Nucleic Acids Research, 2005, Vol. 33, Database issue ***D29–D33***
doi:10.1093/nar/gki098

The EMBL Nucleotide Sequence Database

Carola Kanz*, Philippe Aldebert, Nicola Althorpe, Wendy Baker, Alastair Baldwin, Kirsty Bates, Paul Browne, Alexandra van den Broek, Matias Castro, Guy Cochrane, Karyn Duggan, Ruth Eberhardt, Nadeem Faruque, John Gamble, Federico Garcia Diez, Nicola Harte, Tamara Kulikova, Quan Lin, Vincent Lombard, Rodrigo Lopez, Renato Mancuso, Michelle McHale, Francesco Nardone, Ville Silventoinen, Siamak Sobhany, Peter Stoehr, 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 14, 2004; Revised October 6, 2004; Accepted October 14, 2004

ABSTRACT

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl), maintained at the European Bioinformatics Institute (EBI) near Cambridge, UK, is a comprehensive collection of nucleotide sequences and annotation from available public sources. The database is part of an international collaboration with DDBJ (Japan) and GenBank (USA). Data are exchanged daily between the collaborating institutes to achieve swift synchrony. Webin is the preferred tool for individual submissions of nucleotide sequences, including Third Party Annotation (TPA) and alignments. Automated procedures are provided for submissions from large-scale sequencing projects and data from the European Patent Office. New and updated data records are distributed daily and the whole EMBL Nucleotide Sequence Database is released four times a year. Access to the sequence data is provided via ftp and several WWW interfaces. With the web-based Sequence Retrieval System (SRS) it is also possible to link nucleotide data to other specialist molecular biology databases maintained at the EBI. Other tools are available for sequence similarity searching (e.g. FASTA and BLAST). Changes over the past year include the removal of the sequence length limit, the launch of the EMBLCDSs dataset, extension of the Sequence Version Archive functionality and the revision of quality rules for TPA data.

INTRODUCTION

The European Bioinformatics Institute (EBI) is an outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. It is located on the Wellcome Trust Genome Campus near Cambridge, UK.

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 free access by the scientific community (1).

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) is Europe's primary nucleotide sequence resource. This database is the European part of an international collaboration with DDBJ (Japan) (2) and GenBank (USA) (3) (INSDC, International Nucleotide Sequence Database Collaboration). Data are exchanged on a daily basis between the collaborating institutes. The data in the EMBL Nucleotide Sequence Database originates from a combination of large-scale genome sequencing projects, direct submissions from individual scientists and the European Patent Office. There is a quarterly release of the whole database and new and updated records are distributed daily.

Over the last year, the size of EMBL Nucleotide Sequence Database has increased from 27.2 million entries in Release 76, September 2003 to 42.3 million entries in Release 80, September 2004, of which 4.4 million entries are WGS (Whole Genome Shotgun) data. There are now over 185 000 organisms represented in the database.

In 2004, the limit on sequence length has been dropped, the EMBLCDSs dataset containing all coding sequences annotated in the EMBL Nucleotide Sequence Database was launched, the data collection rules for Third Party Anotation

*To whom correspondence should be addressed. Tel: +44 1223 494453; Fax: +44 1223 494468; Email: ckanz@ebi.ac.uk

(TPA) data were revised and the functionality of the Sequence Version Archive was extended further.

Other databases provided by the EBI include the protein resource UniProt (4), InterPro, a database of protein families, domains and functional sites (5), the Macromolecular Structure Database E-MSD (6), the automatic genome annotation database Ensembl (7), Genome Reviews, curated versions of complete Genomes from the EMBL Database, the Enzyme database IntEnz (8) and the database for protein interaction data, IntAct (9).

SUBMISSIONS TO THE EMBL NUCLEOTIDE SEQUENCE DATABASE

Why is it essential to submit new sequence?

Printing sequence data as part of a publication is neither sensible nor manageable, hence journals prefer to cite only the accession number assigned by the INSD Collaboration. Most journals have a mandatory submission procedure such that papers will only be accepted if they have an accession number. The nucleotide sequence is considered part of the publication and therefore almost all nucleotide sequences are publicly available. Having your sequence in the database means it is readily available to the scientific user community. A repository of primary nucleotide sequence data that is freely accessible is essential for computational analysis and genome research.

How to submit new sequences to the EMBL Nucleotide Sequence Database?

The primary tool for submission of nucleotide sequence data is Webin. For alignment data, it is Webin-Align. Projects with large-scale submissions can open a project account allowing direct updates.

Information for submitters can be found here: http://www.ebi.ac.uk/embl/Documentation/information_for_submitters.html. For submission guidelines please see http://www.ebi.ac.uk/embl/Submission/.

Webin

Webin is the preferred submission tool for nucleotide sequences and biological information. It should also be used for TPA submissions. Webin allows fast submissions of single, multiple and very large numbers of sequences (bulk submissions) and is available at http://www.ebi.ac.uk/embl/Submission/webin.html.

Genome project submissions

Large-scale sequencing projects can open a project account to deposit and update data directly using email or ftp. Groups producing large volumes of sequence data are advised to contact the database at datasubs@ebi.ac.uk. More information is available at http://www.ebi.ac.uk/embl/Submission/genomes.html.

Alignment submissions

Webin-Align (10) is the dedicated submission tool for multiple nucleotide and protein alignments. It accepts all common alignment formats and is available at http://www.ebi.ac.uk/embl/Submission/align_top.html.

WGS submissions

WGS data submission is not a continuous process—WGS datasets are normally not updated more often than once every few months. Therefore email or ftp accounts are not opened for the submission of WGS data, but submissions are dealt with on a one-by-one basis. Potential submitters are advised to contact the EMBL database at datasubs@ebi.ac.uk.

How to update entries in the EMBL Nucleotide Sequence Database?

The editorial rights to an entry in the EMBL Nucleotide Sequence Database remain with the original submitter(s). The EBI team adds value to entries, e.g. via cross-references, but the data itself is archival and is not updated by the EBI. Submitters are advised to update their own entries via the update form (http://www.ebi.ac.uk/embl/webin/update.html).

DATA IN THE EMBL NUCLEOTIDE SEQUENCE DATABASE

Data in the EMBL Nucleotide Sequence Database are grouped into divisions, according to either the methodology used in their generation (e.g. EST and HTG divisions) or taxonomic origin of the sequence source (e.g. HUM and PRO divisions). There are also some specialized entry types.

Whole Genome Shotgun (WGS) data

Methods using WGS data are used to gain a large amount of genome coverage for an organism. The sequences of all contigs originating from one experiment are grouped in a set. WGS entries have the standard EMBL format, with accession numbers clearly distinct from those of non-WGS entries. The accession numbers of all entries in each WGS set share the same prefix.

Third Party Annotation (TPA) data

The Third Party Annotation data set was launched in response to requests from the research community to submit entries that include either re-annotation of existing data, or combinations of novel sequence, existing primary sequence, trace archive and WGS data.

To distinguish TPA entries from primary data, the abbreviation 'TPA' appears at the beginning of each description (DE) line and in the keyword list. The link to the primary data information is given in the linetypes AH and AS that have been created for TPA entries. The following flatfile extract is taken from entry BN000024:

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	c

Constructed (CON) entries and expanded CONs

CON entries do not contain a sequence but an assembly of contigs, i.e. the sequence is to be constructed from segments of smaller sequences.

The format of a CON entry is similar to that of a standard entry, with the additional CO linetype to accommodate the assembly information. A CON entry does not have any annotation apart from source features.

The following example of an assembly is taken from entry BX470249:

```
CO join(BX640423.1:1..348251,BX640424.1:51. .349146,BX640425.1:51..348257,
CO BX640426.1:51..348866,BX640427.1:51..348997,BX640428.1:51..348525,
CO BX640429.1:51..344321,BX640430.1:51..348014,BX640431.1:51..347894,
CO BX640432.1:51..346301,BX640433.1:51..349305,BX640434.1:51..344805,
CO BX640435.1:51..346259,BX640436.1:51..255260)
```

Recently, the expanded forms of CON entries (CONFF) have been made available via SRS and ftp. In this format, the sequence defined by the assembly and the annotation of the segments are imposed onto the constructed sequence.

EMBLCDSs dataset

Following requests from database users, a new subset of EMBL data, the EMBLCDSs database, has been created during the last year. Every CDS (coding sequence) feature annotated in EMBL entries is displayed as a single entry.

More details are provided in the New Developments section below.

ACCESSING THE EMBL NUCLEOTIDE SEQUENCE DATABASE

The EMBL Nucleotide Sequence Database is available from the EBI via various WWW interfaces, ftp and email (for more information see http://www.ebi.ac.uk/embl/Access).

Sequence Retrieval System (SRS)

The EMBL Nucleotide Sequence Database can be accessed via the EBI SRS server (11,12) at http://srs.ebi.ac.uk/. In SRS, the data are available in the libraries shown in Table 1.

Table 1. SRS data libraries

Library	Content
EMBL	Entire EMBL Nucleotide Sequence Database apart from Contig and expanded Contig data
EMBL (Release)	The latest public release of the EMBL Nucleotide Sequence Database
EMBL (Updates)	All entries that are new or updated since the latest public release
EMBL (Third Party Annotation)	TPA data
EMBL (Contig)	CON entries
EMBL (Contigs Expanded)	Expanded CON entries
EMBL (Coding Sequences)	CDS data
EMBLALIGN (under 'Nucleotide related databases')	Alignment data

WGS data are not represented in a separate library any more, but is part of EMBL (Release) and EMBL (Updates). WGS entries can be identified via the keyword 'WGS'.

SRS also links to other databases, with cross-references to UniProt and publications available online, for example.

FTP Server

Release data, daily updates and cumulative files of all data types can be freely obtained from the ftp server at ftp://ftp.ebi.ac.uk/pub/databases/embl/. Please see the README file for further information.

To create and maintain a local copy of the cumulative file, the syncron tool (ftp://ftp.ebi.ac.uk/pub/software/unix/listtools/) can be used to download automatically newly available incremental data files from the ftp site and to merge them locally.

Dbfetch

Dbfetch (database fetch) is a tool for simple sequence retrieval via http. It can be used to retrieve up to 50 entries from various databases. Dbfetch can be found at http://www.ebi.ac.uk/cgi-bin/dbfetch.

Wsdbfetch provides programmatic access to the Dbfetch functionality. The service is described using Web Services Description Language (WSDL) and uses the Simple Object Access Protocol (SOAP) to communicate with other systems. For further information on Wsdbfetch please see http://www.ebi.ac.uk/Tools/webservices/WSDbfetch.html.

EMBL Sequence Version Archive

The EMBL Sequence Version Archive (SVA) (13) is a repository of all versions of any entry that have been distributed to the public from the EMBL Nucleotide Sequence Database. An interactive web-based interface to the SVA can be accessed at http://www.ebi.ac.uk/cgi-bin/sva/sva.pl.

Entries from the SVA can also be retrieved using dbfetch.

Completed genome sequences

Direct access to completely sequenced genomic components is available via the EBI Genomes server at http://www.ebi.ac.uk/genomes/. At the time of writing (September 2004) there are 162 completed genomes of bacteria, 19 archaea, 36 eukaryota, 540 organelles, 136 phages, 204 plasmids, 903 viruses and 36 viroids available.

Sequence searching

A comprehensive set of sequence analysis and database search algorithms is available at http://www.ebi.ac.uk/Tools/. The most commonly used algorithms available are FASTA (14) and WU-BLAST (15), permitting comparisons between query sequences and the nucleotide, translated nucleotide and protein databases.

Sequence similarity searches are available interactively over the WWW as well as by email. Instructions for email searches can be obtained by sending a message with the word HELP in its body to gpfasta@ebi.ac.uk.

Access via email

Data can also be retrieved by email using netserv (netserv@ebi.ac.uk). To get started send an email to netserv@ebi.ac.uk with 'HELP' in the message body.

NEW DEVELOPMENTS

Sequence length limit

In the past, the sequence length of a database record was limited to 350 000 bp. In June 2004, this restriction was lifted and entries of any length are now permitted in the database. Complete genomic units such as entire chromosomes can now be represented in a single entry. To represent unsequenced gaps, the new 'gap' feature is used. Some genomes that were split in the past in order to comply with the 350 000 bp limit have now been updated into single entries, e.g. AE000516.

Third Party Annotations—new rules

Following a decision taken at the 2004 Collaborative Meeting, the INSD Collaboration has increased the stringency for acceptance of data into the TPA dataset. The aim is to ensure that the TPA dataset includes the highest quality sequence and biological annotation.

To achieve this aim, the similarity between the TPA sequence and the contributing primary sequences is checked at the time of submission. We aim to achieve a similarity of at least 90%. In addition, there can be no more than 50 bp of the TPA sequence that does not correspond to primary entry(ies). All TPA records are manually curated and checked prior to public release.

To be released into the public TPA dataset, entries must also meet the following requirements:

(i) The study must have been published in a peer-reviewed journal.
(ii) The study must be supported by biological experimental evidence.

Further details may be found at: http://www.ebi.ac.uk/embl/Documentation/third_party_annotation_dataset.html. and http://www.ebi.ac.uk/webin/webin_help.html.

EMBL Sequence Version Archive—extended functionality

In February 2004, a new 'batch retrieval' functionality has been added to the SVA. Multiple entries can now be retrieved by supplying a list of accession numbers with either entry version number, sequence version number (user-indicated in the interface) or no version details for the most recent entry.

By the end of 2004, expanded CON entries will be included in the SVA.

A warning has been added to report the suppression date for entries that have been suppressed in the database.

EMBLCDSs dataset

Following requests from database users, a new subset of EMBL data, EMBLCDSs database, has been created during the year. Every CDS (coding sequence) feature annotated in EMBL entries is displayed as a single entry.

Entries are presented in an EMBL-like flatfile format, with addition of new line types (Figure 1).

The primary identifier of the entry given in the ID line is the protein_id of the CDS feature, the IV (identifier version) line gives protein_id and version. The accession number and

```
ID   CAD19988 standard; genomic DNA; FUN; 1839 BP.
XX
IV   CAD19988.1
XX
PA   AJ426417.1
XX
DE   Gibberella fujikuroi carotene cyclase
XX
OS   Gibberella fujikuroi
OC   Eukaryota; Fungi; Ascomycota; Pezizomycotina; Sordariomycetes;
OC   Hypocreomycetidae; Hypocreales; Nectriaceae; Gibberella;
OC   Gibberella fujikuroi complex.
OX   NCBI_TaxID=5127;
XX
FH   Key                     Location/Qualifiers
FH
FT   CDS                     join(AJ426417.1:801..823,AJ426417.1:874..1416,
FT                           AJ426417.1:1470..1643,AJ426417.1:1692..2790)
FT                           /db_xref="GOA:Q8X0Z1"
FT                           /db_xref="UniProt/TrEMBL:Q8X0Z1"
FT                           /gene="carRA"
FT                           /product="carotene cyclase"
FT                           /function="essential role in carotenoid biosynthesis"
FT                           /protein_id="CAD19988.1"
FT                           /translation="MGWEYAQVHLKYTIPFGVVLAAVYRPLMSRLDVFKLVFLITVSFF
...
XX
SQ   Sequence 1839 BP; 444 A; 433 C; 424 G; 538 T; 0 other;
     atgggctggg aatatgccca agtgcacctg aaatacacga taccgtttgg tgttgttttg 60
     gcggcggttt acagaccgtt gatgtcacgg ctggatgttt ttaagcttgt gtttttgata 120
...
//
```

Figure 1. A sample entry from the EMBLCDSs dataset.

sequence version of the parent EMBL entry can be found in the PA line. The DE line is created automatically and comprises the organism and product names. The taxonomic information is taken from the parent entry. The CDS annotation itself contains all qualifiers that belong to the feature, nucleotide locations being given in relation to the parent entry(ies). The nucleotide sequence of the feature is shown last in the entry.

The EMBLCDSs dataset is available via SRS [library: EMBL (Coding Sequences)] and ftp (ftp://ftp.ebi.ac.uk/pub/databases/embl/cds).

Finishing whole genome shotgun sets

Data from the WGS projects where the sequencing and assembling process is finished are moved into the main section of the database. At the time of writing only 5 out of 120 relatively small projects have been finished (example: *Nanoarchaeum equitans Kin4-M*, WGS project prefix: AACL, newly created entry in the main section: AE017199). In all cases, accession numbers of the WGS entries are added as secondary accession numbers to newly created entries in the main section to help track the data.

XML format

The International Nucleotide Sequence Database Collaboration INSDC has adopted a first draft for a common XML format for nucleotide data. The DTD can be found at http://www.ebi.ac.uk/embl/Documentation/DTD/INSDSeq_v1.dtd.txt.

CITING THE EMBL NUCLEOTIDE SEQUENCE DATABASE

The preferred form for citation of the EMBL Nucleotide Sequence Database is: Kanz,C. *et al*. (2005) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res*., 33, D29–D33.

CONTACTING THE EMBL DATABASE

Contact by email: data submissions: datasubs@ebi.ac.uk; other enquiries: datalib@ebi.ac.uk; data 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.
Telephone: data submissions, +44 1223 494499; general, +44 1223 494444.
Fax: general, +44 1223 494468.

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. Miyazaki,S., Sugawara,H., Ikeo,K., Gojobori,T. and Tateno,Y. (2004) DDBJ in the stream of various biological data. *Nucleic Acids Res*., **32**, D31–D34.
3. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2004) GenBank: update. *Nucleic Acids Res*., **32**, D23–D26.
4. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
5. 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.
6. Golovin,A., Oldfield,T.J., Tate,J.G., Velankar,S., Barton,G.J., Boutselakis,H., Dimitropoulos,D., Fillon,J., Hussain,A., Henrick,K. *et al*. (2004) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res*., **32**, D211–D216.
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. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res*., **32**, D434–D437.
9. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudai,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P. and Apweiler,R. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
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. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server—new features. *Bioinformatics*, **18**, 1149–1150.
12. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server—recent developments. *Bioinformatics*, **18**, 368–373.
13. Leinonen,R., Nardone,F., Oyewole,O., Redaschi,N. and Stoehr,P. (2003) The EMBL sequence version archive. *Bioinformatics*, **19**, 1861–1862.
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.