Nucleic Acids Research, Vol. 20, Supplement 2071–2074

The EMBL Data Library

Desmond G.Higgins, Rainer Fuchs, Peter J.Stoehr and Graham N.Cameron
European Molecular Biology Laboratory, Meyerhofstrasse 1, 6900 Heidelberg, Germany

INTRODUCTION

The EMBL Data Library is part of the European Molecular Biology Laboratory in Heidelberg, Germany. It was established in 1980 and its principal role is to maintain and distribute a database of nucleotide sequences (the EMBL Nucleotide Sequence Database). It is also involved in maintaining other biological databases such as the protein sequence database SWISS-PROT and distributes other databases of interest to molecular biologists.

The Data Library acts also as a major distributor of other databases of interest to molecular biologists. These are not maintained by the Data Library but are distributed quarterly on tape and CD-ROM or are made available from the EMBL file server (7). Table 1 shows a list of these databases and indicates the distribution mechanism for each (magnetic tape, CD-ROM or from the file server).

THE DATABASES

The EMBL Nucleotide Sequence Database

The Nucleotide Sequence Database (1) is the main activity of the group. This work is done in collaboration with GenBank® (2) (Los Alamos and Mountain View, U.S.A.) and 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 daily basis. The exponential growth in size of the database continues and the latest release (release 30; February 1992) reports 83.6 million bases from 63,378 entries. The database approximately doubles in size every 18 months. Improved data handling methods and compulsory data submission policies on the part of most of the major molecular biology journals make it possible to deal with the increasing volumes of sequence data that are being generated. This year, the first sequences were incorporated from large scale genome sequencing projects. In the future, it is expected that this will be the major source of sequence data. The complete database is available every three months on magnetic tape or compact disc (CD-ROM). New sequences are also available by computer networks as soon as they have been processed by the Data Library staff.

The SWISS-PROT protein sequence database

The SWISS-PROT database (3) is maintained collaboratively by the EMBL Data Library and Dr. Amos Bairoch of the University of Geneva. It is distributed in the same file format as the Nucleotide Sequence Database, with which it is fully cross-referenced. Release 20 of SWISS-PROT (December 1991) contained 7.5 million amino acids from 22,654 sequences. The data in SWISS-PROT are derived from translations of DNA sequences in the EMBL database, adapted from the Protein Identification Resource collection (4) (PIR, Washington, D.C.), extracted from the literature and directly submitted by researchers. Its strengths are the quality and consistency of its annotation and the cross references to other databases, especially PROSITE (5), PDB (6) and the EMBL nucleotide database. SWISS-PROT is distributed on magnetic tape and CD-ROM every 3 months, and new entries can be retrieved between releases from the EMBL file server.

DATA DISTRIBUTION

Every 3 months, the databases available from EMBL are distributed on compact disc (CD-ROM) and magnetic tape to users around the world. CD-ROM has become the preferred medium because of its low cost and convenience and because the disks can be read by users with access to personal computers. However, the rapid growth of the databases makes immediate access to latest data between releases increasingly important, and is provided by EMBL through the EMBL file server and EMBnet.

CD-ROM

The CD-ROM is written in the ISO 9660 standard format and can therefore be used by a wide range of computer systems. The main contents of the EMBL CD-ROM are the nucleotide and protein sequence databases as distributed on magnetic tape. In a new collaboration with Oxford University Press the CD-ROM now also contains most of the databases described in the annual Sequence Supplement of Nucleic Acids Research (see Table 1).

Query software for MS-DOS is provided for retrieving sequences by keyword, author name, species and free text, among other criteria. Well-documented index files for most database fields have been added recently to allow software developers to create their own retrieval programs and facilitate access to the raw data. Based on these index files query software for Macintosh systems is being developed and it is planned to begin distributing during 1992. Copies of the main databases are included in a format that can be used by the widely available FASTA package (24) with MS-DOS or Macintosh systems. MS-DOS software and index files (25) for quickly screening nucleotide sequences for strong similarity to database sequences are also provided.

Network Access

To provide users with immediate access to the sequence databases between releases the EMBL file server (7) was set up in 1988. Nucleotide and protein sequence data are available as soon as the Data Library staff have completed an entry. The service operates by electronic mail and is available free of charge to users around the world. All one needs to make use of it is an electronic mail connection. Further information can be obtained by sending

Table 1. List of the databases distributed by EMBL and the mechanism of distribution in each case.

Database		Reference	Distribution		
			Mag.Tape	CD-ROM	File Server
EMBL	nucleotide sequence database	(1)	●	●	●
SWISS-PROT	protein sequence database	(2)	●	●	●
ENZYME	database of EC nomenclature	(8)	●	●	●
ECD	E. coli map database	(9)	●	●	●
EPD	Eukaryotic promoter database	(10)	●	●	●
FLYBASE	Drosophila genetic map database	(11)	●	●	●
PROSITE	protein pattern database	(5)	●	●	●
REBASE	Restriction enzyme database	(12)	●	●	●
TFD	Transcription factor database	(13)		●	●
TRNA	tRNA sequences	(14)		●	●
RRNA	Small subunit rRNA sequences	(15)		●	●
METHYL	Site specific methylation	(16)		●	●
SMALLRNA	Small RNA sequences	(17)		●	●
HAEMB	Haemophilia B database of mutations	(18)		●	●
BERLIN	5S rRNA sequences	(19)		●	●
CUTG	Codon usage tabulated from GenBank	(20)		●	●
LIMB	Listing of mol. biology databases	(21)			●
ALU	ALU sequences and alignments	(22)			●
SEQANALREF	Articles dealing with sequence analysis(23)				●
PDB	Brookhaven protein structures database	(6)			●

a mail message to the Internet address Netserv@EMBL-Heidelberg.DE, with the word HELP in the body of the message. A full set of instructions will be returned automatically by e-mail.

The main function of the file server is to provide access to sequence entries. To get a sequence, all one needs to know is the entry name or accession number. E.g. to get the sequence with entry name ECFUCOSE and accession number X15025, one would send the command GET NUC:X15025 or GET NUC:ECFUCOSE to the file server. Several index files are updated daily to assist users in finding accession numbers of new sequences of interest to them. Further, all new sequences that are processed by the Data Library since the last issue of Nucleic Acids research are listed at the back of this journal.

The file server also provides access to the databases indicated in Table 1, a large collection of free molecular biology software and a network-based sequence database homology search service (28, 29). The software collection now has over 170 programs for the most commonly used computer systems (MS-DOS, Macintosh, VAX/VMS and UNIX).

EMBnet

The European Molecular Biology Network (EMBnet) was initiated in 1988 as an attempt to increase the availability and usefulness of the various databases within Europe. It consists currently of a total of 20 nodes from 14 countries and these are listed in Table 2. Progress so far includes the establishment of network connections and systems to update copies of the nucleotide and protein databases in each node on a daily basis. 14 of the nodes act as national centres and make the databases and analytical software available to users within their countries, along with training and user support. Other network services such as conferencing systems and remote access to specialised facilities are being investigated collaboratively and the original implementation is being broadened to include other types of node such as database providers, service nodes, research nodes and user nodes.

Table 2: Participants in the EMBnet project (as of February 1992).

National EMBnet nodes	
IMP, Vienna	Austria
BioBase, Aarhus	Denmark
CSC, Espoo	Finland
CITI2, Paris	France
DKFZ, Heidelberg	Germany
IMBB, Crete	Greece
Weizmann Institute, Rehovot	Israel
CNR, Bari	Italy
CAOS/CAMM Centre, Nijmegen	Netherlands
Biotechnology Centre of Oslo	Norway
CNB, Madrid	Spain
Biomedical Centre, Uppsala	Sweden
Biozentrum, Basel	Switzerland
SERC Daresbury Lab., Warrington	United Kingdom
Other nodes	
EMBL	Heidelberg
European Patent Office	The Hague, Netherlands
Hoffman-La Roche	Basel, Switzerland
ICGEB	Trieste, Italy
MIPS	Martinsried, Germany
UK, HGMP	Harrow, United Kingdom

DATA ACQUISITION

In the past, the primary source of sequence data was the molecular biology literature. Abstracting such information from journals is time consuming, error-prone and inefficient. Starting with Nucleic Acids Research in 1988, some journals made it a condition of publication that authors of sequence containing papers submit their data directly to the databases. There are several advantages to this approach. Firstly, the annotation of each sequence entry can be largely carried out by the researchers involved, who know more about the data than anyone else. Secondly, the time-lag between publication of a sequence and its appearance in the database can be completely eliminated.

Sequences can be made available by computer network within a week of direct submission by authors. Finally, journals can decide to stop printing large sequences, because the data are available in electronic form in the databases. Over 80% of all of the data entered into the Nucleotide Sequence Database now come as direct submissions from the authors. Because of the ever increasing rate of creation of the data, however, the work of abstracting the remaining data from the literature remains an enormous task.

How to submit data

Researchers who intend to submit data to any of the sequence databases should either 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 or use the AUTHORIN program, described below. The form exists both in a paper version and as a computer-readable text file which can be completed using a text editor or word-processor. Many molecular biology journals distribute the paper version to authors of manuscripts reporting sequence data, and a few journals publish it periodically (26, 27). For example, it is printed regularly in this journal. The computer readable version of the form is distributed with all releases of the EMBL and GenBank databases and can be obtained via computer network using the EMBL file server. Alternatively, either of these versions can be obtained by contacting the Data Library in any of the ways listed at the end of this paper. A data submission should include the sequence data in computer-readable form (computer network mail, magnetic tape, MS-DOS or Macintosh diskette) and a completed data submission form for each submitted sequence. Data can be sent to the Data Library via computer network, telefax or by normal post.

The AUTHORIN program allows users of MS-DOS or Macintosh computers to create data submission forms interactively. The software is simple to use and results in machine readable files which can be automatically processed in the Data Library. This is increasingly becoming the preferred means of data submission.

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 from Genome projects

One of the most exciting developments in 1991 was the incorporation into the databases of the first data from the international nematode genome project (30) and the European yeast chromosome III sequencing project (31). This is the start of what will become the major source of data for the sequence databases in the future. Automatic procedures to allow the direct submission and incorporation of such data were developed and applied successfully. There are now three cosmids totalling over 120 thousand bases of DNA from the nematode project: cosmid B0303 (accession number M77697), cosmid ZK637 (acc. no. Z11115) and cosmid ZK643 (acc. no. Z11126). The complete sequence of yeast chromosome III is being assembled at the Martinsried Institute of Protein Sequences (MIPS) and will be put into the public database in the near future. Similar procedures for direct submissions of large volumes of data are currently being developed in collaboration with other European sequencing initiatives.

Sequence data in patents

Arrangements have been made with the European Patent Office (EPO) to make public the sequence data contained within patent applications. During 1992, the Data Library will start processing the backlog of European patent data, approximately 2000 applications containing 10,000 sequences. Patent applications with first priority in the USA are being processed by the U.S. National Center for Biotechnology Information (NCBI) with whom data will be exchanged. Cross-references to the patent literature will be provided in sequence database entries. This is an important source of nucleotide and protein sequence data which has been neglected up to now by the public sequence data banks.

Data exchange with other sequence data banks

New nucleotide database entries are exchanged on a daily basis between EMBL, GenBank and DDBJ. The Data Library has invested considerable effort in eliminating existing problems stemming from data exchange and data reformatting. At present, the EMBL database covers more than 99% of the data produced by GenBank and DDBJ, clearly showing that there is a 'functional equivalence' of the three collaborating databases.

In 1992, the American National Center for Biotechnology information (NCBI) will take over responsibility for the GenBank data bank and will begin the public distribution of a new database, the GenInfo backbone database, aiming at a complete coverage of sequence information published in scientific journals. To guarantee the best possible service to the user community, NCBI and the EMBL Data Library have already started into a close collaboration, a first result being the work on mechanisms for data exchange between these groups based on the jointly refined definitions of molecular biological data types using Abstract Syntax Notation 1 (ASN.1).

How to contact the EMBL Data Library

Network:	Datasubs@EMBL-Heidelberg.DE (for data submissions) Datalib@EMBL-Heidelberg.DE (for questions requiring a personal response) Netserv@EMBL-Heidelberg.DE (the automatic file server)
Postal address:	Data Submissions, EMBL Data Library, Postfach 10.2209, W-6900 Heidelberg, 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) *CABIOS*, **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. 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.

7. Stoehr, P. and Omond, R. (1989) *Nucl. Acids Res.* , **17**, 6763–6764.
8. Bairoch, A. (1990) University of Geneva, Geneva.
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. Ashburner, M. (1990) University of Cambridge, Cambridge.
12. Roberts,R.J. (1985) *Nucl. Acids Res.*, **13**, r165–r200.
13. Ghosh, D. (1990) *Nucl. Acids Res.*, **18**, 1749–1756.
14. Sprinzl, M, Dank, N., Nock, S. and Schön, A. (1991) *Nucl. Acids Res..*, **19,** 2127–2171.
15. Neefs, J.-M., Van de Peer, Y., De Rijk, P. Goris, A. and De Wachter, R. (1991) *Nucl. Acids Res..*, **19,** 1987–2015.
16. Nelson, M. and McClelland, M. (1991) *Nucl. Acids Res..*, **19,** 2045–2071.
17. Gupta, S. and Reddy, R. (1991) *Nucl. Acids Res..*, **19,** 2073–2075.
18. Giannelli, F., Green, P.M., High, K.A., Sommer, S., Lillicrap, D.P., Ludwig, M., Olek, K., Reitsma, P.H., Goosens, M., Yoshioka, A. and Brownlee, G.G. (1991) *Nucl. Acids Res.*, **19,** 2193–2220.
19. Specht, T., Wolters, J. and Erdmann, V.A. (1991) *Nucl. Acids Res..*, **19,** 2189–2191.
20. Wada, K., Wada, Y., Doi, H., Ishibashi, F., Gojobori, T. and Ikemura, T. (1991) *Nucl. Acids Res.*, **19,** 1981–1986.
21. Lawton, J.R., Martinez, F.A. and Burks, C. (1989)*Nucl. Acids Res.*, **17**, 5885–5899.
22. Jurka, J. and Smith, T. (1988) *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 74775–4778.
23. Bairoch, A. (1991) University of Geneva, Geneva.
24. Pearson, W.R. and Lipman, D,J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448
25. Higgins, D.G. and Stoehr, P.J. (1992) *CABIOS*, **8**, in press.
26. Kahn, P. and Hazledine, D. (1988) *Nucl. Acids Res.* **16** (10), i.
27. Kahn, P., Hazledine, D. and Cameron G. (1988) *Plant Molecular Biology* 11, 541.
28. Fuchs, R., Stoehr, P., Rice, P., Omond, R. and Cameron, G. (1990) *Nucl. Acids Res.* **18** (15), 4319–4323.
29. Fuchs, R. (1990) *CABIOS*, **6**, 120–121.
30. Coulson, A., Sulston, J., Brenner, S. and Karn, J. (1986)*Proc. Natl. Acad. Sci. U.S.A.*, **83**, 7821–7825.
31. Goffeau, A. and Van Hoeck, F. (1990) European HUGO meeting: Genome Analysis. From Sequence to Function.