Volume 16 Number 5 1988 Nucleic Acids Research

The EMBL data library

Graham N.Cameron

European Molecular Biology Laboratory, Meyerhofstrasse 1, 6900 Heidelberg, FRG
Submitted August 17, 1987

BACKGROUND

The EMBL Data Library was established in 1980 to collect, organize and distribute published nucleic acid sequence data. At that time GenBank® was being established in America with an almost identical brief. The two groups decided to collaborate, and this collaboration still flourishes today.
Twelve releases of the Nucleotide Sequence Data Library have now been distributed, and the collection has grown from half a million to about 14 million bases (Release 12, July 1987). The data are distributed to about 200 sites in 25 countries, and many software developers redistribute them with analysis software. In all, the nucleotide sequence collection is in use by some 10000 scientists throughout the world.
The Data Library Group within EMBL has 8 full time staff, and is funded entirely within the budget of the Laboratory.

SERVICES PROVIDED

EMBL currently distributes the following databases on magnetic tape:

1. The EMBL Nucleotide Sequence Data Library
2. The Protein Identification Resource database
3. The SWISS-PROT protein database
4. A Restriction Enzyme database.

The EMBL Nucleotide Sequence Data Library is produced by the EMBL Data Library in collaboration with GenBank. The Protein Identification Resource (PIR) database is provided by the National Biomedical Research Foundation (USA), and simply redistributed by EMBL. The SWISS-PROT database was developed by Amos Bairoch at the University of Geneva. It is a protein sequence collection whose format is compatible with the EMBL Nucleotide Sequence Data Library, and is built by reformatting data provided in the PIR database and translating the EMBL nucleotide sequence database. Over the coming months EMBL will take over an increasing responsibility for the production of the SWISS-PROT database. The restriction enzyme database is prepared by Dr. R. J. Roberts at Cold Spring Harbor Laboratory.
An annual subscription, designed to recover our distribution costs, is charged as summarised in table 1.
Subscribers receive four releases of the EMBL Nucleotide Sequence Data Library each year, and may elect to receive copies of the latest PIR and SWISS-PROT releases as they become available during the year. Researchers can request the latest release of any of these databases at any time at a charge of one quarter of the annual subscription.

Table 1: Subscription charges

	Annual subscription (databases)	Documentation only
Academic Users (Member states)	DM 400	DM 200
Academic Users (Countries not members of EMBL)	DM 800	DM 200
Commercial Users	DM 2000	DM 200

```
ID   CTGL01     standard; DNA; 945 BP.
XX
AC   X00920;
XX
DT   01-APR-1986  (author review)
DT   25-SEP-1985  (FT modified)
DT   18-JAN-1985  (first entry)
XX
DE   Chironomus thummi thummi globin gene for globin IV
XX
KW   signal peptide; globin.
XX
OS   Chironomus thummi (midge, chironome, Muecke)
OC   Eukaryota; Metazoa; Arthropoda; Insecta; Diptera.
XX
RN   [1]   (bases 1-945; enum. 1 to 945)
RA   Antoine M., Niessing J.;
RT   "Intron-less globin genes in the insect Chironomus thummi thummi";
RL   Nature 310:795-798(1984).
XX
CC   Data kindly reviewed (17-APR-1985) by J. Niessing
XX
FH   Key         From     To       Description
FH
FT   PRM          228     231      TATA-box
FT   CAP          260     260      cap site
FT   CDS          306     350      signal peptide
FT   CDS          351     758      globin IV
FT   SITE         819     824      polyA signal
FT   POLYA        842     842      polyA site
XX
SQ   Sequence  945 BP;  294 A;  185 C;  160 G;  306 T;
     ctttatttat gtggaaattt tttttccaga atatcgagca gaatatcact agtattgaaa
     aagaggtaat taaataagct caaattatta tagagtttgt tgacctttc taatgattat
     gtggttgaaa acagtaaaaa aaacaaaata gaaaatctct tttgattgca taacgatgtt
     tcttatctca cagcttttca caataatgtc ttctcaaaat ttttaagtat aaatggagca
     caaatttcga tagtaaatca gttcttcaat tcgtttcaaa gttgtaactt cacaaaccaa
     tcaaaatgaa actcctcatt cttgccttgt gcttcgccgc tgcctcagcc ttgactgctg
     accaaatcag cacagtccaa tcatcatttg ctggagttaa gggagatgct gttggtatcc
     tctatgccgt tttcaaagct gatccatcaa tccaagccaa attcacacaa ttcgctggaa
     aggacctcga ctcaatcaag ggatcagctg atttctcagc tcatgccaac aaaattgtcg
     gattcttctc aaagatcatc ggagaccttc caaacattga tggagatgtc accacattcg
     ttgcctcaca cacaccccgt ggagttacac atgatcaatt gaacaacttc cgtgctggat
     tcgtcagcta catgaaggct cacaccgact tcgctggagc cgaagctgcc tggggtgcaa
     ctcttgatgc tttcttcgga atggtcttcg ccaagatgta aatcttttaa atatcaatga
     tatttattag tagtgcctta atttatgaca aacatggaaa taaaaaaaaa tatcgtttat
     ggtttaaatt tttgttgttt tatcttgaat ttcctatgac ttattggaaa aagatttcag
     aacgttgatt gtacttgttt atagtgaagc atataattct caagc
//
```

Figure 1: A sample entry from the EMBL Nucleotide Sequence Data Library

NUCLEOTIDE SEQUENCE DATA LIBRARY

Release 12 of the Nucleotide Sequence Data Library is currently being distributed (August 1987). It contains almost 13000 sequences, reporting a total of 13.6 million bases drawn from nearly nine thousand references. Data are submitted to the Data Library on our Data Submission Forms, designed to solicit all the information necessary to create a database entry. These forms are distributed to researchers by journals and are available from the Data Library in both machine readable and paper versions. In addition, in collaboration with GenBank, all major journals are scanned and data not already received as direct submissions are included.

In the future we will be encouraging researchers to take more responsibility for the completeness and accuracy of the database. We plan, for example, to distribute programs which query the user for the information necessary to build entries for submission to the database. This is desirable both because it will enable us to cope with the ever increasing flood of data, and because we can never expect to annotate data as well as the researcher working with them.

Structure of the Nucleotide Sequence Data Library

The data are distributed as "flat" text files containing entries, where each entry comprises a single contiguous sequence and its accompanying annotation. Different line types, each with its 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 (CTGL01 in fig. 1), and across releases by its Accession Number list (X00920 in fig. 1). References to EMBL Data Library entries should always cite the primary (first) accession number, which identifies the same data from release to release.

SWISS-PROT PROTEIN SEQUENCE DATA LIBRARY

Release 5 of the SWISS-PROT protein sequence database was readied for distribution in September 1987. The distribution format is, as far as possible, analagous to that of the Nucleotide Sequence Data Library. Entries are distributed as flat files with two letter line type codes, and are citable by accession number. Future releases will be prepared shortly after Nucleotide Sequence data releases and will include translations of the newest data.

FUTURE DIRECTIONS

SWISS-PROT is an exciting addition to the services offered by EMBL. We plan to take over more and more responsibility for this collection, and are committed to the provision of amino acid and nucleic acid sequence data in a mutually compatible form. We hope soon to be able to offer access to the data via computer networks, in addition to our regular magnetic tape releases. We are also engaged in a complete restructuring of our database which will include its installation in the ORACLE database management system. In the first instance we hope that all the user will see will be a more efficient service, but ultimately it will enable an online query system to be offered. EMBL aims to expand its role in the field of information services providing, in the longer run, online access to a broad range of molecular biological data and a comprehensive consultancy service.

CONTACT

Enquiries about the EMBL Data Library should be addressed to:

The EMBL Data Library	Telephone	(06221) 387258
Postfach 10 22 09	Telex	461613 (embl d)
6900 Heidelberg	Telefax	(06221) 387306
West Germany	Computer network	datalib@embl.earn

Data submitted should be sent either by post to "Data Submissions" at the above address, or preferably electronically to datasubs@embl.earn.