D16–D20 *Nucleic Acids Research, 2007, Vol. 35, Database issue* *Published online 5 December 2006*
doi:10.1093/nar/gkl913

EMBL Nucleotide Sequence Database in 2006

Tamara Kulikova*, Ruth Akhtar, Philippe Aldebert, Nicola Althorpe, Mikael Andersson, Alastair Baldwin, Kirsty Bates, Sumit Bhattacharyya, Lawrence Bower, Paul Browne, Matias Castro, Guy Cochrane, Karyn Duggan, Ruth Eberhardt, Nadeem Faruque, Gemma Hoad, Carola Kanz, Charles Lee, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Dariusz Lorenc, Hamish McWilliam, Gaurab Mukherjee, Francesco Nardone, Maria Pilar Garcia Pastor, Sheila Plaister, Siamak Sobhany, Peter Stoehr, 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, 2006; Revised and Accepted October 16, 2006

ABSTRACT

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl) at the EMBL European Bioinformatics Institute, UK, offers a large and freely accessible collection of nucleotide sequences and accompanying annotation. The database is maintained in collaboration with DDBJ and GenBank. Data are exchanged between the collaborating databases on a daily basis to achieve optimal synchrony. Webin is the preferred tool for individual submissions of nucleotide sequences, including Third Party Annotation, alignments and bulk data. Automated procedures are provided for submissions from large-scale sequencing projects and data from the European Patent Office. In 2006, the volume of data has continued to grow exponentially. Access to the data is provided via SRS, ftp and variety of other methods. Extensive external and internal cross-references enable users to search for related information across other databases and within the database. All available resources can be accessed via the EBI home page at http://www.ebi.ac.uk/. Changes over the past year include changes to the file format, further development of the EMBLCDS dataset and developments to the XML format.

INTRODUCTION

The EMBL Nucleotide Sequence Database is the European node of the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/) between DDBJ (1), EMBL and GenBank (2). The collaborative aim is to collect and present nucleotide sequence and annotation as comprehensively as possible.

The EMBL Nucleotide Sequence Database (EMBL) is maintained at the European Bioinformatics Institute, which hosts several other core biological databases (3).

The main goal of the EMBL Nucleotide Sequence Database is to accept, process and make freely available sequence data from individual researchers, research groups and the European Patent Office (EPO). Collected nucleotide sequences and accompanying annotation are made available via the EBI Sequence Retrieval System (SRS), ftp, web services and similarity search tools.

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

The database is presented as individual entries, each carrying sequence or information on sequence construction, submission information (submission and update dates, version numbers and submitter details), literature citations and annotation in the form of a feature table. Full details of database flatfile format are available in the user manual. Details of feature table format are available in the INSDC Feature Table Definition. Data are also presented in XML formats via the web tools, dbfetch and ftp.

Each entry in the database belongs to one of the several entry types, which differ in either data format or handling of data by the database. Entry types include standard (STD), constructed (CON), third party annotation (TPA), whole genome shotgun (WGS), annotated constructed (ANN) and mass genome annotation library (MGA). New entry types are created as new types of data arrive at the database.

Over the past year, the size of the EMBL Nucleotide Sequence Database has increased from 58.7 million entries in Release 84, September 2005 to 80.5 million entries in Release 88, September 2006, of which 18 million entries are WGS data. The WGS entries now account for >50% of the nucleotide content of the database—80.3 Gbp out of 146.5 Gbp in September 2006. There are now over 260 000 organisms represented in the database.

*To whom correspondence should be addressed. Tel: +44 01223 494463; Fax: +44 1223 494468; Email: kulikova@ebi.ac.uk

During the last year, an important EMBL flatfile format change was completed and there were further developments to XML formats, XML distribution and tools and the TPA dataset.

A detailed and up-to-date description of EMBL Nucleotide Sequence Database activities can be found at http://www.ebi.ac.uk/embl/; a list of relevant URLs is presented in Table 1.

DATA COLLECTION

Sequence submission

EMBL database submission procedures are briefly described below. Full details of procedures are available at 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 data) and is available at http://www.ebi.ac.uk/embl/Submission/webin.html. Bulk data submission in the fasta format is possible via Webin, where the fasta format is sufficient to describe all differences between submitted entries in terms of sequence and annotation fields.

TPA submissions are accepted via Webin; a modification of Webin is also available that is able to accept alignment submissions for inclusion into the EMBL-Align dataset (4). This service is available at http://www.ebi.ac.uk/embl/Submission/align_top.html.

Genome project submissions

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

EPO data processing

Sequence data extracted from biotechnology patent application submissions to the EPO are received, processed and made available weekly in the EMBL Nucleotide Sequence Database. A stable link between the patent document number, the sequence number within the document and the accession number is maintained. The EMBL Nucleotide Sequence Database processes both nucleotide and protein sequences from the EPO, but the distribution methods, collaborative data exchange mechanisms and exchange frequency for protein sequences differ from those of nucleotide sequences.

Data acquisition via data exchange

All new and updated database records are exchanged on a daily basis between EMBL, DDBJ and GenBank. WGS datasets are exchanged when they become available or have been updated and the rest of the data are exchanged daily. In addition to data exchange, lists of accession numbers are exchanged weekly to achieve maximum synchrony in data availability at all three sites.

Table 1. Relevant URLs and emails for EMBL nucleotide sequence database

Access	URL of emails	Comments
Submissions		
New submissions	http://www.ebi.ac.uk/embl/Submission/webin.html	For direct submissions of small-scale sequencing projects, bulk data (e.g. rRNA and EST), large genomes, TPA, etc.
Updates	http://www.ebi.ac.uk/webin/update.html	For updates to existing entries
Project accounts and WGS submissions	datasubs@ebi.ac.uk	Contact database to request project account or WGS submission
Retrieval		
SRS	http://srs.ebi.ac.uk	Data retrieval by term search and through links to/from other databases
Homology search	http://www.ebi.ac.uk/Tools/similarity.html	Data retrieval by sequence similarity and homology
SVA	http://www.ebi.ac.uk/cgi-bin/sva/sva.pl	Access to current and historic data by accession number or protein_id
FTP	ftp://ftp.ebi.ac.uk/pub/databases/embl/	Access to release, update, EMBL CDS, etc. data in the flatfile format and XML format
Genomes	http://www.ebi.ac.uk/genomes/	Completed genomes, links out to proteomes, Integr8 data, etc.
Dbfetch	http://www.ebi.ac.uk/cgi-bin/emblfetch	Retrieval by accession number through web browser
Wsdbfetch	http://www.ebi.ac.uk/Tools/webservices/WSDbfetch.html	Retrieval by accession number through web service
Netserv	netserv@ebi.ac.uk	Data via email
Access via map	http://www3.ebi.ac.uk/Services/EMBLWorld/EMBLWorld.pl	Geographical origin of sequenced samples
Custom datasets	datasubs@ebi.ac.uk	Request a datasets not yet provided in the course of normal productions
General		
General information	http://www.ebi.ac.uk/embl/	Website, including all documents
News	http://www.ebi.ac.uk/embl/News/news.html	Database news
Forthcoming changes	http://www.ebi.ac.uk/embl/Documentation/forthcomingchanges.html	Forthcoming data and format changes
Database statistics	http://www3.ebi.ac.uk/Services/DBStats/	Various statistics, updated daily
XML documentation	http://www.ebi.ac.uk/embl/Documentation/xml/	INSDC and EMBL documentation
Specific help	datasubs@ebi.ac.uk	

Data access

Main access method to EMBL Nucleotide Sequence Database data is SRS (5,6); the FTP server, homology search tools, the Genomes web server (for completely sequenced genomes) and sequence retrieval by accession number (Dbfetch, Wsdbfetch and netserv) are also available (7). Access to all versions, current and historical, of EMBL Nucleotide Sequence Database entries including CON, TPA and WGS data are available via the Sequence Version Archive, SVA (8).

In addition to these facilities that offer a range of ways to search and download data, there are several sites that mirror EMBL Nucleotide Sequence Database data, which provide distributed ftp access.

NEW DEVELOPMENTS

Important changes to the flatfile format

Since release 87 (JUN-2006) the format of the EMBL flat file has undergone a change: the ID line now has a different structure (see below) and the SV line has been removed.

The changes to the ID line structure were as follows:

All tokens are separated by a semicolon, the entry name is not displayed (in its place there will be the primary accession number), the sequence version is indicated in the ID line, the topology is a distinct token and is indicated for both circular and linear molecules and both the data class and the taxonomic divisions are displayed.

Below is an example of the new ID line:

```
ID CD789012; SV 4; linear; genomic DNA; HTG; MAM; 500 BP.
     [1]      [2]   [3]        [4]      [5]  [6]   [7]
```

The tokens represent:

[1] Primary accession number; [2] 'SV' + sequence version number; [3] Topology: 'circular' or 'linear'; [4] Molecule type; [5] Data class: ANN, CON, PAT, EST, GSS, HTC, HTG, MGA, WGS, TPA, STS, STD, 'normal' entries have have 'STD' for 'standard'; [6] Taxonomic division: HUM, MUS, ROD, PRO, MAM, VRT, FUN, PLN, ENV, INV, SYN, UNC, VRL, PHG'; [7] Sequence length + 'BP.'

An explanation of dataclass and taxonomic division, represented in the ID line by three-letter abbreviation, is available in the release notes.

The entry name is no longer displayed in the ID line. Since EMBL release 3 (December 1983), the stable identifier for an entry has been the primary accession number.

A mapping file (deprecated entry name to accession number) was provided via the ftp server for those entries where the entry name did not coincide with the accession number at the point of change.

Two other changes that are linked to the ID line change, both related to the way the data are represented on the ftp server: release data and the cumulative file (file containing all the data that are created or updated since the last release) are split into smaller files according to data class and taxonomic division. Full details on the way in which data are split on the ftp are available in the ftp directories and in the release notes.

XML development

In the past year, INSDC-specific XML was developed further; in spring 2006, the decision was taken to stabilize the production version of the DTD in order to facilitate external developments based on it. The current production version of the XML is INSDSeq v1.4 and can be obtained from http://www.insdc.org/documents.html.

Development of the EMBL-specific EMBLXML has continued and has been extended to EMBLCDS dataset. CDS are now distributed via the ftp server in the XML format in addition to the flatfile distribution. To support further the external use of the INSDC and EMBL XML formats, a web-based tool for instantaneous conversions between each XML and flatfile formats has been created.

EMBLCDS development

The EMBLCDS dataset was created in response to user requests for whole database dumps of coding sequence. EMBLCDS is now offered as a dataset updated daily, available by anonymous FTP, via SRS and via sequence similarity searches. There are currently 5.4 million EMBLCDS entries and 4.8 million items in the non-redundant EMBLCDSnr. To produce the non-redundant dataset, sequence checksums are used to collapse sequences with the same checksum into a single record.

Over the past year, several ways of grouping entries within the EMBLCDS dataset, apart from the grouping by checksum, were introduced: groups by gene name, by species and by shared exons. Grouping indices are available from the ftp server and are used in SRS views to link related records together.

As mentioned earlier in the 'XML development' section, EMBLXML has been extended to cover data from the EMBLCDS dataset.

Access to the data by map

In 2005, the International Nucleotide Sequence Database Collaboration introduced the lat_lon (latitude-longitude) qualifier. The qualifier allows submitters to specify precisely where the sequenced specimen was collected. The data collected so far can now be seen plotted on the world map at http://www3.ebi.ac.uk/Services/EMBLWorld/EMBLWorld.pl (Figure 1).

Cross-references

The EMBL Nucleotide Sequence Database continued to extend the number and diversity of its cross-references to other databases. The number of cross-referenced databases was 27 in the September 2006 release and the number of individual cross-references was over 62 million.

Cross-referenced databases include UniProt (9), InterPro (10), GOA (11) and a few other major databases, along with more specific databases. The cross-referenced database GeneDB (http://www.genedb.org/), for example, holds the latest sequence data and annotation for organisms sequenced by the PSU (Pathogen Sequencing Unit) at The Wellcome Trust Sanger Institute.

'Intradatabase' cross-references where introduced in December 2005 and are internal to the EMBL database. They include EMBL-TPA, EMBL-ANN, EMBL-CON, EMBL-ALIGN and EMBL-JOIN and show some of relationships between the entries in the database that are otherwise

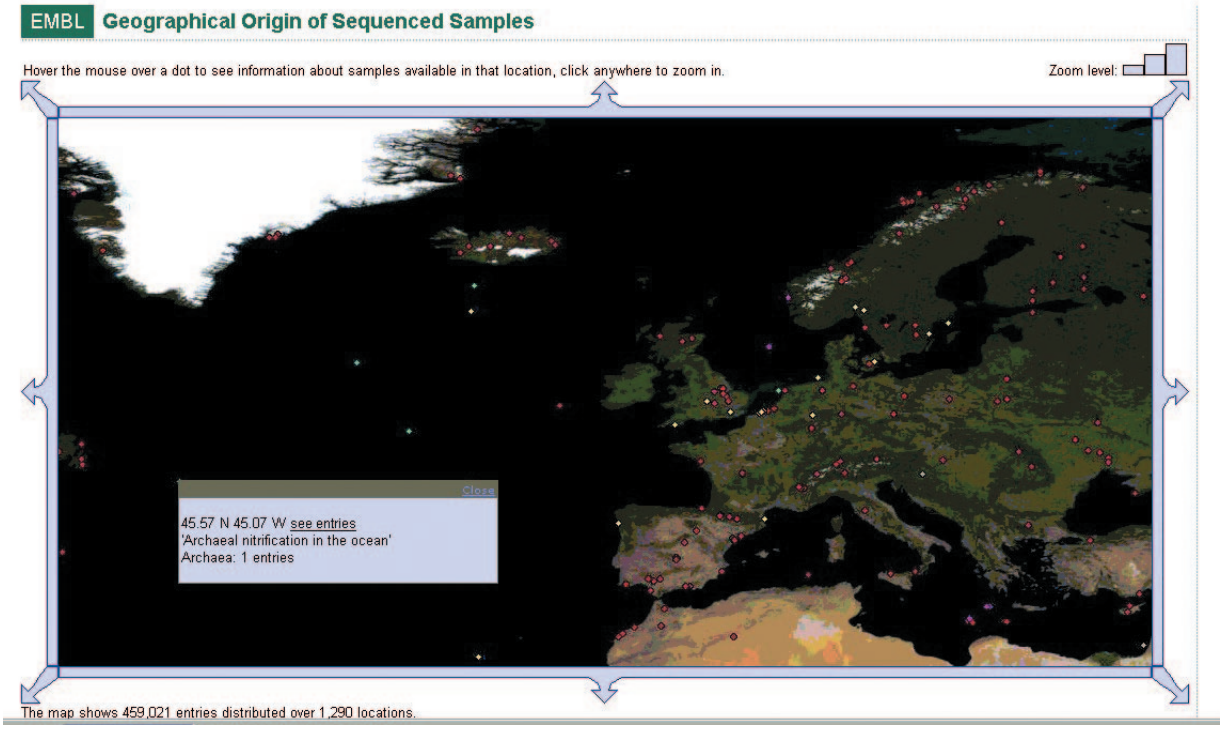

Figure 1. There are three levels of zoom to the map to allow viewing at greater magnification. Using the same geographical information, SRS views of EMBL entries link data to googlemaps.

difficult for users to infer; for example, EMBL-TPA cross-reference:

DR EMBL-TPA; BN000249.

will appear in a standard entry that serves as primary source for a TPA entry BN000249. Explanation for each type of the intradatabase cross-reference is given in the EMBL database release notes.

Further development of the TPA dataset

TPA records are submitted to the International Nucleotide Sequence Databases as part of the process of publishing biological studies that include the annotation of existing nucleotide sequences in the primary sequence database. Over the past year, the TPA dataset was divided into two tiers, TPA:experimental and TPA:inferential to distinguish between annotation supported by wet laboratory experimental evidence and inferred annotation, where the source molecule or its products have not been the subject of direct experimentation (12).

Enhanced evidence system

In order to enable users to see evidence for a particular annotation and make an informed judgment about its validity, the evidence tagging system was improved over the year. In place of the old qualifier 'evidence', two new qualifiers, 'experiment' and 'inference' were introduced in the course of the year. 'Experiment' value is a free text naming the experimental techniques used; 'inference' is a highly structured qualifier that details how the annotation was inferred. The structure of the qualifier is

TYPE[(same species)][:EVIDENCE_BASIS]

where TYPE is one of the following:

'non-experimental evidence, no additional details recorded'
'similar to sequence'
'similar to AA sequence'
'similar to DNA sequence'
'similar to RNA sequence'
'similar to RNA sequence, mRNA'
'similar to RNA sequence, EST'
'similar to RNA sequence, other RNA'
'profile'
'nucleotide motif'
'protein motif'
'ab initio prediction'

The optional text '(same species)' can be included when the inference comes from the same species as the entry.

The optional 'EVIDENCE_BASIS' is either a reference to a database entry (including accession and version) or an algorithm (including version), e.g. 'INSD:AACN010222672.1', 'InterPro:IPR001900', 'ProDom:PD000600', 'Genscan:2.0', etc.

A complete list of all features and qualifiers is available at http://www.ebi.ac.uk/embl/WebFeat/index.html.

The new evidence tagging system described above have been available since December 2005 and has at the time of writing been applied in 1662 entries, with over 145 000 instances of the new qualifiers containing meaningful values

(i.e. containing values different from ''[non-] experimental evidence, no additional details recorded'').

ACKNOWLEDGEMENTS

Funding to pay the Open Access publication charges for this article was provided by EMBL.

Conflict of interest statement. None declared.

REFERENCES

1. Okubo,K., Sugawara,H., Gojobori,T. and Tateno,Y. (2006) DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Res*., **34**, D6–D9.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.L. (2006) GenBank. *Nucleic Acids Res*., **34**, D16–D20.
3. Brooksbank,C., Camon,E., Harris,M.A., Magrane,M., Martin,M.J., Mulder,N., O'Donovan,C., Parkinson,H., Tuli,M.A., Apweiler,R. *et al*. (2003) The European Bioinformatics Institute's data resources. *Nucleic Acids Res*., **31**, 43–50.
4. 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.
5. Zdobnov,E.M., Lopez,R., Apweiler,R. and Etzold,T. (2002) The EBI SRS server-new features. *Bioinformatics*, **18**, 1149–1150.
6. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*., **266**, 114–128.
7. Harte,N., Silventoinen,V., Quevillon,E., Robinson,S., Kallio,K., Fustero,X., Patel,P., Jokinen,P. and Lopez,R. (2004) Public web-based services from the European Bioinformatics Institute. *Nucleic Acids Res*., **32**, W3–W9.
8. Leinonen,R., Nardone,F., Oyewole,O., Redaschi,N. and Stoehr,P. (2003) The EMBL SVA. *Bioinformatics*, **19**, 1861–1862.
9. Wu,C.H., Apweiler,R., Bairoch,A., Natale,D.A., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al*. (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res*., **34**, D187–D191.
10. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Binns,D., Bradley,P., Bork,P., Bucher,P., Cerutti,L. *et al*. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res*., **33**, D201–D205.
11. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*., **32**, D262–D266.
12. Cochrane,G., Bates,K., Apweiler,R., Tateno,Y., Mashima,J., Kosuge,T., Mizrachi,I.K., Schafer,S. and Fetchko,M. (2006) Evidence standards in experimental and inferential INSDC third party annotation data. *OMICS*, **10**, 105–113.