Nucleic Acids Research, 2004, Vol. 32, Database issue ***D27–D30***
DOI: 10.1093/nar/gkh120

The EMBL Nucleotide Sequence Database

Tamara Kulikova*, Philippe Aldebert, Nicola Althorpe, Wendy Baker, Kirsty Bates, Paul Browne, Alexandra van den Broek, Guy Cochrane, Karyn Duggan, Ruth Eberhardt, Nadeem Faruque, Maria Garcia-Pastor, Nicola Harte, Carola Kanz, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Renato Mancuso, Michelle McHale, Francesco Nardone, Ville Silventoinen, Peter Stoehr, Guenter Stoesser, Mary Ann Tuli, Katerina Tzouvara, Robert Vaughan, Dan Wu, Weimin Zhu and Rolf Apweiler

EMBL Outstation, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 15, 2003; Revised October 9, 2003; Accepted October 20, 2003

ABSTRACT

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/), maintained at the European Bioinformatics Institute (EBI), incorporates, organizes and distributes nucleotide sequences from public sources. The database is a part of an international collaboration with DDBJ (Japan) and GenBank (USA). Data are exchanged between the collaborating databases on a daily basis to achieve optimal synchrony. The web-based tool, Webin, is the preferred system for individual submission of nucleotide sequences, including Third Party Annotation (TPA) and alignment data. Automatic submission procedures are used for submission of data from large-scale genome sequencing centres and from the European Patent Office. Database releases are produced quarterly. The latest data collection can be accessed via FTP, email and WWW interfaces. The EBI's Sequence Retrieval System (SRS) integrates and links the main nucleotide and protein databases as well as many other specialist molecular biology databases. For sequence similarity searching, a variety of tools (e.g. FASTA and BLAST) are available that allow external users to compare their own sequences against the data in the EMBL Nucleotide Sequence Database, the complete genomic component subsection of the database, the WGS data sets and other databases. All available resources can be accessed via the EBI home page at http://www.ebi.ac.uk.

INTRODUCTION

The mission of the Service Programme at the EBI is the building, maintenance and provision of biological databases and other information services to support data deposition and access by the scientific community (1). Databases provided at the EBI include the EMBL Nucleotide Sequence Database, the protein databases Swiss-Prot, TrEMBL (2) and UniProt (3), InterPro (4), the Macromolecular Structure Database (E-MSD) (5), the gene expression database ArrayExpress (6) and the Ensembl automatic genome annotation database (7).

In Europe, most nucleotide sequence data and supporting bibliographical and biological data generated are collected and distributed by the EMBL Nucleotide Sequence Database. The EMBL database is a member of the International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank. The main sources of data in the EMBL database are large-scale genome sequencing projects, direct submissions from individual scientists and sequence data extracted from biotechnology patent applications to the European Patent Office. To achieve optimal worldwide synchrony, all new and updated database records are exchanged on a daily basis between EMBL, DDBJ (8) and GenBank (9). Third Party Annotation (TPA) and CONstructed (CON) records are also exchanged daily, while Whole Genome Shotgun (WGS) data sets are exchanged when they become available or have been updated.

EMBL database releases, with accompanying release notes, are produced quarterly.

Within the last 12 months the database size has increased from 18.3 million entries comprising 23 Gb (Release 72, September 2002) to 27.2 million entries comprising over 33 Gb (Release 76, September 2003). The number of organisms represented in the database is now ~150 000.

During the course of 2003, the EMBL Sequence Version Archive was launched, the WGS data collection and distribution procedure was further developed and the data collection rules for the TPA data set continued to be revised. A detailed and up-to-date description of EMBL Nucleotide Sequence Database activities can be found at http://www.ebi.ac.uk/embl/.

SUBMISSIONS TO THE EMBL NUCLEOTIDE SEQUENCE DATABASE

A repository of primary nucleotide sequences is an essential requirement for computational analysis and genome research.

*To whom correspondence should be addressed. Tel: +44 1223 494463; Fax: +44 1223 494468; Email: kulikova@ebi.ac.uk

Furthermore, molecular biologists depend on free access to such a repository. Many journals require authors to submit sequence information to the EMBL, GenBank or DDBJ database prior to publication in order to ensure its availability to scientists.

An introduction to database submission procedures is described below. For comprehensive details of procedures, please see http://www.ebi.ac.uk/embl/Submission/.

Webin

Webin is the preferred submission system for nucleotide sequence and biological annotation. Webin has been designed to allow rapid submission of single, multiple or very large numbers of sequences (bulk submissions) and is available at http://www.ebi.ac.uk/embl/Submission/webin.html. Webin has been modified to accept TPA submissions.

Genome project submissions

Database entries produced at the sequencing site can be deposited and updated directly by the submitters using FTP or email. Groups producing large volumes of genome sequence data over an extended period of time are advised to contact the database at datasubs@ebi.ac.uk.

Alignment submissions

EMBL-Align (10) is a public data set of both protein and nucleotide multiple sequence alignments and can be queried from the EBI Sequence Retrieval Server (SRS) server. It was developed in response to the need for permanent electronic storage and standardized presentation of alignment data from phylogenetic and population analyses. Webin-Align is a dedicated web-based tool for submission of multiple sequence alignments in all common alignment formats. Webin-Align is available at http://www.ebi.ac.uk/embl/Submission/align_top.html.

ACCESSING NUCLEOTIDE SEQUENCE DATA AND RELATED DATA AT THE EBI

The EMBL Nucleotide Sequence Database is available from the EBI network services, interactively via the WWW or by email, using netserv (netserv@ebi.ac.uk). EMBL data sets are freely available from the EBI FTP server at ftp://ftp.ebi.ac.uk/pub/databases/embl/. For more information see http://www.ebi.ac.uk/embl/Access/.

For a complete list of internet-based resources of the EMBL Nucleotide Sequence Database see Supplementary Table 1.

Completed genome sequences and proteome analysis

Direct access to several thousand completely sequenced genomic components is available via the EBI Genomes server at http://www.ebi.ac.uk/genomes/. Proteome analysis information on all completely sequenced organisms is available at http://www.ebi.ac.uk/proteome/ (11).

Whole Genome Shotgun (WGS) data

WGS data are available at ftp://ftp.ebi.ac.uk/pub/databases/embl/wgs/. At the time of writing (September 2003), data sets from 70 separate WGS projects are available. The largest data set is that of *Rattus norvegicus*. While many of the WGS data sets are not annotated, some biological features are present in some of the sets. The WGS data set for *Anopheles gambiae* strain *PEST* is an example with annotation. WGS data sets can now be searched using the FASTA algorithm (see below).

Sequence Retrieval System (SRS)

The EMBL Nucleotide Sequence Database can be accessed via the EBI SRS server (12,13) at http://srs.ebi.ac.uk/. In SRS, the data are available in the following libraries:

(i) EMBL: the database in its entirety by means of a virtual library comprising EMBLRELEASE, EMBLNEW, EMBLTPA and EMBLWGS;
(ii) EMBLRELEASE: library containing the latest official release of the EMBL Nucleotide Sequence Database;
(iii) EMBLNEW: library containing updated and new entries created since the last
official release;
(iv) EMBLTPA: library containing TPA entries;
(v) EMBLWGS: library containing WGS entries;
(vi) EMBLCON: library containing CON entries.

Sequence searching

A comprehensive set of sequence analysis and database search algorithms is available at http://www.ebi.ac.uk/Tools/. Sequence similarity searches are available interactively over the WWW as well as by email. Users can search the EMBL Nucleotide Sequence Database as a whole or by individual taxonomic division.

The most commonly used algorithms available are FASTA (14) and WU-BLAST (15), permitting comparisons between nucleotide query sequences and the nucleotide or protein databases as well as searches of protein query sequences against the nucleotide database.

The FASTA service for genomes and proteomes (http://www.ebi.ac.uk/fasta33/genomes.html) enables users to search interactively completed genomes and proteomes. The same searches can be performed by email (gpfasta@ebi.ac.uk). User instructions are available by sending an email with the word HELP in the body of the message to gpfasta@ebi.ac.uk. WGS data sets are now available for searching.

Sequence analysis

Sequence analysis programs offered include multiple sequence alignment and inference of phylogenies using ClustalW (16), protein classification using InterProScan (17) and others. The EBI also provides interactive sequence analysis resources based on the European Molecular Biology Open Software Suite (18) (EMBOSS) (http://www.emboss.org/).

DEVELOPMENTS

Sequence length limits

Currently, database records are limited in length to 350 000 bp. At the DDBJ/EMBL/GenBank collaborative meeting of May 2003, a decision was taken to remove the size restriction on database records in June 2004.

This development will allow the entire sequence derived from a naturally occurring biological unit to be stored as a single database entry, thus eliminating the need to split long sequences into segments and create CON entries to store the

assembly information (19). Currently, ~3% of all base pairs in the database are stored in the constituent segment entries of CON entries.

Third Party Annotation (TPA) data set

Until recently, the collaborative databases have collected and distributed only primary nucleotide sequence and annotation data resulting from direct sequencing of such molecules as cDNAs, ESTs and genomic DNA. 'Primary data' is defined as annotated sequence that has been determined by submitters and their teams. Primary database entries remain in the ownership of the original submitter and the co-authors of the submission publication(s). The owners of database entries have privileges to implement updates to the data.

In response to demand from the research community, the collaborative databases have created the TPA data set. The types of data that make up the TPA data set include reannotations of existing entries, combinations of novel sequence and existing primary entries and annotation of trace archive and WGS data.

TPA data are submitted using Webin. Submitters are required to provide DDBJ/EMBL/GenBank accession and version numbers and nucleotide locations for all primary entries to which their TPA entry relates. For TPA sequences composed from trace archive data, the identifier (e.g. TI123445566) and corresponding nucleotide locations must be provided.

TPA entries can be distinguished easily from their primary counterparts. The abbreviation 'TPA:' appears at the beginning of each description (DE) line and the keywords 'Third Party Annotation' and 'TPA' appear in the keyword (KW) line.

```
AH TPA_SPAN PRIMARY_IDENTIFIER PRIMARY_SPAN COMP
AS 1-251    BE529226.1         1-251
AS 68-450   BE524624.1         1-383
AS 394-1086 AJ420881.1         1-693
AS 826-1211 AV561543.1         1-386
```

The flat-file extract shown above (from BN000024) shows the two new line types that have been created for TPA entries. The Assembly Header (AH) line provides column headings for the assembly information. The Assembly (AS) lines provide information on the composition of the TPA sequence by listing base span(s) of the TPA sequence together with identifiers and base span(s) of contributing sequences.

In order to ensure sequence annotation of the highest quality, entries that are yet to be discussed in peer-reviewed publications are held confidential and are not visible to database users. This is an important difference from our policy of data release for primary entries.

At the time of writing (September 2003), 457 TPA entries are publicly available, of which 150 entries are of human origin. The second most common source organism for this type of entry is mouse, with 95 entries, showing that, so far, the TPA data set follows the same pattern as the primary dataset. Statistics for all EMBL Nucleotide Sequence Database data, including top 10 organisms by base count, can be found at http://www3.ebi.ac.uk/Services/DBStats/. Further information on the TPA dataset can be found at http://www.ebi.ac.uk/embl/Documentation/third_party_annotation_dataset.html and instructions on data submission can be found at http://www.ebi.ac.uk/webin/webin_help.html.

EMBL Sequence Version Archive (SVA)

The EMBL SVA (20) was created to provide access to all versions of EMBL Nucleotide Sequence Database entries, including CON, TPA and WGS data. There were 145 million entry versions in the archive by September 2003, and new versions are being added every day. Entries from all past EMBL Nucleotide Sequence Database releases, starting with the first release in 1982, have been loaded into the archive.

Each time an EMBL database entry is created or modified it is loaded into the archive, where it can be accessed and compared with other versions of the same entry. If an entry is updated, corrected or extended as a result of new findings from recent experiments, the entry version is incremented. Changes in the taxonomic lineage, or flat-file formatting changes are not reflected in the entry version. For this reason, the archive may contain several variants of an entry with the same entry version number.

The archive can be accessed interactively at http://www.ebi.ac.uk/embl/sva/ and programmatically at http://www.ebi.ac.uk/cgi-bin/dbfetch.

Entries can be retrieved interactively using accession numbers, protein identifiers and sequence versions. The user chooses to view either the complete chronological history of an entry or the entry version that was current at a specified date. The resulting entry versions can be viewed, downloaded and compared. The interactive interface can also be reached by following hyperlinks from the EBI SRS query results page when working with EMBL Nucleotide Sequence Database and EMBL-Align entries. As an example of programmatic entry retrieval, the following URL returns the latest EMBL entry having the accession number AC067752: http://www.ebi.ac.uk/cgi-bin/dbfetch?db=SVA&id=AC067752&format=default.

XML format for data exchange

The EMBL Nucleotide Sequence Database has initiated efforts to produce an XML format for the distribution of entries. The development of this format will be carried out in collaboration with DDBJ and GenBank with the aim of developing a common representation for the distribution of data.

CITING THE EMBL NUCLEOTIDE SEQUENCE DATABASE

The preferred form for citation of the EMBL Nucleotide Sequence Database is: Kulikova, T. *et al.* (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **32**, D27–D30.

CONTACTING THE EMBL NUCLEOTIDE SEQUENCE DATABASE

Computer network: data submissions, datasubs@ebi.ac.uk; other inquiries, datalib@ebi.ac.uk; updates/publication notifications, update@ebi.ac.uk. Postal address: EMBL Nucleotide Sequence Database, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Tel: data submissions, +44

1223 494499; general, +44 1223 494444. Fax: general, +44 1223 494468.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

REFERENCES

1. Brooksbank,C., Camon,E., Harris,M.A., Magrane,M., Martin,M., Mulder,N., O'Donovan,C., Parkinson,H., Tuli,M., Apweiler,R. *et al.* (2003) The European Bioinformatics Institute's data resources. *Nucleic Acids Res.*, **31**, 43–50.
2. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
3. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
4. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
5. Boutselakis,H., Dimitropoulos,D., Fillon,J., Golovin,A., Henrick,K., Hussain,A., Ionides,J., John,M., Keller,P.A., Krissinel,E. *et al.* (2003) E-MSD: the European Bioinformatics Institute Macromolecular Structure Database. *Nucleic Acids Res.*, **31**, 458–462.
6. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P., Lara,G.G. *et al.* (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.
7. Clamp,M., Andrews,D., Barker,D., Bevan,P., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res.*, **31**, 38–42.
8. Miyazaki,S., Sugawara,H., Gojobori,T. and Tateno,Y. (2003) DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res.*, **31**, 13–16.
9. Benson,D.A., Karsch-Mizrachi,I., Lipman D.J., Ostell,J. and Wheeler D.L. (2003) GenBank. *Nucleic Acids Res.*, **31**, 23–27.
10. Lombard,V., Camon,E.B., Parkinson,H.E., Hingamp,P., Stoesser,G. and Redaschi,N. (2002) EMBL-Align: a new public nucleotide and amino acid multiple sequence alignment database. *Bioinformatics*, **18**, 763–764.
11. Pruess,M., Fleischmann,W., Kanapin,A., Karavidopoulou,Y., Kersey,P., Kriventseva,E., Mittard,V., Mulder,N., Phan,I., Servant,F. *et al.* (2003) The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. *Nucleic Acids Res.*, **31**, 414–417.
12. Zdobnov,E.M., Lopez,R., Apweiler, R. and Etzold,T. (2002) The EBI SRS server—new features. *Bioinformatics*, **18**, 1149–1150.
13. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server—recent developments. *Bioinformatics*, **18**, 368–373.
14. Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.*, **24**, 307–331
15. Lopez,R., Silventoinen,V., Robinson,S., Kibria,A. and Gish,W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.*, **31**, 3795–3798.
16. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
17. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.
18. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
19. Stoesser,G., Baker,W., van den Broek,A., Garcia-Pastor,M., Kanz,C., Kulikova,T., Leinonen,R., Lin,Q., Lombard,V., Lopez,R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res.*, **31**, 17–22.
20. Leinonen,R., Nardone,F., Oyewole,O., Redaschi,N. and Stoehr,P. (2003) The EMBL sequence version archive. *Bioinformatics*, **19**, 1861–1862