the backbone database and the annotation databases, the database producers will provide the fundamental data but that the software enabling researchers to manipulate and analyze this information will be supplied by independent groups.

ACKNOWLEDGEMENTS

This work was financially supported by the Services of the Commission of the European Community under the Biotechnology Action Programme (BAP). The development of this model would not have been possible without significant contributions from Patricia Kahn. We also like to thank Howard Bilofsky and Peter Stoehr for critically reading the manuscript.

REFERENCES

- ALWIN, J. (1990) United Kingdom human genome mapping project: Background, development, components, coordination and management, and international links of the project. *Genomics* **6**, 386–388.
- ANDERSON, A. (1989) Full sequence for *E. coli*. *Nature* **338**, 283.
- BACHMANN, B. J. (1990) Linkage map of *Escherichia coli* K-12, edition 8. *Microbiol. Rev.* **54**, 130–197.
- BAIROCH, A. (1991) PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res.* **19** (Suppl.), 2241–2245.
- BAIROCH, A. and BOECKMANN, B. (1991) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.* **19** (Suppl.), 2247–2249.
- BARKER, W. C., GEORGE, D. G. and HUNT, L. T. (1990) Protein sequence database. *Meth. Enzymol.* **183**, 31–49.
- BARKER, W. C., GEORGE, D. G., HUNT, L. T. and GARAVELLI, J. S. (1991) The PIR protein sequence database. *Nucleic Acids Res.* **19** (Suppl.), 2231–2236.
- BARNHART, B. J. (1989) The Department of Energy (DOE) human genome initiative. *Genomics* **5**, 657–660.
- BENSON, D., BOGUSKI, M., LIPMAN, D. J. and OSTELL, J. (1990) The National Center for Biotechnology Information. *Genomics* **6**, 389–391.
- BENTON, D. (1990) Recent changes in the GenBank On-line Service. *Nucleic Acids Res.* **18**, 1517–1520.
- BILOFSKY, H. S., BURKS, C., FICKETT, J. W., GOAD, W. B., LEWITTER, F. I. *et al.* (1986) The GenBank genetic sequence databank. *Nucleic Acids Res.* **14**, 1–4.
- BITENSKY, M. (1986) *Sequencing the Human Genome Workshop*, Los Alamos, Los Alamos National Laboratory.
- BUCHER, P. and TRIFONOV, E. N. (1986) Compilation and analysis of eukaryotic POL II promotor sequences. *Nucleic Acids Res.* **14**, 10009–10026.
- BURKE, D. T., CARLE, G. F. and OLSON, M. V. (1987) Cloning of large segments of exogenous DNA into yeast using artificial-chromosome vectors. *Science* **236**, 806–812.
- BURKS, C., FICKETT, J. W., GOAD, W. B., KANEHISA, M., LEWITTER, F. I. *et al.* (1985) The GenBank nucleic acid sequence database. *Comput. Applic. Biosci.* **1**, 225–233.
- BURKS, C., CINKOSKY, M., GILNA, P., HAYDEN, J. E.-D., ABE, Y. *et al.* (1990) GenBank: Current status and future directions. *Meth. Enzymol.* **183**, 3–22.

- CAMERON, G. (1989) Data Library Group. *EMBL research reports 1990*, EMBL, Heidelberg.
- CAMERON, G., KAHN, P. and PHILIPSON, L. (1989) Journals and databanks. *Nature* **342**, 848.
- CEFIC (1990a) Bio-Informatics in Europe 1. Strategy for a European biotechnology information infrastructure. Confederation of European Chemical Industries, Bruxelles.
- CEFIC (1990b) Bio-Informatics in Europe 2. Strategy for a European biotechnology information infrastructure. Confederation of European Chemical Industries, Bruxelles.
- CHEE, M. S., BANKIER, A. T., BECK, S., BOHNI, R., BROWN, C. M. *et al.* (1990) Analysis of the protein coding content of the sequence of human cytomegalovirus strain AD169. *Curr. Top. Microbiol. Immunol.* **154**, 125–169.
- CHURCH, G. M. and KIEFFER-HIGGINS, S. (1988) Multiplex DNA sequencing. *Science* **240**, 185–188.
- CODD, E. F. (1970) A relational model of data for large shared data banks. *CACM* **13**.
- CONNELL, C., FUNG, S., HEINER, C., BRIDGHAM, J., CHAKERIAN, V. *et al.* (1987) Automated DNA sequence analysis. *BioTechniques* **5**, 342–348.
- COULSON, A., SULSTON, J., BRENNER, S. and KARN, J. (1986) Towards a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc. natn. Acad. Sci. U.S.A.* **83**, 7821–7825.
- DANIELS, D. L. and BLATTNER, F. R. (1987) Mapping using gene encyclopaedias. *Nature* **325**, 831.
- DAVISON, D. B. and CHAPPELEAR, J. E. (1990) The GenBank-Server at the University of Houston. *Nucleic Acids Res.* **18**, 1571–1572.
- DAYHOFF, M. O. (1966) *Atlas of Protein Sequences and Structure*. National Biomedical Research Foundation, Silver Springs.
- EDMAN, P. and BEGG, G. (1967) A protein sequenator. *Eur. J. Biochem.* **1**, 80–91.
- EDWARDS, A., VOSS, H., RICE, P., CIVITELLO, A., STEGEMANN, J. *et al.* (1990) Automated DNA sequencing of the HPRT locus. *Genomics* **6**, 593–608.
- EMBL DATA LIBRARY (1990a) EMBnet Workshop 1990 Uppsala. *EMBL Data Library Technical Document*, EMBL, Heidelberg.
- EMBL DATA LIBRARY (1990b) EMBnet: European Molecular Biology Network. *EMBL Data Library Technical Document*, EMBL, Heidelberg.
- EMBL DATA LIBRARY AND GENBANK (1990) The DDBJ/EMBL/GenBank Feature Table: Definition. Version 1.02.
- ERDMANN, V. A. and WOLTERS, J. (1987). The Berlin RNA databank. *Protein Seq. Data Anal.* **1**, 127.
- ETZOLD, T. (1990) SRS—a fast and easy to handle sequence retrieval system for the sequence collections from EMBL, GenBank and SwissProt. *Genes, Proteins & Computers. An International Conference on Biocomputing in Molecular Biology*, Chester.
- FUCHS, R., STOEHR, P., RICE, P., OMOND, R. and CAMERON, G. (1990) New services of the EMBL Data Library. *Nucleic Acids Res.* **18**, 4319–4323.
- GEORGE, D. G. and BARKER, W. C. (1990) A functional definition of molecular sequence data. *Computer Applications in Biosciences Workshop*, Martinsried, MPI für Biochemie.
- GHOSH, D. (1990) A relational database of transcription factors. *Nucleic Acids Res.* **18**, 1749–1756.
- GILBERT, W. (1991) Towards a paradigm shift in biology. *Nature* **349**, 99.
- GILNA, P. (1991) Data bank quality enhanced by curator program. *News from GenBank* **4**(1), 1–2.
- GOFFEAU, A. and VAN HOECK, F. (1990) Genome research in the European Community. *European HUGO meeting: Genome Analysis. From Sequence to Function*, Frankfurt, DEHEMA.
- GRAUSZ, D. (1991) A new European effort—Techniques that analyze complex genomes (TACpG). *Genomics* **9**, 560–562.
- GRAY, P. M., PATON, N. W., KEMP, G. J. and FOTHERGILL, J. E. (1990) An object-oriented database for protein structure analysis. *Protein Engng.* **3**, 235–243.
- GRIBSKOV, M. (1990) Molecular biology FTP and server list (16 December 1990), Rel. 1.1, published electronically on the BIOSCI bionews bulletin board. Available from the EMBL File Server.
- HAMM, G. H. and CAMERON, G. N. (1986) The EMBL data library. *Nucleic Acids Res.* **14**, 5–9.
- HOLLEY, R. W., APGAR, J., EVERETT, G. A., MADISON, J. T., MARGUISEE, M. *et al.* (1965) The base sequence of yeast alanine transfer RNA. *Science* **147**, 1462–1465.
- HORTON, M. and ADAMS, R. (1987) RFC 1036: Standard for interchange of USENET messages. Network Working Group, Internet Network Information Center.
- INNIS, M. A., MYAMBO, K. B., GELFAND, D. H. and BROW, M. A. D. (1988) DNA sequencing with Taq DNA polymerase and direct sequencing of polymerase chain reaction-amplified DNA. *Proc. natn. Acad. Sci. U.S.A.* **85**, 9436–9440.
- ISO 8824 (1987) Information processing systems—Open systems interconnection—Specification of Abstract Syntax Notation One (ASN.1). International Organization for Standardization, Geneva.
- ISO 8825 (1987) Information processing systems—Open systems interconnection—Specification of basic encoding rules for Abstract Syntax Notation One (ASN.1). International Organization for Standardization, Geneva.
- ISO 9660 Volume and file structure of CD-ROM for information interchange. International Organization for Standardization, Geneva.
- JORDAN, B. R. (1991) The French human genome program. *Genomics* **9**, 562–563.
- KAHN, P. and CAMERON, G. (1990) EMBL Data Library. *Meth. Enzymol.* **183**, 23–31.
- KANTOR, B. and LAPSLEY, P. (1986) RFC 977: networks news transfer protocol. A proposed standard for the stream-based transmission of news. Network Working Group, Internet Network Information Center.
- KNOELOCH, D. W., HILDEBRAND, C. E., MOYZIS, R. K., LONGMIRE, J. L. and SIROTKIN, K. M. *et al.* (1987) Robotics in the human genome project. *BioTechnology* **5**, 1284–1287.
- KOHARA, Y., AKIYAMA, K. and ISONO, K. (1987) The physical map of the whole *E. coli* chromosome: Application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50**, 495–508.
- KRÖGER, M., WAHL, R. and RICE, P. (1990) Compilation of DNA sequences of *Escherichia coli* (update 1990). *Nucleic Acids Res.* **18** (Suppl.), 2549–2552.
- LAWTON, J. R., MARTINEZ, F. A. and BURKS, C. (1989) Overview of the LiMB database. *Nucleic Acids Res.* **17**, 5885–5899.
- MADDOX, J. (1989a) Making authors toe the line. *Nature* **342**, 855.

- MADDOX, J. (1989b) Making good databanks better. *Nature* **341**, 277.
- MANIATIS, T., FRITSCH, E. F. and SAMBROOK, J. (1982) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory, New York.
- MARSHALL, E. (1990) Data sharing: a declining ethic? *Science* **248**, 952–957.
- MAXAM, A. M. and GILBERT, W. (1977) A new method for sequencing DNA. *Proc. natn. Acad. Sci. U.S.A.* **74**, 560–564.
- McKUSICK, V. A. (1989) The Human Genome Organisation: History, purposes, and membership. *Genomics* **5**, 385–387.
- MOORE, J. (1988) AUTHORIN—New GenBank submission software. *News from GenBank* **1**(5), 2.
- MOROWITZ, H. J. and SMITH, T. (1987) *Matrix of Biological Knowledge Workshop*. Santa Fe, Santa Fe Institute.
- NCBI (1990) GenInfo Backbone Database. Version 1.59, draft copy. National Center for Biotechnology Information, Bethesda.
- OLSON, M., HOOD, L., CANTOR, C. and BOTSTEIN, D. (1989) A common language for physical mapping of the human genome. *Science* **245**, 1434–1435.
- PABO, C. O. (1987) New generation databases for molecular biology. *Nature* **327**, 467.
- PEARSON, W. R. and LIPMAN, D. J. (1988) Improved tools for biological sequence comparison. *Proc. natn. Acad. Sci. U.S.A.* **85**, 2444–2448.
- PIR (1990) Document ZQSTAPE-1290. *PIR Technical Document*. National Biochemical Research Foundation (NBRF), Washington.
- PONGOR, S. (1988) Novel databases for molecular biology. *Nature* **332**, 24.
- RAWLINGS, C. J. (1988) Designing databases for molecular biology. *Nature* **334**, 477.
- ROBERTS, R. J. (1989) Benefits of databases. *Nature* **342**, 114.
- ROODE, D., LIEBSCHUTZ, R., MAULIK, S., FRIEDEMANN, T., BENTON, D. *et al.* (1988) New developments at BIONET. *Nucleic Acids Res.* **16**, 1857–1859.
- RUDD, K. E., MILLER, W., WERNER, C., OSTELL, J., TOLSTOSHEV, C. *et al.* (1990) Mapping sequenced *E. coli* genes by computer: software, strategies and examples. *Nucleic Acids Res.* **19**, 637–647.
- SAIKI, R. K., SCHARF, S., FALOONA, F., MULLIS, K. B., HORN, G. T. *et al.* (1985) Enzymatic amplification of β -globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science* **230**, 1350–1354.
- SANGER, F. (1956) The structure of insulin. In: *Currents in Biochemical Research* (ed. D. E. GREEN), Wiley Interscience, New York.
- SANGER, F., NICKLEN, S. and COULSON, A. R. (1977) DNA sequencing with chain-terminating inhibitors. *Proc. natn. Acad. Sci. U.S.A.* **74**, 5463–5468.
- SCHROEDER, J. L. and BLATTNER, F. R. (1982) Formal description of a DNA oriented computer language. *Nucleic Acids Res.* **10**, 69–84.
- SMITH, C. L., WARBURTON, P. W., GAAL, A. and CANTOR, C. R. (1986a) Analysis of genome organization and rearrangement by pulsed field gradient gel electrophoresis. In *Genetic engineering* (eds J. K. SETLOW and A. HOLLAENDER), Plenum, New York.
- SMITH, C. L., ECONOMO, J. G., SCHUTT, S., KLCO, S. and CANTOR, C. R. (1987) A physical map of the *Escherichia coli* genome. *Science* **236**, 1448–1453.
- SMITH, D. H., BRUTLAG, D., FRIEDLAND, P. and KEDES, L. H. (1986b) BIONET™: national computer resource for molecular biology. *Nucleic Acids Res.* **14**, 17–20.
- SMITH, R. H., GOTTESMAN, S., HOBBS, B., LEAR, E., KRISTOFFERSON, D. *et al.* (1991) A mechanism for maintaining an up-to-date GenBank® database via Usenet. *Comput. Appl. Biosci.* **7**, 111–112.
- SMITH, T. F. (1990) The history of the genetic sequence databases. *Genomics* **6**, 701–707.
- SMITH, T. F., GRUSKIN, K., TOLMAN, S. and FAULKNER, D. (1986) The molecular biology computer research resource. *Nucleic Acids Res.* **14**, 25–29.
- STADEN, R. (1980) A new method for storage and manipulation of DNA gel reading data. *Nucleic Acids Res.* **8**, 3673–3694.
- STOEHR, P. J. and OMOND, R. A. (1989) The EMBL network file server. *Nucleic Acids Res.* **17**, 6763–6764.
- U.S. CONGRESS, OFFICE OF TECHNOLOGY ASSESSMENT (1988) *Mapping our Genes—The Genome Projects: How Big, How Fast?* Government Printing Office, Washington, D.C.
- U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES AND U.S. DEPARTMENT OF ENERGY (1990) *Understanding our genetic inheritance. The U.S. Human Genome Project: The first five years FY 1991–1995*. National Technical Information Service, U.S. Department of Commerce, Springfield, VA.
- WALKER, R. (1990) DNA sequence policy editorial. *Nucleic Acids Res.* **18**, 6195.
- WATERMAN, M. S. (1990) Genomic sequence databases. *Genomics* **6**, 700–701.
- WATSON, J. and JORDAN, E. (1989) The Human Genome Program at the National Institutes of Health. *Genomics* **5**, 654–656.
- WHITE, T. J., ARNHEIM, N. and ERLICH, H. A. (1989) The polymerase chain reaction. *Trends in Genetics* **5**, 185–189.
- YUDIN, K. (1990) Database services available from e-mail servers. *News from GenBank* **3**(1), 1–2.