Nucleic Acids Research, 1997, Vol. 25, No. 1 7–13

The EMBL Nucleotide Sequence Database

Guenter Stoesser*, Peter Sterk, Mary Ann Tuli, Peter J. Stoehr and Graham N. Cameron

EMBL Outstation, the EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Received September 22, 1996; Accepted September 22, 1996

ABSTRACT

The EMBL Nucleotide Sequence Database is a comprehensive database of DNA and RNA sequences directly submitted from researchers and genome sequencing groups and collected from the scientific literature and patent applications. In collaboration with DDBJ and GenBank the database is produced, maintained and distributed at the European Bioinformatics Institute (EBI) and constitutes Europe's primary nucleotide sequence resource. Database releases are produced quarterly and are distributed on CD-ROM. EBI's network services allow access to the most up-to-date data collection via Internet and World Wide Web interface, providing database searching and sequence similarity facilities plus access to a large number of additional databases.

INTRODUCTION

The EMBL Nucleotide Sequence Database is a central activity of the European Bioinformatics Institute (EBI), an EMBL outstation located at the Wellcome Trust Genome Campus at Hinxton, near Cambridge, UK. Database services provided by the EBI (1) are a continuation and extension of the former EMBL Data Library (2), in Heidelberg, Germany. Additional to the production of the nucleotide sequence database, the EBI maintains and distributes the SWISS-PROT protein Sequence Database (3) in collaboration with Amos Bairoch of the University of Geneva, TREMBL (a SWISS-PROT supplement consisting of translations from EMBL database coding sequences), the Radiation Hybrid Database (Rhdb) and many other additional specialist molecular biology databases, many of which are produced in collaboration with the EBI.

The EMBL Nucleotide Sequence Database

The EMBL Data Library was established in 1980 to collect, organize and distribute a database of nucleotide sequence data and related information. Since 1982 this work has been done in collaboration with GenBank (4) (NCBI, Bethesda, USA) and the DNA Database of Japan (Mishima). Each of the three international collaborating databases DDBJ/EMBL/GenBank collects a portion of the total sequence data reported world-wide. Procedures are in place to ensure that all new and updated database entries are exchanged between the collaborating databases on a daily basis (Fig. 1). The explosive growth of the database

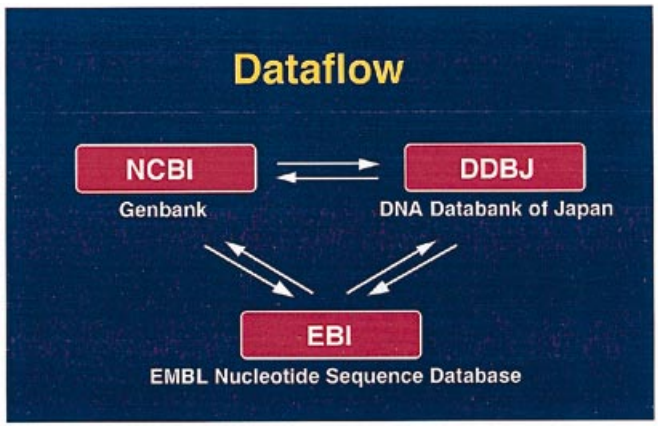

Figure 1. Dataflow between international databases DDBJ/EMBL/GenBank.

continues—sequencing technology has produced means of reading DNA almost like bar-code (Fig. 2). EMBL Release 48 (September 1996) reports 931 582 sequence entries comprising 609 302 252 nucleotides and it is expected that by the end of 1996 the database will include over 1 000 000 entries. The recent increase of sequence data is mainly a consequence of large sequencing projects, like the Human Genome Project. These projects yield an enormous amount of DNA sequence data now available in public databanks.

Top five organisms from EMBL Release 48 calculated on base count:

Homo sapiens	40.17%
Caenorhabditis elegans	7.48%
Mus musculus	6.54%
Saccharomyces cerevisiae	5.29%
Arabidopsis thaliana	2.35%

In particular, EST data (expressed sequence tags) are generated in large amounts and are incorporated into the database. Instead of sequencing a whole genome or a part of it, the idea is to sequence short pieces of DNA which represent genes expressed in particular cells, organs or tissues of different organisms yielding a 'dynamic' picture of gene expression patterns. The total number of EST entries in EMBL Release 48 is 606 286 and due to ongoing efforts this number will be increasing exponentially. The EMBL Nucleotide Sequence Database has ongoing collaborations with an increasing number of genome sequencing groups, who produce large quantities of new sequence data.

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

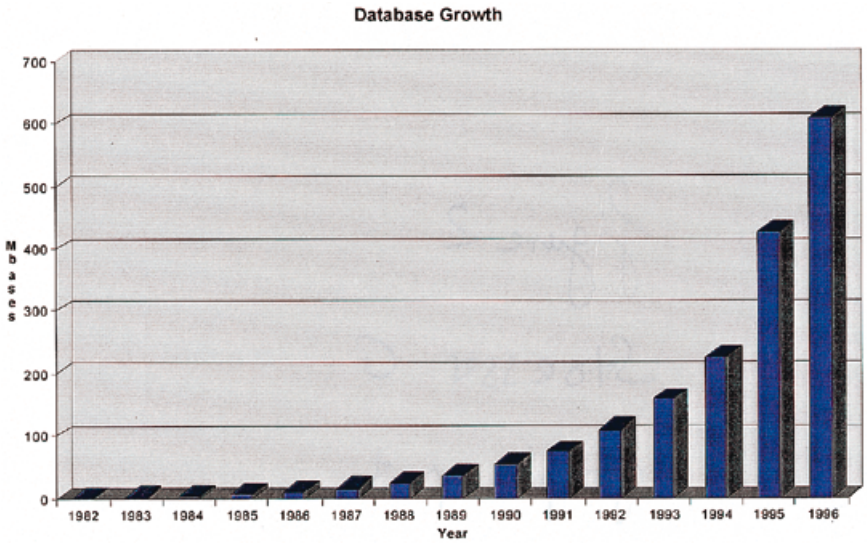

Figure 2. EMBL Nucleotide Sequence Database growth.

Direct data submissions from individual scientists are facilitated by various electronic mechanisms, e.g. the EBI WWW Submission Tool, allowing submission of sequence data and according descriptive biological information to the database in electronic form prior to publication. Most journals now expect that sequence data which appear in journal articles will be submitted to the International Nucleotide Sequence Databases before publication. This mandatory submission policy, regular publishing of database accession numbers in papers, as well as early distribution of 'Table of contents' listings by some of the major journals ensure availability and distribution of new sequence data in a timely fashion.

The EMBL Nucleotide Sequence Database continues to scan all major European molecular biology journals in the context of updating bibliographic references in already existing database entries. The ongoing collaboration with the European Patent Office (EPO) resulted in the capture of more than 25 000 nucleotide and protein sequences, which were published in patent documents between 1960 and 1993 (Patent Backfile) and previously not publicly available in electronic form. A new focus in the context of the collaboration between EPO and EBI now exists for integration of patent data received at the EPO in electronic form since 1994 (Patent Front file).

Some major future developments will include the creation of mechanisms in order to process and organize the growing amount of data originating from genome projects, procedures for handling very long sequences (e.g. complete chromosomes) and also representation of virtual sequences entries.

The complete EMBL Nucleotide Sequence Database is distributed in quarterly releases on a set of compact discs. Software for data query and retrieval is also provided on CD-ROM. The database including daily additions of all new and updated entries is available via the EBI network services (see below) and from nodes of the European Molecular Biology network (EMBnet, see below). EMBL nucleotide sequence database entries are grouped into divisions based on taxonomy. Database entries are distributed in EMBL flat-file format, which is supported by most sequence analysis software packages. Each line of a flat file entry begins with a two-character line code, indicating the type of information contained in the line. A User Manual document is available from the EBI WWW pages providing the complete information on the respective line codes and according descriptions. A typical database entry contains a sequence, a brief description for cataloging purposes, the taxonomic description of the source organism, reference information, and the feature table, containing locations of coding regions and other biologically significant sites. The feature table follows the unified DDBJ/EMBL/GenBank Feature Table Definition (a copy of which can be retrieved from the EBI network server). Where appropriate, entries are cross-referenced to SWISS-PROT, Eukaryotic Promoter database (5), TransFac (6) or Flybase (7). The feature table qualifier '/db_xref' represents cross-references to external databases. For example, a cross-reference from a CDS feature to the database 'FLYBASE' indicates that this feature corresponds to the entity (e.g. gene name) in the FLYBASE database with the given identifier, e.g. /db_xref='FLYBASE:FBgn0012052'.

Data submissions methods

The EBI provides a number of different mechanisms for the direct submission of data (Table 2). Direct submission of sequence data to the nucleotide sequence databases is the primary means of data acquisition, and the most reliable means of ensuring that entries accurately and completely reflect the underlying data. Due to the now standard practice among researchers of submitting their data directly to one of the collaborating databases, there has been an unprecedented reduction in the delay between determination of a sequence and the appearance of that sequence in the database compared with earlier years.

Sequences submitted to the database can be released either immediately after processing or upon publication, depending on the specifications by the submitter. 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.

World wide web

Many individual submissions are now received through EBI's WWW data submission tool (URL: http://www.ebi.ac.uk/subs/emblsubs.html). Using a series of forms submitters are assisted through the submission process. Any world wide web browser supporting forms (Netscape, MacWeb, Lynx, Mosaic) can be used. The web-based submission tool has become the preferred submission medium.

Authorin

Authorin remains a popular submission mechanism, although since the availability of the WWW submission tool, the numbers of Authorin submissions have decreased. Authors prepare their data interactively using MS-DOS or Macintosh computers. One of the main advantages of the Authorin program is that the resulting submission can be automatically processed by the database curation staff. The Authorin program can be obtained electronically from the EBI FTP server (see Appendix 1)

Bulk submissions

This procedure is suitable for those groups submitting a large number of similar sequences only once. Authors planning to submit in this way should contact database staff prior to submission in order to discuss the most appropriate mechanism.

Forms

The Direct Submission Form can also be used for nucleotide sequence submissions. It can be obtained from the EBI network server or by contacting the EBI directly, and a copy is also published periodically in relevant journals.

Submission accounts

For groups producing large volumes of nucleotide sequence data over an extended period, submission accounts can be established with the EBI. A submission protocol is agreed upon and database entries produced at the research site can be deposited and updated directly by the originating group via FTP or electronic mail. A number of new genome projects and research groups have established submission accounts in the past few years, and the procedure has demonstrated itself to be flexible and efficient both for the research groups and for database staff. Each submission account is 'curated' by EBI biologists, who check to ensure that new entries follow database annotation conventions and are consistent with other entries from the same project. The curator also serves as an informed liaison between the sequencing group and the database. A list of groups who already submit data using this method or are expected to begin doing so in the near future is given below.

- EST project, Genexpress, France
- EST project, Genexpress, Germany
- Human Genome Mapping Project, HGMP-RC, UK.
- HIV, Amsterdam, Netherlands
- MHC, Tübingen, Germany
- Human EST, Padova, Italy
- European *Drosophila* Mapping Consortium, Cambridge, UK.
- Fugu GSS, MRC/HGMP-RC, UK.
- Human Genome Mapping Project, Sanger Centre, UK.
- *Caenorhabditis elegans*, Sanger Centre, UK.
- *Mycoplasma capricolum*, NCHGR, USA.
- *Schizosccharomyces pombe*, Sanger Centre, UK.
- *Mycobacterium tuberculosis*, Sanger Centre, UK.
- *Ciona intestinalis*, Sanger Centre, UK.
- *Brugia malayi*, Sanger Centre, UK.

Sequences from patent literature

The capture of data reported in the patent literature since 1960 has continued under contract from the European Patent Office (EPO). The number of entries produced through these activities has turned out to be substantially higher than initially expected, with more than 25 000 protein and nucleotide sequences recovered to date. It should be noted that only a portion of the patent entries are suitable for inclusion in the EMBL Nucleotide Sequence Database; the remaining data are made available in a separate file. After finishing inclusion of 'backfile' EPO sequences, the EBI and EPO have begun collaborating on new means of ensuring that patent sequences appear in the public databases in a timely fashion. Since September 1993, the EPO requires that protein and nucleotide sequences appearing in patent applications be submitted to the EPO in electronic form, which will greatly facilitate the speedy incorporation of these sequences into the database as they become publicly available. New focus now exists in the context of the collaboration between EPO and EBI on integration of patent data received at the EPO in electronic form since 1994 (Patent Front file).

Journal-scanning activities

Mandatory sequence submission requirements on the parts of many journals, the regular practice of publishing database accession numbers in papers, as well as early distribution of 'Table of Contents' listings by some of the most important journals, have greatly enhanced the effectiveness of EBI journal scanning activities over the past years. The EBI continues to scan all major European molecular biology journals, but the activity is directed more toward updating bibliographic references in existing (submitted) entries than toward capturing new sequences. There is still, unfortunately, a small percentage of published sequence data which have not been submitted to any of the three collaborating databases. When these sequences are identified, the authors are contacted and asked to submit their data. The database regularly makes use of entries produced by the NCBI journal scanning operations, both for updating bibliographic references in existing entries, and for including the NCBI entries in the database when no submission exists.

DATA ACCESS

CD-ROM

A set of CD-ROMs is distributed quarterly in the international ISO 9660 standard format. The main contents are the nucleotide and protein sequence databases. Software for data query and retrieval is also provided on the CD-ROM (12). The program EMBL-Search for Macintosh and Windows (13) allows data access by entry name, accession number, keyword, citation, author name, taxonomic classification, database cross-reference, free text and date. EMBL-Search also accesses the Prosite and Enzyme databases, and enables navigation between related entries via the cross-references built into the databases. It uses binary indices whose structure is documented and therefore

Table 1. Databases distributed by the EBI

Database	Description	Ref
3D ali	Structure-based sequence alignments	16
AAtDB	Arabidopsis thaliana genome database	17
Alu	ALU sequences and alignments	18
ANDROGEN	Androgen receptor mutations database	19
Berlin RNA	5S rRNA sequences	20
Bio-Catalog	Directory of molecular biology and genetics software	11
Blocks	Protein Blocks Database	21
CODONUSAGE	Codon usage tables	22
CpGisle	CpG islands database	23
Cutg	Codon usage tabulated from GenBank	24
dbEST	Expressed sequence tags	9
dbSTS	Sequence tagged sites	25
DSSP	Secondary structure assignments of pdb files	26
ECDC	Escherichia coli database collection	27
EMBL	Nucleotide sequence database	1
Enzyme	Database of EC nomenclature	28
EPD	Eukaryotic promoter database	5
FANS_REF	Reference database in the field of functional analysis of nucleotide sequences	29
FlyBase	Drosophila genetic map database	7
FSSP	Families of structurally similar proteins	30
HaemA	Haemophilia A database	31
HaemB	Haemophilia B database	32
HAMSTeR	Haemophilia A mutations database	33
HLA	HLA class I and II sequence database	34
HSSP	Protein structure-sequence alignments	35
IMGT	Immunogenetics database	8
Kabat	Proteins of immunological interest	36
LiMB	List of molecular biology databases	37
Lista	Yeast protein coding sequences	38
Methyl	Site-specific methylation	39
Misfolded	Deliberately misfolded protein models	40
NRL3D	Sequence-structure database	41
NRSub	Non-redundant Bacillus subtilis genome database	42
Nucleosomal DNA	Nucleosomal DNA sequences	43
P53	P53 mutations database	44
P53APC	P53 and APC mutations database	45
PDB	Brookhaven protein structures database	46
PDB Select	Representative list of PDB chain identifiers	47
PIR	Protein sequence database	48
PKCDD	Protein kinase catalytic domain sequence database	49
PLMITRNA	Compilation/classification of higher plant mitochondrial tRNA genes	50.
Primers	PCR primers database	51
Prints	Protein motif fingerprint database	52
Prodom	Protein sequence modules (recurring domains)	53
Prosite	Prosite pattern database	54
PUU	Database of structural domains	55
RDP	Ribosomal database project	56
REBASE	Restriction enzyme database	57
RELibrary	Comprehensive restriction enzyme lists	58
RepBase	Prototypic human repetitive DNA sequences	59
RHdb	Radiation hybrid database	11
RLDB	Reference library database	60
rRNA	Small subunit rRNA sequences	61
SBASE	Protein domain database	62
SeqAnalRef	Sequence analysis bibliography	63
SmallRNA	Compilation of small RNA sequences	64
SRP	Signal recognition particle database	65
SubtiList	Bacillus subtilis database collection	66
SWISS-PROT	Protein sequence database	3
TFD	Transcription factor database	67
TransFac	Eukaryotic cis-acting regulatory DNA elements and trans-acting factors	6
Transterm	Translational termination signal database	68
TREMBL	EMBL coding region translations not integrated in SWISS-PROT	3
tRNA	Database of tRNA sequences	69
UTR-DB	5' and 3' mRNA untranslated regions databases (Metazoa)	70
Yeast	Yeast chromosome database	71

available for other software systems. The sequence databases are also provided on a separate CD-ROM in FastA format for use with software such as FastA on Macintosh and PC systems.

EBI NETWORK SERVICES

The EBI is dedicated to developing network services which take full advantage of the rapid progress in computer network technologies. The EBI databases and software archives are currently accessible via the world wide web, electronic mail fileserver and FTP. New and updated entries from the sequence databases are added daily to the network servers, making it possible to retrieve entries and perform sequence similarity searches on the very latest nucleotide data. A collection of more than 50 additional specialist molecular biology databases is also available (Table 1). These databases are produced by other groups in Europe and world-wide, many in collaboration with the European Bioinformatics Institute. One example is the ImmunoGenetics database (IMGT), a database (8) containing nucleotide sequence information of genes important in the function of the immune system. Complementing these extensive data resources is a collection of freely available molecular biology software for MS-DOS, Macintosh, VMS and UNIX accessible from the EBI WWW pages and network servers (ftp://ftp.ebi.ac.uk/pub/software). Also available is the Bio-Catalog, a list of software of general interest in molecular biology and genetics. First developed at CEPH/Genethon (10), it is now maintained and distributed by the EBI (11). Additionally, documents such as subscription and submission forms, and the DDBJ/EMBL/GenBank Feature Table Definition, can also be retrieved.

Table 2. Summary of submission mechanisms for the EMBL database

Method	Platforms	Notes
Submission Form	Post E-Mail	Printed copies from: 1. the first issue of *Nucl. Acid. Res.* each year 2. from the EBI by request Electronic copies: 1. from the EBI file servers • *http://www.ebi.ac.uk/ebi_docs/embl_db/ebi/dataform.txt* • *ftp://ftp.ebi.ac.uk/pub/databases/embl/release/doc/datasub.txt* 2. with each EMBL release 3. from the EBI on a Macintosh or PC formatted disk by request
Authorin	Macintosh (1) PC(2)	Ftp: 1. *ftp://ftp.ebi.ac.uk/pub/software/mac/authorin.hqx* 2. *ftp://ftp.ebi.ac.uk/pub/software/dos/authorin.exe* E-mail: Send an email to *netserv@ebi.ac.uk* with a single line containing one of the following: 1. *GET Mac_software:authorin.hqx* 2. *GET Dos_software:authorin.uaa*
World Wide Web	Most common platforms	Any WWW browser that supports forms (eg. Netscape, MacWeb, lynx, Mosaic) • *http://www.ebi.ac.uk/subs/emblsubs.html*

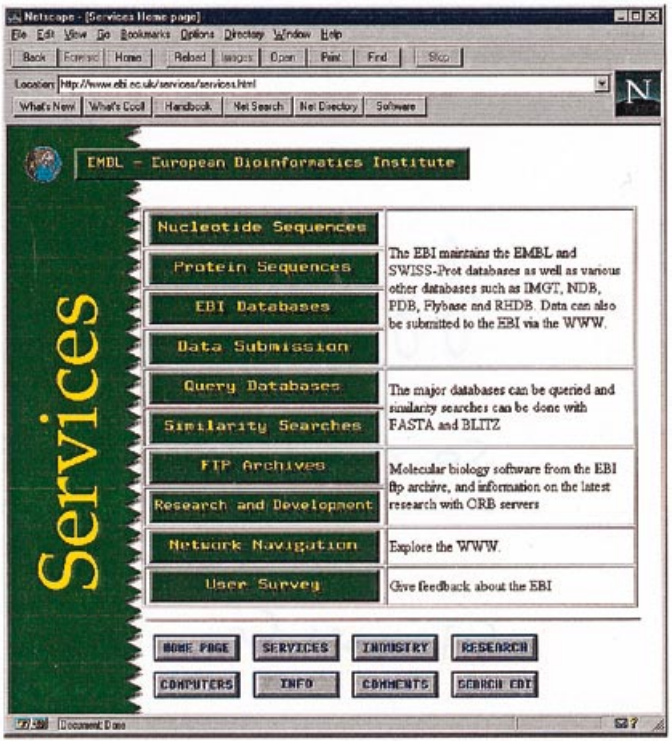

Figure 3. European Bioinformatics Institute (EBI) Services WWW home page.

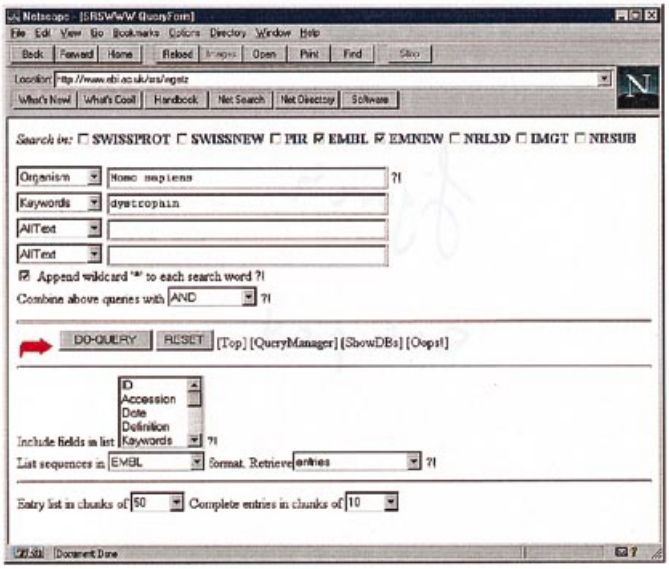

Figure 4. EBIs sequence retrieval system (SRS).

World Wide Web

The EBI WWW server provides the most advanced network access to a broad range of molecular biology information resources (Fig. 3). In addition to the EBI molecular biology archives, FastA and BLITZ sequence similarity search, database query/retrieval and protein structure prediction services are offered. Connect to the EBI WWW server using the URL: http://www.ebi.ac.uk .

Network fileserver

The EBI Network fileserver enables access via electronic mail (e-mail) to the full collection of databases, public domain software and documentation maintained by EBI. Items are retrieved from the server by sending a command in an e-mail message to the fileserver address. Detailed instructions on using the fileserver, and a current list of contents, can be obtained by sending a message to the Internet address netserv@ebi.ac.uk with the word HELP in the body of the message. A full set of instructions will be returned automatically.

FTP server

The EBI anonymous file transfer protocol (FTP) server enables navigation through the directories of the EBI molecular biology

Table 3. Sites maintaining daily updated copies of EMBL Nucleotide Sequence Database

National nodes	Contact Addresses
Austria e-mail contact: grabner@cc.univie.ac.at	Bio Computing Center, University of Vienna, Dr. Bohr-Gasse 9, A-1030 Vienna Austria
Belgium e-mail contact: rherzog@ulb.ac.be	Belgian EMBnet Node, Université Libre de Bruxelles, C.P.300, Paardenstraat 67,B-1640 Sint Genesius Rode, Belgium
Denmark e-mail contact: hum@biobase..dk	BioBase, Ole Worms alle, Building 170, Aarhus Universitet, DK-8000 Aarhus C, Denmark
France e-mail contact: dessen@infobiogen.fr	INFOBIOGEN, 7 rue Guy Môquet, BP 8, F-94801 Villejuif Cedex, France
Finland e-mail contact: heikki.lehvaslaiho@csc.fi	Centre for Scientific Computing, PO Box 405, SF- 02101 Espoo, Finland
Germany e-mail contact: w.chen@genius.embnet.dkfz-heidelberg.de	DKFZ, Im Neuenheimer Feld 280, D-69120 Heidelberg, Germany
Greece e-mail contact: savakis@myia.imbb.forth.gr	IMBB, PO Box 1527, Heraklion GR-71110, Crete, Greece
Israel e-mail contact: lsestern@weizmann.ac.il	Biological Computing Division, Weizmann Institute of Science, Rehovot 76100, Israel
Italy email contact: attimonelli@area.ba.cnr.it - e-mail contact: pongor@icgeb.trieste.it	CNR Area di Ricerca di Bari, Via Amendola 166/5, I-70125, Bari, Italy ICGEB, Area Research Park, Padriciano 99, I-34012 Trieste, Italy
Netherlands e-mail contact: noordik@caos.kun.nl -	CAOS/CAMM Center, University of Nijmegen, Toernooiveld, 6525 ED Nijmegen, The Netherlands European Patent Office, P.B. 5818, Patentlaan 2, 2280 HV Rijswijk (ZH), The Netherlands
Norway e-mail contact: rodrigol@biotek.uio.no	Norwegian EMBnet Node, Biotechnology Center of Oslo, Gaustadaleen 21, N-0371 Oslo, Norway
Spain e-mail contact: carazo@samba.cnb.uam.es	Centro nacional de biotecnologia, CSIC, Universidad Autonoma de Madrid, 28049 Madrid, Spain
Sweden e-mail contact: gad@bmc.uu.se	Computing Department, Biomedical Centre, Box 570, S 751 23, Uppsala, Sweden
Switzerland e-mail contact: embnet@ch.embnet.org - e-mail contact: daniel.doran@roche.com	Biocomputing, Biozentrum der Universitaet Basel, Klingelbergstrasse 70, CH 4056 Basel, Switzerland Hoffman-La Roche Ltd., Pharma Preclinical Res., CH 4002 Basel, Switzerland
United Kingdom e-mail contact: bleasby@dl.ac.uk	SEQNET, SERC Daresbury Lab., Keckwick Lane, Warrington WA4 4AD, UK

archives and retrieval of files. Most directories have 'README' files to help with orientation. Users should connect to the anonymous FTP server at the address ftp.ebi.ac.uk using the username 'anonymous,' and their e-mail address as the password.

Sequence search facilities

The EBI provides a number of services that allow external users to compare their own sequences against the most currently available data in the EMBL Nucleotide Sequence Database and SWISS-PROT. BLITZ is based on the MPsrch program of Collins and Sturrock (Edinburgh University) which uses the well-known Smith and Waterman (14) algorithm for sensitive searches of the protein sequence databases. It is implemented on a MasPar massively-parallel computer at the EBI. Detailed instructions can be obtained by sending an e-mail message to the address blitz@ebi.ac.uk with the word HELP in the body of the message. Mail-FastA is based on Pearson's FastA program (15). It performs sensitive comparisons of nucleotide or amino acid sequences against the database. Further information can be obtained by sending an email to the address fasta@ebi.ac.uk , with the word HELP in the body of the message.

Database query/retrieval

The EBI provides a query/retrieval system using SRS (65), the Sequence Retrieval System (Fig. 4). Specific query forms are accessable at the URL: http://www.ebi.ac.uk/srs/srsc

EMBnet

The European Molecular Biology Network 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 3 gives a full listing of sites maintaining daily updated copies of the EMBL Nucleotide Sequence Database.

The preferred form for citation of the EMBL Nucleotide Sequence Database is:

The EMBL Nucleotide Sequence Database,
Stoesser,G., Sterk,P., Tuli,M.A., Stoehr,P.J. and Cameron,G.N.
Nucleic Acids Res. 1997, Vol. **25**, No. 1, pages 7–13.

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 (E-mail)	datasubs@ebi.ac.uk
Data submissions (WWW)	http://www.ebi.ac.uk /subs
Corrections to nucleotide entries (E-mail)	update@ebi.ac.uk
Corrections to nucleotide entries (WWW)	http://www.ebi.ac.uk/ebi_docs/update.html
EBI network fileserver	netserv@ebi.ac.uk
FastA sequence search server	fasta@ebi.ac.uk
MPsrch protein sequence search server	blitz@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.
Telephone:	+44 (1223) 494444
Telefax:	+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 Rice,C.M., Fuchs,R., Higgins,D.G., Stoehr,P.J. and Cameron,G.N. (1993) *Nucleic Acids Res*., **21**, 2967–2971.
3 Bairoch,A. and Apweiler,R. (1996) *Nucleic Acids Res*., **24**, 21–25.
4 Benson,D., Lipman,D.J. and Ostell,J. (1994) *Nucleic Acids Res*., **22**, 3441–3444.
5 Bucher,P. and Trifonov,E.N. (1986) *Nucleic Acids Res*., **14**, 10009–10026.
6 Wingender,E. (1988) *Nucleic Acids Res*., **16**, 1879–1902.
7 The FlyBase Consortium (1994) *Nucleic Acids Res*., **22**, 3456–3458.
8 Lefranc,M.-P., Giudicelli,V., Busin,C., Malik,A., Mougenot,I., Delhais,P. and Chaume,D. (1995) *Ann. New York Acad. Sci.*, **764**, 47–50.
9 Boguski,M.S., Lowe,T.M.J. and Tolstoshev,C.M. (1993) *Nature Genet.*, **4**, 332–333.
10 Rodriguez-Tomé,P. and Caterina,D. (1993) CEPH/Généthon.
11 Rodriguez-Tomé,P. (1995) EMBL-EBI.
12 Fuchs,R. and Stoehr,P.J. (1993) *CABIOS* **9**, 71–77.
13 EMBL-EBI (1995).
14 Smith,T.F. and Waterman,M.S. (1981) *J. Mol. Biol*., **147**, 195–197.
15 Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448
16 Pascarella,S. and Argos,P. (1992) *Protein Engng*, **5**, 121–137.
17 Arabidopsis thaliana Database Project (1995) Massachusetts General Hospital, Harvard Medical School, USA
18 Jurka,J. and Smith,T. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 4775–4778.
19 Patterson,M.N., Hughes,I.A., Gottlieb,B. and PinskyL. (1994) *Nucleic Acids Res*. **22**, 3560–3562.
20 Specht,T., Wolters,J. and Erdmann,V.A. (1991) *Nucleic Acids Res*., **19**, 2189–2191.
21 Wallace,J.C. and Henikoff,S. (1992) *CABIOS*, **8**, 249–254.
22 CherryM. (1992) Massachusetts General Hospital, USA
23 Larsen,F., Gundersen,G., Lopez,R. and Prydz,H. (1992) *Genomics*, **13**, 1095–1107.
24 Wada,K., Wada,Y., Ishibashi,F., Gojobori,T. and Ikemura,T. (1992) *Nucleic Acids Res*., **20**, 2111–2118.
25 Olson,M., Hood,L., Cantor,C. and Botstein,D. (1989) *Science*, **254**, 1434–1435.
26 Sander,C. (1993) EMBL, Heidelberg.
27 Wahl,R., Rice,P., RiceC.M. and Kröger,M. (1994) *Nucleic Acids Res*., **22**, 3450–3455.
28 Bairoch,A. (1994) *Nucleic Acids Res*., **22**, 3626–3627.
29 GelfandM.S. (1991) USSR Academy of Sciences, Puschino, USSR
30 Holm,L., Ouzounis,C., Sander,C., Tuparev,G. and Vriend,G. (1992) *Protein Sci*., **1**, 1691–1698.
31 Tuddenham,E.G., Schwaab,T., Seehafer,J., Millar,D.S., Gitschier,F., Higuchi,M., Bidichandani,S., Connor,J.M., Hoyer,L.W. and Yoshioka,A. (1994) *Nucleic Acids Res*., **22**, 4851–4868.
32 Giannelli,F., Green,P.M., Sommer,S.S., Lillicrap,D.P., Ludwig,M., Schwaab,R., Reitsma,P.H., Goossens,M., Yoshioka,A. and Brownlee,G.G. (1994) *Nucleic Acids Res*., **22**, 3534–3546.
33 Haemostasis Research Group (1996) Hammersmith Hospital London, UK.
34 Bodmer,J.G., Marsh,S.G., Albert,E.D., Bodmer,W.F., Dupont,B., Erlich,H.A., Mach,B., Mayr,W.R., Parham,P. and Sasazuki,T. (1994) *Tissue Antigens*, **44**, 1–18.
35 Sander,C. and Schneider,R. (1994) *Nucleic Acids Res*., **22**, 3597–3599.
36 Kabat,E.A., Wu,T.T., Perry,H.M., Gottesman,K.S. and Foeller,C. (1992)
37 Keen,G., Redgrave,G., Lawton,J., Cinkosky,M., Mishra,S., Fickett,J. and Burks,C. (1992) *Mathl. Comp. Modelling*, **16**, 93–101.
38 Dölz,R., Mossé,M.-O., Bairoch,A., Slonimski,P.P. and Linder,P. (1994) *Nucleic Acids Res*., **24**, 66–91.
39 McClelland, M, Nelson,M. and Raschke,E. (1994) *Nucleic Acids Res.*, **22**, 3640–3659.
40 Holm,L. and Sander,C. (1992) *J. Mol. Biol*., **225**, 93–105.
41 Pattabiraman,N., Namboodiri,K., Lowrey,A. and Gaber,B.P. (1990) *Protein Seq. Data Anal*., **3**, 387–405.
42 Perriere,G., Gouy,M. and Gojobori,T. (1994) *Nucleic Acids Res*., **22**, 5525–5529.
43 Isohikhes,I. and Trifonov,E.N. (1993) *Nucleic Acids Res*., **21**, 4857–4859.
44 Hollstein,M., Rice,K., Greenblatt,M.S., Soussi,T., Fuchs,R., Sorlie,T., Hovig,E., Smith-Sorenson,B., Montesano,R. and Harris,C.C. (1994) *Nucleic Acids Res*., **22**, 3551–3555.
45 Sommer,S.S. and LiaoD. (1995) Mayo Clinic/Foundation, Rochester, USA.
46 Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) *J. Mol. Biol*., **112**, 535–542.
47 Hobohm,U., Scharf,M., Schneider,R. and Sander,C. (1992) *Protein Sci.*, **1**, 409–417.
48 George,D.G., Barker,W.C., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1994) *Nucleic Acids Res*., **22**, 3569–3573.
49 Hanks,S.K. and Quinn,A.M. (1991) *Methods Enzymol*., **200**, 38–62.
50 Ceci,L.R. (1996) C.N.R., Bari, Italy.
51 Shomer,B. (1996) EMBL-EBI.
52 Attwood,T.K., Beck,M.E., Bleasby,A.J. and Parry-Smith,D.J. (1994) *Nucleic Acids Res*., **22**, 3590–3596.
53 Sonnhammer,E.L.L. and Kahn,D. (1994) *Protein Sci*., **3**, 482–492.
54 Bairoch,A.and Bucher,P. (1994) *Nucleic Acids Res*., **22**, 3583–3589.
55 Holm,L. and Sander,C. (1994) *Proteins*, **19**, 256–268.
56 Maidak,B.L., Larsen,N., McCaughey,M.J., Overbeek,R., Olsen,G.J., Fogel,K., Blandy,J. and Woese,C.R. (1994) *Nucleic Acids Res*., **22**, 3485–3487.
57 Roberts,R.J. and Macelis,D. (1994) *Nucleic Acids Res*., **22**, 3628–3639.
58 Raschke,E. (1993) *Genet. Anal. Tech. Applic.*, **10**, 49–60.
59 Jurka,J., Walichiewicz,J. and Milosavljevic,A. (1992) *J. Mol. Evol*., **35**, 286–291.
60 Lehrach,H. (1990) *Genome Anal*., **1**, 39–81.
61 Neefs,J.M., Van de Peer,Y., De Rijk,P., Chapelle,S. and De Wachter,R. (1993) *Nucleic Acids Res*., **21**, 3025–3049.
62 Pongor,S., Hátsági,Z., Degtyarenko,K., Fábián,P., Skerl,V., Hegyo,H., Myrvai,J. and Bevilacqua,V. (1994) *Nucleic Acids Res*., **22**, 3610–3615.
63 Bairoch,A. (1994) University of Geneva.
64 Gu,J. and Reddy,R. (1994) *Nucleic Acids Res*., **22**, 3481–3482.
65 Larsen,N. and ZwiebC. (1993) *Nucleic Acids Res*., **21**, 3019–3020.
66 Moszer,I., Glaser,P. and Danchin,A. (1995) *Microbiology* **141**, 261–268.
67 Ghosh,D. (1992) *Nucleic Acids Res*., **20**, 2091–2093.
68 Brown,C.M., Stockwell,P.A., Dalphin,M.E. and Tate,W.P. (1994) *Nucleic Acids Res*., **22**, 3620–3624.
69 SteinbergS., Misch,A. and SprinzlM. (1993) *Nucleic Acids Res*., **21**, 3011–3015.
70 Pesole,G., Grillo,G. and Liuni,S. (1996) *Comp. Chem.* **20**, 141–144.
71 Liebl,S. and Sonnhammer,E. (1994) MIPS, Germany and Sanger Centre, UK.