18–24 Nucleic Acids Research, 1999, Vol. 27, No. 1

The EMBL Nucleotide Sequence Database

Guenter Stoesser*, Mary Ann Tuli, Rodrigo Lopez and Peter Sterk

EMBL Outstation–The European Bioinformatics Institute (EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received October 1, 1998; Revised October 5 1998; Accepted October 16, 1998

ABSTRACT

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl.html) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications. While automatic procedures allow incorporation of sequence data from large-scale genome sequencing centres and from the European Patent Office (EPO), the preferred submission tool for individual submitters is Webin (WWW). Through all stages, dataflow is monitored by EBI biologists communicating with the sequencing groups. In collaboration with DDBJ and GenBank the database is produced, maintained and distributed at the European Bioinformatics Institute (EBI). Database releases are produced quarterly and are distributed on CD-ROM. Network services allow access to the most up-to-date data collection via Internet and World Wide Web interface. EBI's Sequence Retrieval System (SRS) is a Network Browser for Databanks in Molecular Biology, integrating and linking the main nucleotide and protein databases, plus many specialised databases. For sequence similarity searching a variety of tools (e.g. Blitz, Fasta, Blast etc) are available for external users to compare their own sequences against the most currently available data in the EMBL Nucleotide Sequence Database and SWISS-PROT.

INTRODUCTION

In Europe, the vast majority of the nucleotide sequence data produced is collected, organised and distributed by the EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl.html) located at the European Bioinformatics Institute (Cambridge, UK), an Outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany.

THE WELLCOME TRUST GENOME CAMPUS

Located in the grounds of the Wellcome Trust Genome Campus near Cambridge (UK) are three different institutes, the Sanger Centre, the UK Human Genome Mapping Project Resource Centre (HGMP-RC) and the European Bioinformatics Institute (EBI). The HGMP-RC is a body funded by the UK Medical Research Council to provide computing and biological services in the context of the Human Genome Program. The Sanger Centre constitutes Europe's major Genome Research Centre established jointly in 1992 by the Wellcome Trust and the Medical Research Council to provide a major focus for mapping and sequencing the human genome, and genomes of many other organisms. The European Bioinformatics Institute is part of the European Molecular Biology Laboratory (EMBL) and is funded by 15 European nations and Israel.

THE EUROPEAN BIOINFORMATICS INSTITUTE

The EBI is a research and service organisation serving academic research (in molecular biology, genetics, medicine and agriculture) as well as biotechnological, chemical, agricultural and pharmaceutical industries. The main missions of the EBI (1) centre around biological databases. In this respect a number of databases are operated, namely the EMBL Nucleotide Sequence Database, the Protein Databases (SWISS-PROT and TREMBL) (2), the Radiation Hybrid Database Rhdb (3) and the Macromolecular Structure Database (MSD). The EBI is engaged in an extensive program of applied research and development on software methods for integration and interoperation of biological databases.

THE EMBL NUCLEOTIDE SEQUENCE DATABASE

The EMBL Database in collaboration with GenBank (NCBI, Bethesda, USA) and the DNA Data Bank of Japan (NIG, Mishima) has been processing nucleotide sequence data reported world-wide since 1982. Genome Project data, direct submissions by individual scientists and patent sequence data from the European Patent Office are the main sources contributing to the ongoing database growth. To achieve optimal synchronisation all new and updated database records are exchanged between the International Nucleotide Sequence Collaboration on a daily basis.

EMBL Database entries are grouped into divisions. Divisions are subsets of the database reflecting the areas of interest of the user community and are based mainly on taxonomy (e.g. HUM = human, PLN = plants, PRO = prokaryotes, etc.) with a few exceptions like HTG (High Throughput Genome Sequences), GSS (Genome Survey Sequences) and EST (Expressed Sequence Tags) for which grouping is based on the specific nature of the underlying data.

*To whom correspondence should be addressed. Tel: +44 1223 494 466; Fax: +44 1223 494 472; Email: stoesser@ebi.ac.uk

Database entries are distributed in EMBL flat-file format which is supported by most sequence analysis software packages and also provides a structure usable by human readers. Detailed information on the flat-file format and line-types are provided in the EMBL Database USER MANUAL document available from the EBI network servers. Feature Tables within database entries describe the roles and locations of higher order sequence domains and elements within the genome of an organism. The Feature Table follows the unified DDBJ/EMBL/GenBank Feature Table Definition which is available from the EBI network servers: URL http://www.ebi.ac.uk/ebi_docs/embl_db/ft/feature_table.html

EMBL Database releases are produced quarterly and are distributed on CD-ROM. The most up-to-date data collection is available via Internet and World Wide Web interface.

Identifiers and integration with other databases

Interconnectivity between a growing number of biomolecular databases is becoming an essential prerequisite for utilising the wealth of information becoming available. Where appropriate, EMBL Database entries are cross-referenced to other databases like the Eukaryotic Promoter database (4) TRANSFAC (5), Flybase (6), TREMBL and SWISS-PROT. SWISS-PROT itself is linked to more than 30 different databases thus providing a focal point for database interconnectivity. Cross-references to external databases are represented in the Feature Table qualifier /db_xref. Additionally, protein identifiers can be used by external databases (such as SWISS-PROT) as an identifier onto which cross-references can be built at feature level, e.g., to individual CDS features. Protein identifiers are currently assigned to all CDS features in the nucleotide sequence database to identify an exact protein translation for a coding sequence and are found in the Feature Table qualifier /dbxref, e.g.,

```
FT   CDS    328. .1866
FT         /db_xref='PID:e191449'
```

This identifier remains the same as long as the translation remains the same. When a translation change occurs, however minor, a new PID value is assigned. Requirements for full functionality are: (i) identifiers are maintained collaboratively; (ii) identifiers are portable amongst databases; (iii) users can access exact versions without being tied to one of the collaborative databases.

Confusion resulted amongst users and other databases concerning the relationship between these identifiers and the GenBank 'gi' numbers which for example are additionally assigned to every CDS feature. To clarify this, and to adopt a comparable scheme of identifiers for both nucleotides and proteins, the collaborating databases DDBJ/EMBL/GenBank will introduce a new form of nucleotide and protein identifiers.

New identifiers

Both nucleotide and protein identifiers will consist of a stable part which will not change, and a version part which will be incremented whenever the underlying nucleotide sequence or protein translation changes. The new form of identifiers will allow easier tracking of changes to nucleotide and protein identifiers by external databases compared to the current identifiers.

Nucleotide sequence identifier. Currently, the line type 'NI' contains an identifier (e.g., e1344565) for each nucleic acid sequence. The value of this identifier will only change, when a change in the sequence occurs, while the accession-number on the AC line will remain unchanged.

The new nucleotide sequence identifier will be of the form:

'Accession.Version' (eg, Z86131.1),

where the accession number part will be stable, but the version part will be incremented when the sequence changes. A new linetype 'SV' (Sequence Version) will be introduced to represent this information. For example:

```
ID   DBSELBGEN    standard; DNA; PRO; 2196 BP.
AC   X99911;
SV   X99911.3
```

Protein sequence identifier. The new protein identifier (replacing PIDs, e.g., /db_xref="PID:e123345) will consist of a stable ID portion (3+5 format with 3 position letters and 5 numbers) plus a version number after a decimal point. For example:

/protein_id = 'CAA12345.6'

The version number will change only when the protein sequence coded by the CDS changes, while the stable part will remain unchanged. This qualifier will be valid only on CDS features which translate into a valid protein.

During the transition phase both the old and new forms of identifiers will be provided, e.g.,

```
FT   CDS  1124. .1939
FT        /db_xref='PID:g45266'
FT        /protein_id='CAA12345.6'
FT        /db_xref='SWISS-PROT:P29808'
FT        /gene='aacC3'
FT        /product='aminoglycoside-(3)-N-acetyl-
          transferase isoenzyme
FT        III'
FT        /translation='MTDLNIPHTHAHLVDAFQALGIRAGQAL
FT        MLHASVKAVGAVMGGPNVILQALMDALTPDGTLMMYAGWQDI
FT        PDFIDSLPDALKAVYLEQHPPFDPATARAVRENSVLAEFLRT
FT        WPCVHRSANPEASMVAVGRQAALLTANHALDYGYGVESPLAK
FT        LVAIEGYVLMLGAPLDTITLLHHAEYLAKMRHKNVVRYPCPI
FT        LRDGRKVWVTVEDYDTGDPHDDYSFEQIARDYVAQGGGTRGK
FT        VGDADAYLFAAQDLTRFAVQWLESRFGDSASYG'
XX
```

Implementation schedule. During the Collaborative Meeting held at the EBI in May 1998, the DDBJ/EMBL/GenBank Databases agreed on the conventions concerning protein_id assignments etc., and Nucleotide Sequence Versioning. Subject to synchronisation amongst the international databases we plan to introduce the new form of nucleotide and protein identifiers early in 1999.

Data Acquisition

Genome Project data

A direct dataflow to the EMBL Database from various international sequencing efforts exists to ensure immediate incorporation and distribution of new sequence data and descriptive information.

Particularly noteworthy is the collaboration on high-throughput data acquisition with the genome projects in the Sanger Centre, one of the most productive sequencing centres in Europe and world-wide.

Sanger Centre projects include human chromosomes 1, 6, 20, 22, X, other vertebrates (mouse, chicken, pufferfish), worm

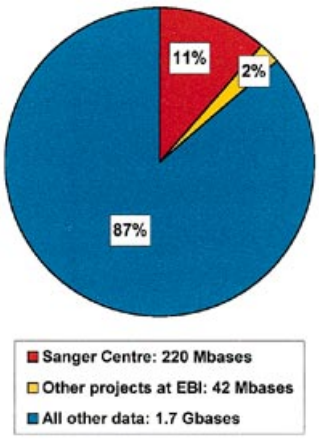

Figure 1. Contributions of Genome Projects at the EBI to the EMBL Database. The Sanger Centre is one of the world's most productive genome centres.

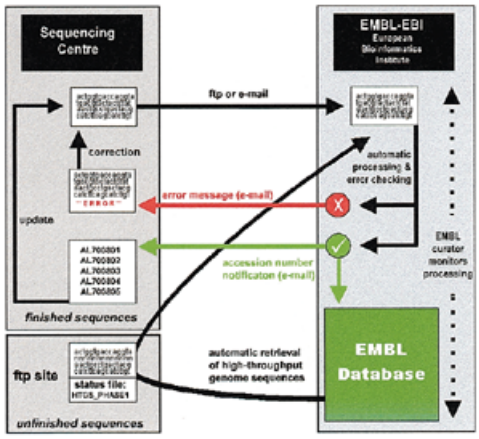

Figure 2. Submissions of genome project data. Data are actively submitted by the project to their account. Data are automatically processed daily. In addition to active submission by the sequencing centre, the EBI can retrieve unfinished sequence data (high-throughput sequences, HTG) from FTP sites. This is currently done for a number of Sanger Centre projects.

Caenorhabditis elegans, several yeasts, the fruitfly *Drosophila melanogaster*, several protozoa and microbes.

Projects from other sequencing centres include: *Fugu rubripes* GSS, HGMP-RC, UK; *Arabidopsis thaliana* ESSA project, European collaboration; Human EST, Padova, Italy; European *Drosophila* Mapping Consortium, Cambridge, UK; Mouse EST, Institute Pasteur, Paris, France (see Fig. 1).

How do Genome Project data get processed? The EMBL Database has developed automatic procedures to allow the direct submission and incorporation of genome sequences such that new projects can be accommodated easily. Through all stages, EBI biologists are communicating with the sequencing groups.

The exact procedure of data acquisition is dependent on whether the sequence data to be incorporated represents '(un)finished' or 'finished' sequence data.

Unfinished HTG data. A consortium of large scale sequencing centres and their funding agencies have reached a consensus agreement (the 'Bermuda Principles') regarding data produced in publicly funded projects. This agreement states that 'unfinished' sequence data be released as soon as it is 'useable' for homology searching and other types of sequence analysis. Based on these guidelines vast amounts of sequence data produced at sequencing centres, e.g., the Sanger Centre, are included into the EMBL Database as soon as they become available from the individual sequencing groups, and are immediately available for homology searches via the EBI network services.

High-throughput sequence records are included in the HTG division and contain keywords to indicate the status of the sequencing (e.g., HTGS_PHASE1).

HTGS_PHASE1: Sequence consists of an unordered set of sequence pieces (typically 7–20), unoriented, unannotated and containing gaps

HTGS_PHASE2: Sequence consists of sequence pieces (typically 2 or 3) for which order and orientation have been established, while gaps remain

HTGS_PHASE3: Sequence is considered to be completed and might contain (some) annotation.

'Unfinished' HTG data are automatically collected on a daily basis as FastA format files by the EBI from the Sanger Centre FTP server. By checking a status file containing all the necessary information to create or update an EMBL HTG entry (e.g., entryname, accession number in case of updates, chromosome number, list of authors, etc.) the sequences are retrieved from the server, flatfiles are automatically generated and loaded into the database. A single accession number is assigned to one clone, and as sequencing progresses and the entry passes from one phase to another, it will retain the same accession number. Because of the transient nature of these data, ongoing updates will occur after initial inclusion. For users, it is important to note that these data are unfinished and do not necessarily represent the correct sequence. Work on the sequence is in progress and the release of this data is based on the understanding that the sequence may change as work continues. Unfinished HTG records can be retrieved via EBI's Sequence Retrieval System (SRS).

On completion by the sequencing group, finished sequences are submitted to the database thus replacing the former 'unfinished' version of the sequence record (see below and Fig. 2).

Finished genome sequence data. 'Finished' genome sequence data are submitted actively by the sequencing groups to the appropriate submission account at the EBI. This applies for both re-submissions of former HTG records and initial completed project submissions. Genome sequencing groups producing large volumes of nucleotide sequence data over an extended period are encouraged to have submission accounts established. Each submission account is curated by EBI biologists. A submission protocol is agreed upon and database entries produced at the research site will be deposited and updated directly by the originating group via FTP or Email. The curator checks to ensure that new entries follow database annotation conventions and are consistent with other entries from the same project. This procedure has demonstrated itself to be flexible and efficient both for the research groups and for database staff. Groups wishing to establish a submission account with the EBI should contact database staff (see Appendix for contact information).

If loading of data is successful, database accession numbers will be assigned to the new entries and communicated to the according sequencing group with the notion that the accession

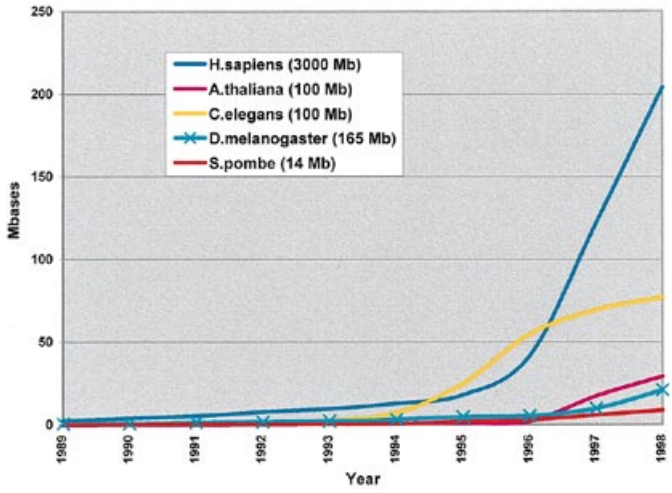

Figure 3. Genome Monitoring Table (Genome MOT). Cumulative progress plot: progress of the major eukaryotic genome projects.

number information should be included in subsequent publications. The corresponding EMBL database entries are made available immediately via the EBI's network services (see below).

If loading of data is unsuccessful (e.g., because of contents or syntax problems), no accession number is assigned, instead a flatfile with error messages is created and forwarded to the sequencing group. If necessary, the database curator will assist the project group in resolving the problems in order to resubmit.

Genome annotation

Initial submissions of genome data by sequencing projects include preliminary gene annotations based on gene prediction programs. The underlying information concerning methods used and matches found are currently not available to the user community. Given that a sequence might have a large number of matches to other database entries and might be run through several different algorithms using different parameters, the sheer quantity of information is currently considered to be beyond the scope of the Feature Table.

Additionally, sequencing groups may not maintain the annotation after a sequencing project has finished. Discussions between representatives of the major species-specific databases and EMBL suggested that genome annotation should be maintained by the community—in particular the community databases (e.g., Flybase, SGD etc.). The sequence database can then link via /db_xref from a given CDS feature to the according external up-to-date annotation and methods/matches database information.

Genome Monitoring Table (MOT). An effort is being made to monitor the progress of a number of large genome sequencing projects. A genome monitoring table (Genome MOT) showing the total amount of finished and unfinished genomic DNA sequences deposited per year into the DDBJ/EMBL/GenBank databases for a number of organisms is updated on a weekly basis and can be found at URL: http://www.ebi.ac.uk/~sterk/genome-MOT/ (Fig. 3).

Direct submissions

Submission of sequence information to a database prior to publication has become standard practice. A unique accession number is assigned by the database which permanently identifies the sequence submitted. The database accession number should be included in the manuscript, preferably in a footnote on the first page of the journal article, or as required by individual journal procedures. This procedure ensures availability and distribution of new sequence data in a timely fashion.

Data confidentiality and release dates. Sequences submitted to the database can be released to the public either immediately or withheld until an author-specified date.

Confidential data are never withheld after publication. In general, unless otherwise directed by the author, submitted sequences are available to the research community several months before these sequences appear in a journal publication.

Submitting and updating sequences via WWW

The preferred submission and update mechanisms (for sending new and modifying existing sequence data and descriptive information) to the EMBL Database are WWW-based systems.

WEBIN. The Web-based sequence submission system 'Webin' is EBI's preferred submission medium. Webin is available from http://www.ebi.ac.uk/submission/webin.html

Webin allows submission of sequence data and descriptive information by navigating the user through a series of WWW forms in an interactive and straightforward way. Information required to create a database entry: (i) submitter information; (ii) release date information; (iii) sequence data, description and source information; (iv) reference citation information; (v) feature information (e.g., coding regions, regulatory signals etc.). Submitters are able to modify and view their data prior to submission in the format in which it will be finally published in the EMBL database.

Database programmers and curators are continuously making improvements to Webin based on feedback from users and internal input, with the aim of further automisation of the submission procedure.

Checking sequences for vector contamination. To assist submitters the EBI now provides a vector screening service using the latest implementation of the BLAST algorithm and a special sequence databank known as EMVEC.

EMVEC is an extraction of sequences from the SYNthetic division of EMBL containing more than 2000 sequences commonly used in cloning and sequencing experiments. EMVEC is by no means a complete vector databank but EBI believes it is representative of the kind of material used in modern sequencing and should be useful to submitters. The databank will be updated with each release of EMBL and made publicly available on the EBI's FTP (ftp.ebi.ac.uk) server for those who wish to have it.

The interactive WWW service can be found at:
http://www.ebi.ac.uk/submission/webin.html
http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/databasehome.html
http://www2.ebi.ac.uk/blastall/vectors.html

The results will list sequences producing significant alignments and associated information like vector name, score, alignment etc. We suggest that you remove vector contamination from your sequence data before submitting to the database.

WEBUP. The responsibility of keeping a database entry up to date lies with the original submitting group. Submitters are encouraged to report changes to the sequence, features, gene or product nomenclature and to send full citation information to the database when their sequence data is published. Updates reported to the database from another party must be approved by the original submitter. The WWW-based update form is available from the EBI at URL: http://www.ebi.ac.uk/ebi_docs/update.html

The EBI is currently developing a new WWW based update system which will enable the user to update their entry interactively. Users will be able to obtain a current version of a public database entry from the EBI's CORBA servers, modify their data and send it back to EMBL as a loadable flatfile. A different version of the tool will allow users to update confidential entries.

Data will be validated during the process in a similar manner to Webin, but unlike Webin the new update system is being written in Java and uses CORBA technology.

SEQUIN. Sequin is a multi-platform (Mac/PC/Unix) stand-alone software tool developed by the NCBI for submitting entries to the EMBL, GenBank or DDBJ sequence databases. The Sequin program, along with detailed downloading and installation instructions, plus general information is available from the EMBL Database via WWW browser and anonymous FTP. Sequin is available from http://www.ebi.ac.uk/subs/allsubs.html ; ftp://ftp.ebi.ac.uk/pub/software/sequin/

Submitting and updating sequences via Email

A computer-readable data submission form is available from the EBI: (i) from the EBI WWW-server http://www.ebi.ac.uk/subs/allsubs.html ; (ii) by electronic mail via the EBI fileserver (netserv@ebi.ac.uk); (iii) with all releases of the EMBL Nucleotide Sequence Database. Please note that we intend to discontinue use of the Email submission form when we are confident that all users have WWW access. A computer-readable update form is available by Email upon request or by anonymous FTP: Email: update@ebi.ac.uk; FTP: ftp://ftp.ebi.ac.uk/pub/databases/embl/release/update.doc

Bulk submissions

Authors planning to submit a large number of similar sequences (i.e., >25) on a single occasion are encouraged to contact the database before submitting the data. Database staff will then assist in making the submission of this specific data as convenient as possible, thus saving the author the time and effort required to complete numerous submission events individually.

Sequence alignment submissions

Sequence alignment data (e.g. from phylogenetic analysis of nucleotide or amino-acid sequences) can also be deposited at the EBI. An alignment-number (e.g. DS8200) will be assigned to the data by the database and it is suggested that this number is printed in the according publication. Alignment data and additional information are available via the EBI FTP and File servers: (i) EBI FTP server: by anonymous FTP from FTP.EBI.AC.UK in directory /pub/databases/embl/align; (ii) EBI File server: by sending an Email message to netserv@ebi.ac.uk including the line HELP ALIGN or GET ALIGN:DS8200.DAT; (iii) EBI WWW pages ftp://ftp.ebi.ac.uk/pub/databases/embl/align/

Currently a compilation of text files, the issue of format standardisation has been discussed by the database staff, external users and experts in the field. There is a wide spectrum of opinions concerning possible alignment formats, but further standardisation (e.g., Nexus file formats) will constitute a major enhancement. (7)

DATA ACCESS

EBI network services

Data access to sequence data at the EBI is granted via Email using the netserver (8) or interactively via the WWW where the main service is composed of an SRS server (9). Databases as well as software can be downloaded from the EBI's FTP server (ftp.ebi.ac.uk).

The netserver understands a series of commands which allow the retrieval of sequence data as well as the content of certain directories on the EBI's FTP server. A user guide can be obtained by sending an Email message with the word 'help' in the body of the message to netserv@ebi.ac.uk.

The Sequence Retrieval System (SRS; 9) server at the EBI contains a comprehensive collection of databanks. The server can be accessed using a suitable WWW browser with support enabled for Java and JavaScript. The SRS server's URL is: http://srs.ebi.ac.uk/

Database searching

The EBI provides a comprehensive set of sequence database searching algorithms. Both interactive and Email submissions are possible for searching EMBL and SWISS-PROT and their inter-release updates as well as PDB. Databases derived from these are also available [i.e., TREMBL and SWALL (aka sp_tr_nrdb)]. EMBL is available for searches as individual taxonomical divisions or as a whole.

Specialised databanks are also available. The search of complete sequence databanks sometimes results in the masking of desirable hits. This occurs when the user intends to compare a known DNA or protein sequence of, for example, an immunoglobin, an HLA sequence region or a GPC receptor component against all other known sequences of the same type but searches using a databank that contains a myriad of sequences from different organisms, many unrelated genes and protein families. For this reason EBI makes searches available of IMGT (10), GPCRDB (11) and EVEC (Lopez *et al*. 1998, EMBnet News Vol. 5.3), which are specialised subsets derived from EMBL or SWISS-PROT with improved annotation and expert support.

The most commonly used algorithms available from EBI are: Fasta3 (12), WU-Blast (13) and NCBI-Blast2 (14). More specialised search methods comprising the BLITZ services, which build around various implementations of the rigorous Smith and Watermann (15) algorithm, are available for protein databank scanning. These are: Compugen's Bic2, Scanps (SCANPS version 2.3.1: Geoffrey J. Barton, University of Oxford, UK) and Ssearch3. Users should refer to the following Email addresses to submit and obtain help by sending an Email to the appropriate address with the word 'HELP' in the body of the message: fasta@ebi.ac.uk; blast@ebi.ac.uk; blitz@ebi.ac.uk.

Table 1. Sites maintaining daily updated copies of the EMBL Nucleotide Sequence Database

AUSTRALIA - ANGIS http://www.angis.su.oz.au/ Contact: Tim Littlejohn Email: tim@angis.su.oz.au Tel: 1800 728 028 or 9531 2948 Fax: 02 9351 5694	**AUSTRIA - Vienna Biocenter** http://www.at.embnet.org/ Contact: Martin Grabner Email: martin.grabner@cc.univie.ac.at Tel: +43-1-79515 / 6108 Fax: +43-1-7986224
BELGIUM - BEN, the Belgian EMBnet Node http://www.be.embnet.org/ Contact: Robert Herzog Email: rherzog@ulb.ac.be Tel: +32-2-6509762 Fax: +32-2-6509767	**CANADA -** http://www.cbr.nrc.ca Contact: Christoph W. Sensen Email: sensencw@niji.imb.nrc.ca Tel: + -902-426-7310 Fax: + -902- 426-9413
CHINA http://www.cbi.pku.edu.cn/ Contact: Jingchu Luo Email: office@cbi.pku.edu.cn	**DENMARK - BioBase** http://biobase.dk/ Contact: Hans Ullitz-Moeller Email: hum@biobase.dk Tel: +45-89-422846 Fax: +45-89-131160
ETI - Expert Center for Taxonomic Identification http://www.eti.uva.nl Contact: Peter H. Schalk Email: pschalk@eti.uva.nl Tel: +31-20-5257239 Fax: +31-20-5257238	**FINLAND - CSC Center for Scientific Computing** http://www.csc.fi/molbio/ Contact: Erja Heikkinen Email: Erja.Heikkinen@csc.fi Tel: +385-0-4572078 Fax: +385-0-4572302
FRANCE - INFOBIOGEN http://www.infobiogen.fr/ Contact: Philippe Dessen Email: dessen@infobiogen.fr Tel: +33-1-45595241 Fax: +33-1-45595250	**GERMANY - GENIUSnet** http://genome.dkfz-heidelberg.de/ Contact: Martin Ebeling Email: M.Ebeling@dkfz-heidelberg.de Tel: +49-6221-422349 Fax: +49-6221-422333
GREECE - IMBB http://www.imbb.forth.gr/ Contact; Babis Savakis Email: savakis@myia.imbb.forth.gr Tel: +30-81-212647 Fax: +30-81-231308	**HUNGARY - Agricultural Biotechnology Center** http://www.abc.hu/ Contact: Endre Barta Email: barta@hubi.abc.hu Tel: +36-28-330127 Fax: +36-28-320096
IRELAND - INCBI http://acer.gen.tcd.ie/ Contact: Andrew Lloyd Email: atlloyd@acer.gen.tcd.ie Tel: +353-1-6081969 Fax: +353-1-6798558	**ISRAEL - Israeli National Node** http://dapsas1.weizmann.ac.il/ Contact: Leon Esterman Email: lsestern@weizmann.weizmann.ac.il Tel: +972-8-343934 Fax: +972-8-344113
ITALY - CNR Area di Ricerca http://area.ba.cnr.it/ Contact: Marcella Attimonelli Email: attimonelli@area.ba.cnr.it Tel: +39-80-5482130 Fax: +39-80-5484467	**THE NETHERLANDS - CAOS/CAMM Center** http://www.caos.kun.nl/ Contact: Jan H. Noordik Email: noordik@caos.kun.nl Tel: +31-24-3653386 Fax: +31-24-3652977
NORWAY - The Norwegian EMBnet Node http://www.no.embnet.org/ Contact: Linda Akselberg Email: Linda.Akselberg@biotek.uio.no Tel: +47-22-958756 Fax: +47-22-694130	**POLAND - IBB** http://www.ibb.waw.pl/ Contact: Piotr Zielenkiwicz Email: piotr@ibbrain.ibb.waw.pl Tel: +48-2-6584703 Fax: +48-39-121623
PORTUGAL - PEN Portuguese EMBnet Node http://www.pen.gulbenkian.pt/ Contact: Pedro Fernandes Email: pfern@pen.gulbenkian.pt Tel: +351-1-4431408 Fax: +351-1-4435625	**RUSSIA - Russian EMBnet Node** http://www.genebee.msu.su/ Contact: Vladimir P. Skulachev Email: skulach@head.genebee.msu.su Fax: +7 -95)-39-0338
SPAIN - Centro Nacional de Biotecnologia http://www.cnb.uam.es/ Contact: Jose-Maria Carazo Email: carazo@samba.cnb.uam.es Tel: +341-585-4543 Fax: +341-585-4506	**SWEDEN - Biomedical Center** http://www.bmc.uu.se/ Contact: Nils E. Eriksson Email: embnetadm@perrier.embnet.se Tel: +46-18-174016 Fax: +46-18-551759
SWITZERLAND - Swiss EMBnet National Node http://www.ch.embnet.org/ Contact: C.Victor Jongeneel Email: Victor.Jongeneel@isrec.unil.ch Tel: +41-61-2672247 Fax: +41-61-2672078	**TURKEY - Turkish EMBnet Node** http://www.rigeb.gov.tr Contact: Muzaffer Taylan Email: taylan@rigeb.gov.tr Tel: +90-262-6412300 ext. 4007 Fax:+90-262-6412309
UK Human Genome Mapping Project http://www.hgmp.mrc.ac.uk/ Contact: Administration Email: admin@hgmp.mrc.ac.uk Tel: +44-1223-494500 Fax: +44-1223-494512	**UNITED KINGDOM - SEQNET Daresbury Laboratory** http://www.dl.ac.uk/ Contact: Alan Bleasby Email: bleasby@dl.ac.uk Tel: +44-1952-603351 Fax: +44-1952-603100
ICGEB Area Science Park http://www.icgeb.trieste.it/ Contact: Sandor Pongor Email: pongor@icgeb.trieste.it Tel: +39-40-3757300 Fax: +39-40-226555	

All the above database searching facilities are available in interactive mode by using a suitable WWW browser and going to the following URL's: http://www2.ebi.ac.uk/fasta3/ ; http://www2.ebi.ac.uk/blast2/ ; http://www2.ebi.ac.uk/blastall/ ; http://www2.ebi.ac.uk/bic_sw/ ; http://www2.ebi.ac.uk/ppsearch/

Sequence analysis

The EBI makes available highly specialised sequence analysis programs. Such services include multiple sequence alignment and inference of phylogenies using clustalw (16), Gene prediction using GeneMark (17), pattern searching and discovery using PRATT (18), as well as applications which have been developed in-house for various projects which the general user community should find useful.

EMBnet

The European Molecular Biology Network (http://www.embnet.org) was initiated in 1988 to link European laboratories using biocomputing and bioinformatics in molecular biology research as well as to increase the availability and usefulness of the molecular biology databases within Europe. Remote copies of the nucleotide and protein sequence databases, updated daily, as well as other molecular biology resources, are held at nationally mandated nodes. As bioinformatics grows, EMBnet plays an important role in support, training, research and development for the European bioinformatics research community. Table 1 gives a full listing of sites maintaining daily updated copies of the EMBL Database.

Citing the EMBL Database. The preferred form for citation of the EMBL Nucleotide Sequence Database is: Stoesser G., Tuli M.A., Lopez R. and Sterk. P. (1999) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **27**, 18–24.

How to contact the European Bioinformatics Institute

Network	
General enquiries:	datalib@ebi.ac.uk
EBI WWW home page	http://www.ebi.ac.uk
Data submissions (WWW)	http://www.ebi.ac.uk/submission/webin.html
Data submissions (Sequin)	http://www.ebi.ac.uk/submission/sequin.html
Data submissions (Email)	datasubs@ebi.ac.uk
Updates (WWW)	http://www.ebi.ac.uk/ebi_docs/update.html
Updates (Email)	update@ebi.ac.uk
EBI network fileserver	netserv@ebi.ac.uk
FastA sequence search server	fasta@ebi.ac.uk
MPsrch protein sequence search	blitz@ebi.ac.uk
Blast sequence search server	blast@ebi.ac.uk
FTP server (anonymous)	ftp.ebi.ac.uk
Software	ftp.ebi.ac.uk/pub/software

Postal address. EMBL Outstation–The EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Tel: +44 1223 494444; Fax: +44 1223 494468.

REFERENCES

1 Emmert,D.B, Stoehr,P.J. Stoesser,G. and Cameron,G.N. (1994) *Nucleic Acids Res*., **22**, 3445–3449.
2 Bairoch,A. and Apweiler,R. (1998) *Nucleic Acids Res*., **26**, 38–42.
3 Lijnzaad P., Helgesen C. and Rodriguez-Tomé P. (1998) *Nucleic Acids Res*., **26**, 102–105.
4 Périer R.C., Junier T. and Bucher P. (1998) *Nucleic Acids Res*., **26**, 353–357.
5 Heinemeyer,T., Wingender,E., Reuter,I., Hermjakob,H., Kel,A.E., Kel,O.V., Ignatieva,E.V., Ananko,E.A., Podkolodnaya,O.A., Kolpakov,F.A., Podkolodny,N.L. and Kolchanov,N.A. (1998) *Nucleic Acids Res*., **26**, 362–367.
6 Flybase Consortium (1998) *Nucleic Acids Res*., **26**, 85–88.
7 Cohen,B.L., Sheps,J.A. and Wilkinson,M. (1998) *Systematic Biol*., **47**, 479–480.
8 Stoehr,P.J. and Omond,R.A. (1989) *Nucleic Acids Res*., **17**, 6763–6764.
9 Etzold,T., Ulyanov,A. and Argos,P. (1996) *Methods Enzymol*., **266**, 114–128.
10 Lefranc,M.-P., Guidicelli,V., Busin,C., Bodmer,J., Müller,W., Bontrop,R., Lemaitre,M., Malik,A. and Chaume,D. (1998) *Nucleic Acids Res*., **26**, 297–303.
11 Horn,F., Weare,J., Beukers,M.W., Horsch,S., Bairoch,A., Chen,W., Edvardsen,O., Campagne,F. and Vriend,G. (1998) *Nucleic Acids Res*., **26**, 275–279.
12 Pearson,W.R. (1994) *Methods Mol. Biol*., **24**, 307–331.
13 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol*., **215**, 403–410.
14 Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) *Nucleic Acids Res*., **25**, 3389–3402.
15 Smith,R.F. and Waterman,M.S. (1981) *Adv. Applied Math*., **2**, 482–489.
16 Higgins,D., Thompson,J.D. and Gibson,T.J. (1994) *Nucleic Acids Res*., **22**, 4673–4680.
17 Borodovsky,M. (1993) *Comput. Chem*., **17**, 123–133.
18 Jonassen,I., Collins,J.F. and Higgins,D.G. (1995) *Protein Sci*., **4**, 1587–1595.