Volume 14 Number 1 1986 Nucleic Acids Research

The EMBL data library

Gregory H.Hamm and Graham N.Cameron

European Molecular Biology Laboratory, Meyerhofstrasse 1, 6900 Heidelberg, FRG

Received 15 October 1985

ABSTRACT

The EMBL Data Library was the first internationally supported central resource for nucleic acid sequence data. Working in close collaboration with its American counterpart, GenBank (1), the library prepares and makes available to the scientific community a comprehensive collection of the published nucleic acid sequences. This paper describes briefly the contents of the database, how it is available, and possible future enhancements of Data Library services.

BACKGROUND

The EMBL Data Library was established in October, 1980, with three primary goals:

- to make freely available a reliable and comprehensive collection of the published nucleic acid sequence data;
- to encourage standardization and free exchange of data in the international molecular biology community;
- to serve as a European focus for efforts devoted toward computing and information services in molecular biology.

At that time, several groups worldwide were engaged in the collection of nucleic acid sequences, but no designated group anywhere had secured central support for the establishment of a permanent resource. Given the rapidly growing body of nucleic acid sequence data, this situation was a cause of concern for many European scientists, who recommended that EMBL host the required resource. The recommendation was implemented, and, following an initial period of research and development, the Data Library made its first release in April, 1982.

The Data Library has a full-time staff of seven, and is presently funded entirely within the budget of the European Molecular Biology Laboratory.

SERVICES PROVIDED

Data Distribution

The EMBL Nucleotide Sequence Data Library is distributed on magnetic tape, free of charge, to anyone interested. Tapes can be ordered by writing directly to the EMBL Data Library. The magnetic tapes are available in various formats so as to be readable on nearly any computer with a 9-track tape drive. Supplied with the collection, both on tape and in printed form, are a User Manual (2), which contains a definition of the data formats and conventions used, and Release Notes (3), which describe the contents of the current release and include an extensive set of indices (e.g., by journal, keyword, and species) for use with the data.

No restrictions are placed on use or redistribution of the data. (The EMBL collection is redistributed by a number of other research resources and commercial firms in the context of software systems or on-line services, including BIONET (4).) Typically 200 user sites in 25 countries request each release. A recent survey suggests that this represents a total user community of some 4000 individual scientists. The present distribution frequency is four releases per year.

A printed compendium (5) of the database is prepared together with GenBank and is presently published yearly as a special publication of this journal. Recently, the EMBL Data Library began serving as a European distribution source for copies of the Protein Identification Resource database (6) kindly provided by the National Biomedical Research Foundation.

Consultation and Training

The Data Library also provides informal support to scientists (particularly in Europe) seeking advice on hardware, software, and data requirements for research involving sequence data. This usually takes the form of telephone consultation, but recently was expanded to include the organisation of an international course in computational sequence analysis. Scientists also visit the Data Library to work on specific projects using the EMBL computing facilities.

NUCLEIC ACID SEQUENCE DATABASE

Contents

The EMBL Data Library is presently distributing Release 6, which contains some 4.5 million bases drawn from more than three thousand references. Most data are abstracted from approximately 35 journals which are routinely reviewed. This work is shared between the EMBL Data Library and GenBank.

Table 1 - Growth of the EMBL Data Library

	Release 1		Release 4		Release 5		Release 6	
References cited	387		1275		1710		3308	
Sequences	Entries	Bases	Entries	Bases	Entries	Bases	Entries	Bases
Artificial	3	4688	11	10163	12	14066	134	41962
Chloroplast	7	5873	31	29495	60	51401	96	83832
Genetic elements	17	17015	22	24471	23	27180	38	36948
Mitochondrial	25	74422	69	150068	98	170201	241	255461
Prokaryotic	61	58288	214	293437	297	392693	627	625207
Viral/Phage	123	164692	272	720041	334	834979	790	1188124
Eukaryotic	332	260455	1079	919530	1538	1362366	2904	2326404
unclassified					16	21607	5	9654
Total	568	585433	1698	2147205	2378	2874493	4835	4567592

A small but growing number of entries are drawn from direct submissions (to either EMBL or GenBank) by authors.

The distribution tapes, including annotation and indices, contain some 22 million characters. The rapid growth in the volume of data in all categories is shown in Table 1. The next release will approach ten times the size (in terms of nucleotides) of the first release three years ago.

Structure

The Nucleotide Sequence Data Library organisation was chosen in an attempt to make the data as easily accessible as possible without restricting their usefulness to a particular type of computing environment. A simple "flat file" organisation is used, so that users with limited computing experience can work with the data easily, while that those requiring a more complex structure can write programs to perform the necessary reorganisation.

The nucleic acid sequence database is composed of entries, each of which corresponds to a single contiguous sequence as contributed or reported in the literature. (Entries are often assembled from several papers reporting overlapping data.) Entries in the database are structured so as to be usable by human readers as well as by computer programs.

A database entry is composed of lines of text. Different types of lines, each with its own format, are used to record the various types of data which make up the entry. The type of each line is indicated by a simple two-character line code. This device allows very simple programs to be written to access particular information without having to deal with other data in the entry. It also permits us to add new data items in the future

```
ID   CTGL01     standard; DNA; 945 BP.
XX
AC   X00920;
XX
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, )
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
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         829     834      polyA signal
FT   POLYA        842     842      polyA site
XX
SQ   Sequence  945 BP;  294 A;  185 C;  306 T;  160 G.
     CTTTATTTAT GTGGAAATTT TTTTTCCAGA ATATCGAGCA GAATATCACT AGTATTGAAA
     AAGAGGTAAT TAAATAAGCT CAAATTATTA TAGAGTTTGT TGACCTTTTC 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 typical entry from the EMBL database, identified by the primary accession number X00920.

without breaking existing user programs.

The sample entry from the database shown in Figure 1 illustrates the different types of lines. For example, the "DE" line contains a brief description of the sequence, whereas the "FT" lines comprise a feature table which records regions of interest in the sequence.

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 will identify the same data regardless of changes which occur from release to release. Also, both the EMBL and GenBank databases use the same accession number for any given entry.

FUTURE DIRECTIONS

Current development efforts are directed toward improvements in the timeliness, completeness and utility of the database. For example, much review work is being automated, so that the data will appear more rapidly following publication. With GenBank, effort is also being devoted toward making the two databases identical in content, with automatic conversion between them. Other work is underway to improve feature representation and nomenclature to simplify computer searches and valid statistical analyses. More generally, the EMBL is presently reviewing its entire Biocomputing Programme with a view toward an expanded role in both the research and services areas. A few of the service enhancements under discussion include the addition of other types of biological databases, the implementation of on-line retrieval services, and the development of an electronic bulletin board for European molecular biologists. Whether or not any of these are pursued, the commitment to the present Data Library will be maintained.

REFERENCES

1. Bilofsky, H.S., Burks, C., Fickett, J.W., Goad, W.B., Lewitter, F.I., Rindone, W.P., Swindell, C.D., and Tung, C-S. (1986) Nucl. Acids Res. 14 (this issue).
2. EMBL Nucleotide Sequence Data Library User Manual, Release 6 (1985), European Molecular Biology Laboratory.
3. EMBL Nucleotide Sequence Data Library Release Notes, Release 6 (1985), European Molecular Biology Laboratory.
4. Smith, D.H., Brutlag, D., Friedland, P., and Kedes, L.H. (1986) Nucl. Acids Res. 14 (this issue).
5. EMBL Data Library and GenBank staff (eds.), Nucleotide Sequences 1985 (1985), IRL Press, Ltd.
6. Barker, W., et al (1986) Nucl. Acids Res. 14 (this issue).