
The EMBL data library

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

INTRODUCTION

The principal role of the EMBL Data Library, since its inception in 1980, has been to maintain and distribute a database of nucleotide sequences (the EMBL Nucleotide Sequence Database). It also supports and maintains the protein sequence database SWISS-PROT and distributes other databases of interest to molecular biologists. The past twelve months have seen developments in a large number of our activities; such as the processing of the first sequences from the patent literature, the incorporation of data from the U.S. National Center for Biotechnology Information (NCBI) journal scanning activity, the introduction of new network services (anonymous FTP, Gopher, and a new fast sequence search facility) and the establishment of the European Bioinformatics Institute (EBI).

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[®], USA (2) and the DNA Database of Japan (DDBJ). 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 explosive growth of the database continues and the latest release (Release 34; March 1993) reports just under 130 million bases from 105,340 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. Close collaboration with genome project databases has resulted in refined procedures for automatic inclusion of genome sequence data into our database. Genome projects are now one of the major sources of sequence data, comprising 20% of the entries in the database.

The complete database is available every three months on compact disc (CD-ROM), magnetic tape and over computer networks. Additionally, new sequences can be retrieved between releases via computer networks as soon as they have been processed.

The SWISS-PROT protein sequence database

The SWISS-PROT database (3) is maintained collaboratively by the EMBL Data Library and 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 25 of SWISS-PROT (April 1993) contains 10.2 million amino acids from 29,955 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 the EMBL nucleotide sequence database, PROSITE (5), PDB (6). SWISS-PROT is distributed on CD-ROM and magnetic tape every 3 months, and new entries can be retrieved between releases via our network servers.

Other databases

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 CD-ROM and magnetic tape or are made available from the EMBL network servers. Table 1 shows a list of these databases and indicates the distribution mechanism for each (magnetic tape, CD-ROM or network servers).

DATA ACQUISITION

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 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. Today, approximately 90% of all data are directly submitted. The work of abstracting the remaining 10% from the literature remains a time-consuming and error-prone task. Additionally, these data tend to be less complete owing to space restrictions on printed publication.

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 or use the AUTHORIN program, described below. A 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 e-mail file server (36). Many molecular biology journals distribute a paper version to authors of manuscripts reporting sequence data, and a few

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

Database	Distribution Mag.Tape	CD-ROM	Network Servers
EMBL nucleotide sequence database (1)	•	•	•
SWISS-PROT protein sequence database (3)	•	•	•
ENZYME database of EC nomenclature (7)	•	•	•
ECD E.coli map database (8)	•	•	•
EPD eukaryotic promoter database (9)	•	•	•
PROSITE protein pattern database (5)	•	•	•
REBASE restriction enzyme database (10)	•	•	•
FLYBASE Drosophila genetic map database (11)		•	•
TFD transcription factor database (12)		•	•
TRNA database of tRNA sequences (13)		•	•
RRNA small subunit rRNA sequences (14)		•	•
Haemophilia B database of mutations (15)		•	•
BERLIN database of 5S rRNA sequences (16)		•	•
SEQANALREF sequence analysis bibliography (17)		•	•
LIMB listing of mol. biology databases (18)		•	•
PKCDD protein kinase catalytic domain database (19)		•	•
CpG Islands database (20)		•	•
SRP signal recognition particle database (21)		•	•
Kabat database of proteins of immunological interest (22)		•	•
RLDB reference library database (23)		•	•
Translational termination signal database (24)		•	•
Site specific methylation database (25)		•	•
Small RNA sequence database (26)		•	•
CUTG codon usage tabulated from GenBank (27)		•	•
Protein blocks database (28)		•	•
HSSP homology-derived protein structures (29)		•	•
DSSP definition of protein secondary structure (30)			•
PDB Brookhaven protein structures database (6)			•
ALU sequences and alignments (31)			•
Functional analysis bibliography (32)			•
3D-ALI 3D alignment database (33)			•
HLA class I and II sequence database (34)			•
RELIB restriction enzyme library (35)			•

journals, including *Nucleic Acids Research*, publish it periodically (37, 38). The form solicits all the information needed for a nucleotide or protein sequence entry and provides instructions on how to submit the data.

The AUTHORIN program allows users of MS-DOS or Macintosh computers to prepare data submissions interactively and the results can be automatically processed by the Data Library. This is our preferred means of data submission. Copies of AUTHORIN can be obtained on floppy diskettes from NCBI (GenBank/NCBI, 8N-803, Bldg 38A, Bethesda, MD 20894) or electronically from the EMBL network servers (see below).

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. Submitters are given the option of withholding data from public availability until they are published. In these cases it is helpful to us if the submitter supplies details of subsequent publication.

The EMBL Data Library encourages users to send corrections and updates to the publicly available data. Additional information to previously submitted data is welcomed. Please use the update address given at the end of this paper.

The changing face of nucleotide sequence data

The past year has seen the inclusion of data from an increasing variety of sources. Historically, the data came from scientists working in individual laboratories, who were planning publication of their sequences as part of their research. Today, however, much of our data comes from other sources and the type and

quality is often very different. Currently, we provide the database as a set of 14 files based on approximate taxonomic divisions. It is clear that mechanisms to allow building of more sophisticated views of the underlying data are required.

Genome Projects

In continuation of existing successful collaborations with genome projects in Europe, this past year saw the inclusion of the first data from Genexpress, Munich, which is part of a sequencing effort to collect all expressed human sequences (complementary DNA clones, cDNAs), and from the French Arabidopsis project, which similarly aims to sequence all expressed sequences (cDNAs) from Arabidopsis. The full list of our genome project collaborations is given below:

- *C.elegans* nematode project
- *S.cerevisiae* yeast project
- Genexpress Genethon
- Genexpress, Munich
- French Arabidopsis cDNA project
- UK Human Genome Mapping Project

Our policy is to release genome project data as early as possible to the user community and therefore automatic procedures for sequence incorporation have been developed and successfully implemented. Data from these projects is of a different quality, often in the form of partial sequences generated from randomly selected cDNA clones, known as expressed sequence tags (ESTs,

Table 2. Sites maintaining daily updated copies of EMBL Nucleotide Sequence Database (May 1993).

National nodes	Contact Addresses
Belgium	Belgian EMBnet Node, Computing Center, C.P.300, 50 Av. Franklin Roosevelt, 1050 Brussels, Belgium e-mail contact: rherzog@ulb.ac.be
Brazil	Brazilian National Bioinformatics Resource, EMBRAPA/CENARGEN, SAIN-Parque Rural CP102372, CEP 70770, Brazil e-mail contact: neshich@cenargen.embrapa.br
Denmark	BioBase, Ole Worms alle, Building 170, Aarhus Universitet, DK-8000 Aarhus C, Denmark e-mail contact: hum@biobase.aau.dk
France	BISANCE, Laboratoire de Biochimie, Ecole Polytechnique, 91128 Palaiseau Cedex, France e-mail contact: dessens@coli.polytechnique.fr
Finland	Centre for Scientific Computing, PO Box 405, SF 02101 Espoo e-mail contact: harper@convex.csc.fi
Germany	DKFZ, Im Neuenheimer Feld 280, W 6900 Heidelberg, Germany e-mail contact: dok419@genius.embnnet.dkfz-heidelberg.de
Greece	IMBB, PO Box 1527, Heraklion 71110, Crete, Greece e-mail contact: savakis@myia.imbb.forth.gr
Israel	Biological Computing Division, Weizmann Institute of Science, Rehovot 76100, Israel e-mail contact: lsestern@weizmann.bitnet
Italy	University of Bari, Dipartimento di Biochimica e Biologica Molecolare, Traversa 200 Re David 4, 170125 Bari, Italy email contact: attimonelli@mvx36.csata.it
—	ICGEB, Area Science Park, 1-34012 Trieste, Italy e-mail contact: pongor@genes.icgeb.trieste.it
Netherlands	CAOS/CAMM Center, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands e-mail contact: noordik@caos.caos.kun.nl
—	European Patent Office, P.B. 5818, Patentlaan 2, 2280 HV Rijswijk (ZH), The Netherlands
Norway	Norwegian EMBnet Node, Biotechnology Center of Oslo, Gaustadaleen 21, N-0371 Oslo, Norway e-mail contact: rodrigol@biomed.uio.no
Spain	Centro nacional de biotecnologia, CSIC, Universidad Autonoma de Madrid, 28049 Madrid, Spain e-mail contact: carazo@cnb.uam.es
Sweden	Biomedical Centre, Box 570, S 751 23, Uppsala, Sweden e-mail contact: gad@perrier.embnnet.se
Switzerland	Biocomputing, Biozentrum der Universit Basel, Klingelbergstrasse 70, CH 4056 Basel, Switzerland e-mail contact: embnet@comp.bioz.unibas.ch
—	Hoffman-La Roche, CH 4002 Basel, Switzerland e-mail contact: doran@embl-heidelberg.de
United Kingdom	SEQNET, SERC Daresbury Lab., Warrington WA4 4AD, UK e-mail contact: bleasby@dl.ac.uk

39). Only occasionally is more biological information available. For this reason, 'EST' data, which are flagged by the use of either of the keywords 'expressed sequence tag' or 'transcribed sequence fragment', will appear in a new database division called EST as from Release 35 (June 1993).

Sequences from patent literature

The EMBL Data Library has recently begun capturing data from the patent literature under contract from the European Patent Office (EPO). Since late 1992, we have been processing this backlog of European patent data, consisting of approximately 2000 patent applications containing about 10,000 nucleotide sequences. Patent applications with first priority in the USA are being processed by NCBI with whom data will be exchanged. References to the patent literature are provided in these sequence database entries. This is an important source of nucleotide and protein sequence information which has been neglected up to now by the public sequence data banks.

NCBI journal-scanning activity

In 1992, NCBI started a project aimed at extracting sequence information as published in scientific journals. The scope of the task includes scanning of journals which only irregularly publish sequence data. This is a useful supplement while direct submission policies are not 100% effective. Therefore we have begun to include data from this new source, taking care to avoid duplication of directly submitted data.

DATA DISTRIBUTION

Every 3 months, the databases available from EMBL are distributed on 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. Immediate access to the latest data between releases is increasingly important, and is provided by the EMBL network servers.

CD-ROM

The CD-ROM is distributed as a two volume set of compact discs written in the international ISO 9660 standard format which enables it to be used on a wide range of computer systems. The main contents are the nucleotide and protein sequence databases. In a collaboration with Oxford University Press, the CD-ROM now also contains most of the databases described in the annual Nucleic Acids Research Supplement (see Table 1).

Query software for MS-DOS and Macintosh (40) 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. Copies of the main databases are included in a format that can be used by the widely available FASTA package (41) for sequence similarity searches. MS-DOS software and index files (42) for quickly screening nucleotide sequences for strong similarity to database sequences are also provided.

NETWORK SERVICES

In response to the increasing importance of direct access to biological information resources using computer networks, the EMBL Data Library, in collaboration with the EMBL computer group, has developed a number of systems which provide the user community with electronic access to our stored data. These services are provided free of charge by EMBL.

Fileservers

To provide users with immediate access to the sequence data created between releases the EMBL e-mail file server (36) was set up in 1988. This service operates by electronic mail and further information can be obtained by sending 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.

The main function of the e-mail server is to provide access to individual sequence entries. For example, to get the sequence with accession number X70047, one would send the command **GET NUC:X70047** to the file server. Several index files are updated daily to assist users in finding accession numbers of new sequences of interest to them. There are weekly-updated listings available from the fileserver of all the newest entries in our database.

Improved network connectivity has promoted the utility of our anonymous FTP server. This supplements the e-mail file server, giving more interactive access to complete databases. These include the quarterly releases of the sequence databases plus weekly batches of updates, as well as many related datasets and an extensive archive of molecular biology software. Users should connect to the anonymous FTP server at the address [FTP.EMBL-Heidelberg.DE](ftp://FTP.EMBL-Heidelberg.DE) using the username **anonymous**, giving their e-mail address as the password.

We also offer access to the FTP archives via the Gopher protocol. Gopher clients should connect to the Gopher server at the address [FTP.EMBL-Heidelberg.DE](ftp://FTP.EMBL-Heidelberg.DE) using port 70. Gopher clients simplify the use of network services by hiding complexity behind a simple graphical user interface. Being part of the EMBnet 'biogopher' network, EMBL's Gopher provides links to other information resources in Europe and elsewhere.

Sequence search facilities

The EMBL Data Library provides a set of services, running on the EMBL computer facilities, that allow external users to compare their new sequence against a regularly updated set of protein and nucleotide sequence databases. The newest service of this kind is based on the MPsrch database similarity search program of Collins and Sturrock at the University of Edinburgh. MPsrch uses the well-known Smith and Waterman (43) algorithm for sensitive searches of protein or nucleic acid databases, implemented on a MasPar massively parallel computer at EMBL. On a serial computer such searches take of the order of one hour to carry out. Using MPsrch however, a complete search of SwissProt with a query sequence of 100 amino acids can be completed in less than 30 seconds. The results are mailed back to the sender. Instructions for use may be obtained by sending a specially formatted e-mail message to the Internet address Blitz@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 instructions describe how to specify

the parameters of the search (gap penalty, weight matrix) and the number of alignments to be returned in the output.

Mail-Quicksearch and Mail-FastA are two alternative services provided on the EMBL computing system. Mail-Quicksearch is based on the GCG software package and is suitable for finding sequences which are very similar to the query sequence. Sending an e-mail message to the Internet address Quick@EMBL-Heidelberg.DE, with the word **HELP** in the body of the message results in the return of a full set of instructions on how to use the service. Mail-FastA is based on Pearson's FastA program (41). It performs sensitive comparisons of nucleotide or amino acid sequences against the database. Further information can be obtained by sending a mail message to the Internet address Fasta@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.

EMBnet

The European Molecular Biology Network (EMBnet) was initiated in 1988 as an attempt to increase the availability and usefulness of the molecular biology databases within Europe. Remote copies of the nucleotide sequence database are held at nationally mandated nodes. These EMBnet nodes receive daily updates from the EMBL Data Library, including all new and updated entries from the collaborating databases (DDBJ/EMBL/GenBank). Nodes receiving such daily data are listed in Table 2. Many of the EMBnet nodes provide on-line interactive services to their academic and commercial user communities. By these means access to a large number of molecular biology databases has been brought in easy reach of most of Europe's scientific community.

The European Bioinformatics Institute (EBI)

In March of this year the EMBL Council made the decision to establish a new institute, an EMBL outstation, which is to be the future home of the Data Library. This new institute, the EBI, will be situated near Cambridge, UK, and should provide the environment necessary to continue to develop and expand the Data Library's existing activities. Continuity of the services is guaranteed throughout the transition period, which is expected to have been completed by summer 1995.

How to contact the EMBL Data Library

Network: Datasubs@EMBL-Heidelberg.DE (for data submissions)
 Datalib@EMBL-Heidelberg.DE (for general enquiries)
 Update@EMBL-Heidelberg.DE (for corrections to nucleotide entries)
 Netserv@EMBL-Heidelberg.DE (file server)
 NetHelp@EMBL-Heidelberg.DE (for network server enquiries)
 [FTP.EMBL-Heidelberg.DE](ftp://FTP.EMBL-Heidelberg.DE) (anonymous FTP and Gopher servers)
 BLITZ@EMBL-Heidelberg.DE (MPsrch protein sequence search server)
 FASTA@EMBL-Heidelberg.DE (FastA sequence search server)
 QUICK@EMBL-Heidelberg.DE (Quicksearch sequence search 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

REFERENCES

1. Higgins, D.G., Fuchs, R., Stoehr, P.J. and Cameron, G. N. (1992) *Nucleic Acids Res.*, **20**, 2071–2074.
2. Burks, C., Cinkosky, M.J., Fischer, W.M., Gilna, P., Hayden, J.E.-D., Keen, G.M., Kelly, M., Kristofferson, D. and Lawrence, J. (1992) *Nucleic Acids Res.*, **20**, 2065–2069.
3. Bairoch, A. and Boeckmann, B. (1993) *Nucleic Acids Res.*, this issue.
4. Barker, W.C., George, D.G., Mewes, H.-W. and Tsugita, A. (1992) *Nucleic Acids Res.*, **20**, 2023–2036.
5. Bairoch, A. (1993) *Nucleic Acids Res.*, this issue.
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. Bairoch, A. (1993) *Nucleic Acids Res.*, this issue.
8. Kröger, M., Wahl, R., Schachtel G. and Rice, P. (1992) *Nucleic Acids Res.*, **20**, 2119–2144.
9. Bucher, P. and Trifonov, E.N. (1986) *Nucleic Acids Res.*, **14**, 10009–10026.
10. Roberts, R.J. and Macelis D. (1985) *Nucleic Acids Res.*, **20**, 2167–2180.
11. Ashburner, M. (1990) University of Cambridge, Cambridge.
12. Ghosh, D. (1992) *Nucleic Acids Res.*, **20**, 2091–2093.
13. prinzl, M., Dank, N., Nock, S. and SchÖn, A. (1991) *Nucleic Acids Res.*, **19**, 2127–2171.
14. D Rijk, P., Neefs, J.-M., Van de Peer, Y. and De Wachter, R. (1992) *Nucleic Acids Res.*, **20**, 2075–2089.
15. 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. (1992) *Nucleic Acids Res.*, **20**, 2027–2063.
16. Specht, T., Wolters, J. and Erdmann, V.A. (1991) *Nucleic Acids Res.*, **19**, 2189–2191.
17. Bairoch, A. (1991) University of Geneva, Geneva.
18. Lawton, J.R., Martinez, F.A. and Burks, C. (1989) *Nucleic Acids Res.*, **17**, 5885–5899.
19. Hanks, S.K., Quinn, A.M. and Hunter, T. (1992) Salk Institute.
20. Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) *Genomics*, **13**, 1095–1107.
21. Zwieb, C. and Larsen, N. (1992) *Nucleic Acids Res.*, **20**, 2207.
22. Wu, T.T. (1992) Technological Institute, Northwestern University, Illinois.
23. Zehetner, G. (1992) ICRF, Genome Analysis Laboratory, London.
24. Brown, C. (1993) *Nucleic Acids Res.*, this issue.
25. Nelson, M. and McClelland, M. (1991) *Nucleic Acids Res.*, **19**, 2045–2071.
26. Shumyatsky, G. and Reddy, R. (1992) *Nucleic Acids Res.*, **20**, 2159–2165.
27. Wada, K., Wada, Y., Ishibashi, F., Gojobori, T. and Ikemura, T. (1992) *Nucleic Acids Res.*, **20**, 2111–2118.
28. Wallace, J.C. and Henikoff, S. (1992) *CABIOS*, **8**, 249–254.
29. Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) *Protein Science*, **1**, 409–417.
30. Sander, C. (1992) EMBL, Heidelberg.
31. Jurka, J. and Smith, T. (1988) *Proc. Natl. Acad. Sci. U.S.A.*, **85**, 4775–4778.
32. Gelfand, M.S. (1991) Institute of Protein Research, USSR Academy of Sciences, Puschino.
33. Pascarella, S. and Argos, P. (1992) *Protein Engineering*, **5**, 121–137.
34. Marsh, S. (1993) ICRF, London.
35. Raschke, E. (1993) *Genetic Analysis, Techniques and Applications*, **10**, in press.
36. Stoehr, P. J. and Omond, R.A. (1989) *Nucleic Acids Res.*, **17**, 6763–6764.
37. The EMBL Data Library (1993) *Nucleic Acids Res.*, **21**, i-vii.
38. The EMBL Data Library (1992) *Plant Molecular Biology*, **18**, 1221–1224.
39. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merrill, C.R., Wu, A., Olde, B., Moreno, R.F., Kerlavage, A.R., McCombie, W.R. and Venter, J.C. (1991) *Science*, **252**, 1651–1656.
40. Fuchs, R. and Stoehr, P. (1993) *CABIOS*, **9**, 71–77.
41. Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.
42. Higgins, D.G. and Stoehr, P.J. (1992) *CABIOS*, **8**, 137–139.
43. Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.