The EMBL data library

Peter J.Stoehr and Graham N.Cameron
European Molecular Biology Laboratory, Meyerhofstrasse 1, 6900 Heidelberg, FRG

INTRODUCTION

The EMBL Data Library was established in 1980 at the European Molecular Biology Laboratory. The EMBL Data Library collects, organises and distributes a database of nucleotide sequences and related descriptive information submitted directly from the scientific community and extracted from publications in scientific journals. It is also involved in the production and distribution of a protein sequence database (SWISS-PROT), and has begun to provide additional related data sets useful to molecular biologists.

The databases

The EMBL Nucleotide Sequence Database

The Nucleotide Sequence Database (1) was the original motivation for the establishment of the group , and continues to be the main endeavour of the Data Library. Since 1982 this work has been done in collaboration with GenBank® (2) (Los Alamos, N.M. and Mountain View, CA), and more recently the DNA Database of Japan (Mishima). Each of the three groups collects a portion of the total reported sequence data and exchanges it with the others on a regular basis.

Each database entry comprises a single contiguous sequence and its accompanying descriptive information (annotation). Different line types, each with their own two letter code, are used to make up an entry. A sample entry is shown in Figure 1. Each entry is uniquely identified within a release by its name (HSCA2 in Fig. 1), and across releases by its accession number list (Y00339 in Fig. 1). References to database entries should always cite the primary (first) accession number.

Release 25 of the database (November 1990) contained 41,580 sequence entries consisting of 52.9 million bases. There is a rapidly accelerating rate of growth, a trend which will undoubtedly continue. Clearly, the resources of the databases cannot increase at the same rate, and therefore we have revised and streamlined our data processing procedures. As part of an effort to achieve this we have installed the database into the ORACLE relational database management system.

The SWISS-PROT protein sequence database

The SWISS-PROT database (3), maintained collaboratively by the EMBL Data Library and Dr Amos Bairoch (University of Geneva), is a collection of amino acid sequences translated from the EMBL Nucleotide Sequence Database, adapted from the Protein Identification Resource collection (4) (PIR, Washington, D.C.), obtained from the literature and directly submitted by research groups. SWISS-PROT is fully annotated and particular efforts are made to eliminate duplicate sequences and to annotate the presence and extent of sequence domains. Recent review articles are used to periodically update the annotation of families or groups of proteins. SWISS-PROT is rich in cross-references to other databases. SWISS-PROT is similar in format to the Nucleotide Sequence Database and therefore the two collections can easily be used together.

Release 16 of the database (October 1990) contains 18,364 sequence entries comprising 5.9 million amino acids.

PROSITE pattern database

PROSITE (5) is a compilation of sites and patterns characteristic of specific biological functions found in protein sequences. It is maintained by Dr Amos Bairoch (University of Geneva). Some of the patterns have been published in the literature, but most have been developed by the database author. Cross-references are provided to instances of the patterns in the SWISS-PROT database. PROSITE is distributed with SWISS-PROT.

ENZYME database of EC nomenclature

ENZYME (6) is a database of characterised enzymes for which an Enzyme Commission (EC) number has been provided. It includes data such as the EC number, recommended and alternative names, catalytic activity and cofactors. Cross-references are provided to the SWISS-PROT database and also to the Mendelian Inheritance in Man database (MIM) (7) for human diseases associated with a deficiency of the enzyme. The main source of data is the recommendations of the Nomenclature Committee of the International Union of Biochemistry (8). ENZYME is distributed with SWISS-PROT.

ECD—E.coli map database

ECD (9) is a compilation of E.coli sequences in the EMBL/GenBank nucleotide sequence databases containing additional information on map locations. It is maintained by Dr Manfred Kroeger (University of Giessen). ECD is distributed with the EMBL Nucleotide Sequence Database. The CD-ROM distribution also includes query software for MS-DOS computer systems.

The Eukaryotic Promoter Database (EPD)

In 1988 we began to distribute a database of eukaryotic promoters (10), prepared by Philipp Bucher (presently at Stanford University, CA). This database contains detailed annotation of eukaryotic transcription start sites present in the Nucleotide Sequence Database and documented in the research literature. The database itself contains no sequences, but rather references to the sequences. It is distributed with the Nucleotide Sequence Database.

Restriction Enzyme Data (REBASE)
The database of restriction enzymes (11) provided by Dr. Rich Roberts (Cold Spring Harbor Laboratory) is distributed with all releases of the nucleotide sequence data.

Links between databases

Rapid growth in both the volume and complexity of sequence data makes it increasingly impractical for central data banks to maintain a pool of expertise capable of annotating all sequences; indeed such annotation may be interpretive work of a kind more appropriate to specialised research groups. Databases maintained remote from, but coordinated with, a centralised sequence collection, provide a model whereby the detailed biological annotation can be carried out at sites where the appropriate expertise is present. Figure 2 illustrates the current state of links between the EMBL nucleotide and SWISS-PROT protein sequence databases and other specialised collections. These links

```
ID   HSCA2      standard; RNA; PRI; 1523 BP.
XX
AC   Y00339;
XX
DT   19-SEP-1987 (Rel. 13; Last updated; Version 1)
DT   19-SEP-1987 (Rel. 13; Created)
XX
DE   Human mRNA for carbonic anhydrase II (EC 4.2.1.1)
XX
KW   CA2 gene; carbonic anhydrase II.
XX
OS   Homo sapiens (human)
OC   Eukaryota; Animalia; Metazoa; Chordata; Vertebrata; Mammalia;
OC   Theria; Eutheria; Primates; Haplorhini; Catarrhini; Hominidae.
XX
RN   [1] (bases 1-1523)
RA   David Hewett-Emmett Ph.D.;
RT   ;
RL   Submitted (01-JUL-1987) on tape to the EMBL Data Library by:
RL   David Hewett-Emmett Ph.D., University of Texas Health Science
RL   Center at Houston, Genetics Centers , Graduate School of Biomedical
RL   Sciences, P.O. Box 20334, Houston, TX 77225, U.S.A..
XX
RN   [2]
RA   Montgomery J.C., Venta P.J., Tashian R.E., Hewett-Emmett D.;
RT   "Nucleotide sequence of human liver carbonic anhydrase II cDNA";
RL   Nucleic Acids Res. 15:4687-4687(1987).
XX
DR   SWISS-PROT; P00918; CAH2$HUMAN.
XX
CC   *source: tissue=liver; library=lambda gt11;
CC   *source: clone=pHCAII38.3 and pHCII14.1;
CC   **map: 8q22;
XX
FH   Key             Location/Qualifiers
FH
FT   CDS             39..818
FT                   /product="carbonic anhydrase II"
FT   polyA_signal    1079..1084
FT                   /note="polyA signal"
FT   polyA_signal    1266..1271
FT                   /note="alternative polyA signal"
FT   polyA_signal    1476..1481
FT                   /note="alternative polyA signal"
FT   polyA_signal    1506..1511
FT                   /note="alternative polyA signal"
XX
SQ   Sequence  1523 BP;  456 A; 309 C; 319 G; 439 T; 0 other;
     gtgccgattc ctgccctgcc ccgaccgcca gcgcgaccat gtcccatcac tgggggtacg
     gcaaacacaa cggacctgag cactggcata aggacttccc cattgccaag ggagagcgcc
     agtcccctgt tgacatcgac actcatacag ccaagtatga cccttccctg aagcccctgt
     acagattgat tcagtttcac tttcactggg gttcacttga tggacaaggt tcagagcata

        .          .          .          .          .          .
        .          .          .          .          .          .

     atatatttat agcaaagtta tcttaaatat gaattctgtt gtaatttaat gacttttgaa
     ttacagagat ataaatgaag tattatctgt aaaaattgtt ataattagag ttgtgataca
     gagtatattt ccattcagac aatatatcat aacttaataa atattgtatt ttagatatat
     tctctaataa aattcagaat tct
//
```

Figure 1. A sample entry from the EMBL Nucleotide Sequence Database

are manifested in database entries as pointers to stable objects in other databases, for example to primary accession numbers in the sequence databases.

Nucleotide Sequence Acquisition

The staff of the nucleotide sequence databases have devoted a great amount of effort to developing systems which encourage researchers to submit their newly-determined sequences and related data directly to the databases, preferably in computer-readable form.

Direct submission is important for many reasons:

- Abstracting the information from the research literature is labour intensive.
- Entries can appear in the database much sooner if we get the information from the authors early in the publication process.
- Machine readable submissions reduce the chance of us introducing errors.
- Authors can bring far more expertise to bear on annotating their own data than the database staff can.
- Interpretive annotation is more a research than a database activity.
- The scale of present day sequencing is reaching the point where journals are finding it inappropriate to print the actual sequences. If no mechanism to ensure their deposition in the databases is in place, research papers will be published whose underlying data are unavailable to the research community.
- As an increasing number of journals publish sequence related papers without printing the sequences, scanning the literature to locate sequence data will no longer work.

Progress with direct submissions has been good, with almost 80% of the new data coming as submissions in the latter half of 1990. Several journals now not only request that authors submit sequence data to the database but require it as a condition for publication. While these developments are encouraging, the accelerating rate at which sequence data is generated ensures that the work of entering and annotating the remaining (non-submitted) data continues to be an enormous task. The goal of ensuring more direct submissions therefore remains a high priority.

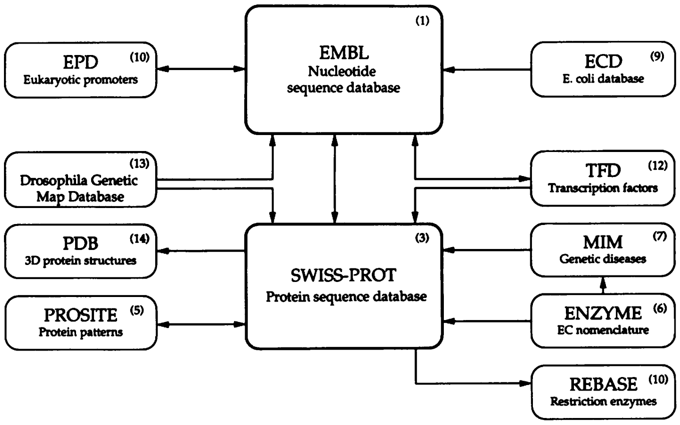

Figure 2. Current cross-referencing between databases

How to submit nucleotide sequence data to the EMBL Data Library

Researchers who intend to submit data to any of the sequence databases should get a copy of a Sequence Data Submission Form, which solicits all the information needed for a nucleotide or protein sequence entry and provides instructions on how to submit the data. The form exists both in a paper version and as a computer-readable text file which can be completed using a text editor. Many molecular biology journals distribute the paper version to authors of manuscripts reporting sequence data, and a few journals publish it periodically.(15,16) The computer-readable version of the form is distributed with all releases of the EMBL and GenBank® databases and can also be obtained via computer network (using the EMBL File Server, see below). Alternatively, either of these versions can be obtained by contacting the Data Library in any of the ways listed at the end of this document. A data submission should include the sequence data in computer-readable form (computer network mail, magnetic tape or MS-DOS or Macintosh floppy diskette) and a completed data submission form for each submitted sequence. Data can be sent to the Data Library via computer network, telefax or normal post.

Complete submissions are processed within a few days and the authors are given accession numbers which are permanent references to the data and a means of citation. When submissions are incomplete, authors may be contacted for further information before an accession number is assigned. Submitters are given the option of withholding data from public availability until they are published.

Data Distribution

The main way of distributing copies of the entire database is by mailing of CD-ROMs and magnetic tapes every three months. Much of this distribution is done by the Data Library and some is done by secondary distributors, such as groups which supply the data along with sequence analysis software. Further information about subscriptions can be obtained by contacting us at the address given at the end of this document.

CD-ROM is the preferred medium because it represents a cheap way to store large quantities data, and because the devices required to read it are within financial reach of the typical personal computer user. The CD-ROM includes MSDOS software for query and retrieval of data in the sequence databases and the databases are also provided on the CD-ROM in a format compatible with widely-used sequence homology search software (eg FASTA by Lipman and Pearson (17) for MS-DOS and Macintosh systems)

Network Access

The rapid pace of research in molecular biology has generated a requirement for better and more rapid access to the databases than that provided by quarterly releases. As part of an attempt to meet this need, EMBL set up in early 1988 a file server which enables researchers world-wide to retrieve entries from the major databases available at EMBL via computer network (18). Nucleotide and protein sequence data is available over the file server as soon as Data Library staff have completed the entry. This is particularly attractive in combination with the data submission policy, since it enables readers of these journals to access the sequence data in computer-readable form as soon as the issue containing them appears. The file server facility has been steadily extended and now includes access to many other data collections, free molecular biology software, and a sequence database homology search service (19,20). The file server can be used by anyone with access to the BITNET network or to any other network which has a gateway into BITNET (e.g., JANET in the UK, Internet etc.). It is provided free of charge, though users may have to meet some or all of the communication costs, depending on the accounting system of their local computer service.

Use of the facility is simple and involves sending file server commands, one per line, in a standard electronic mail to the address NETSERV@EMBL.BITNET. The most important file server command, to get users started, is HELP. If the file server receives this command, it will return a help file to the sender, explaining in some detail how to use the facility.

EMBnet

Another way in which EMBL is attempting to increase the availability and usefulness of the various databases is by establishing a European molecular biology network (EMBNet) consisting of centres of expertise in molecular biology and biocomputing in Europe.

In 1988 a trial phase of the EMBnet project was initiated with four centres, and since then a gradual expansion has led to the involvement of at least one node in 12 western European countries. Progress so far includes the establishment of network connections and the implementation of systems to update their copy of the nucleotide sequence database on a daily basis. These national centres make these data, along with analytical software, available to researchers within their countries and offer training and support in their use. Other network services are being investigated collaboratively (eg conferencing systems, remote access to specialised facilities) and original implementation is being broadened to embrace other types of nodes (database providers, service nodes, research nodes, user nodes etc.).

How to contact the EMBL Data Library

Network:	datasubs@embl.bitnet (for data submissions); datalib@embl.bitnet (for questions requiring a personal response)
Postal address:	Data Submissions, EMBL Data Library, Postfach 10.2209, 6900 Heidelberg, Federal Republic of Germany
Telephone:	+49-6221-387258
Telefax:	+49-6221-387519 or 387306
Telex:	461613 (embl d)

REFERENCES

1. Hamm, G. and Cameron, G. (1986) *Nucl. Acids Res.*,**14**, 5-10.
2. Burks, C., Fickett, J.W., Goad, W.B., Kanehisha, M., Lewitter, F.I., Rindone, W.P., Swindell, C.D., Tung, C.S. and Bilofsky, H.S. (1985) *C-ABIOS* ,1, 225-233.
3. Bairoch, A. and Boeckmann, B. (1991) *Nucl. Acids Res.*, **19**, 2247-2249..
4. George, D.G., Barker, W.C. and Hunt, L.T. (1986) *Nucl. Acids Res.*, **14**, 11-14.
5. Bairoch,A. (1991) *Nucl. Acids Res.* **19**, 2241-2245.
6. Bairoch, A. (1990) University of Geneva, Geneva.
7. McKusick, V. (1990) *Mendelian Inheritance in Man*, John Hopkins University Press, Baltimore.
8. Enzyme Nomenclature, NC-IUB, Academic Press, New York (1984).
9. Kröger,M., Wahl,R. and Rice,P. (1991) *Nucl. Acids Res.*, **19**, 2023-2043.
10. Bucher,P. and Trifonov,E.N. (1986) *Nucl. Acids Res.*, **14**, 10009-10026.
11. Roberts,R.J. (1985) *Nucl. Acids Res.*, **13**, r165-r200.
12. Ghosh, D. (1990) *Nucl. Acids Res.*, **18**, 1749-1756.
13. Ashburner, M. (1990) University of Cambridge, Cambridge.
14. Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535-542.
15. Kahn, P. and Hazledine, D. (1988) *Nucl. Acids Res.* **16** (10), i.
16. Kahn, P., Hazledine, D. and Cameron G. (1988) *Plant Molecular Biology* 11, 541.
17. Pearson, W.R. and Lipman, D,J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444-2448
18. Stoehr, P. and Omond, R. (1989) *Nucl. Acids Res.* **17** (16), 6763-6764.
19. Fuchs, R., Stoehr, P., Rice, P., Omond, R. and Cameron, G. (1990) *Nucl. Acids Res.* **18** (15), 4319-4323.
20. Fuchs,R. (1990) *CABIOS*, **6**, 120-121.