6–12 Nucleic Acids Research, 1996, Vol. 24, No. 1

The European Bioinformatics Institute (EBI) databases

Patricia Rodriguez-Tomé*, Peter J. Stoehr, Graham N. Cameron and Tomas P. Flores

EMBL Oustation, the EBI, Hinxton Hall, Hinxton, Cambridge CB10 1RQ, UK

Received October 12, 1995; Accepted October 23, 1995

ABSTRACT

The European Bioinformatics Institute (EBI) maintains and distributes the EMBL Nucleotide Sequence database, Europe's primary nucleotide sequence data resource. The EBI also maintains and distributes the SWISS-PROT Protein Sequence database, in collaboration with Amos Bairoch of the University of Geneva. Over fifty additional specialist molecular biology databases, as well as software and documentation of interest to molecular biologists are available. The EBI network services include database searching and sequence similarity searching facilities.

INTRODUCTION

The European Bioinformatics Institute (EBI) is an EMBL Outstation, located at Hinxton Hall, near Cambridge, UK. Since September 1994, all activities previously based at the EMBL Data Library (1) in Heidelberg, Germany are located at the EBI. The database services of the EBI (2) are the continuation and extension of the EMBL Data Library. A central activity of the European Bioinformatics Institute (EBI) is the development and distribution of the EMBL Nucleotide Sequence database. The EBI also maintains and distributes the SWISS-PROT Protein Sequence database (3) in collaboration with Amos Bairoch of the University of Geneva. Over fifty additional specialist molecular biology databases, some produced in collaboration with the EBI, are also distributed through EBI releases and network services.

DATABASES

The EMBL Nucleotide Sequence database

The main activity of the group is the development, maintenance and distribution of a comprehensive database of nucleotide sequences. The EMBL nucleotide sequence database, produced in collaboration with GenBank (4) (NCBI, Bethesda, USA) and the DNA database of Japan (Mishima), is Europe's primary nucleotide sequence data resource. Each of these three groups collect a portion of the total sequence data reported world-wide. All new and updated database entries are exchanged between the groups on a daily basis. The rate of growth of the database continues to accelerate. As an example, release 44 (September 1995), with more than 360 million bases from 506 192 entries, represents an annual increase of ~2.5 times the number of entries and 1.7 times the number of bases. Important sources of data have been secured through collaborations with genomic sequencing projects and other groups, such as phylogenetic research groups, who produce large quantities of new nucleotide sequence data. The ongoing collaboration with the European Patent Office has resulted in the capture of nucleotide and protein sequences which were published in patent documents between 1960 and 1993 and previously not publicly available in electronic form. The complete database is distributed in quarterly releases on compact disc (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).

The nucleotide sequence database entries are distributed in the EMBL flat-file format, which is supported by most sequence analysis software packages. A typical entry contains a sequence, a brief description for cataloging purposes, the taxonomic description of the source organism, bibliographic 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 may also be cross-referenced to SWISS-PROT, Eukaryotic Promoter database (5), TransFac (6) or FlyBase (7). The consistency of database entries reflects the diversity of sequence determination methods. For instance, expressed sequence tag (EST) entries, 'single pass' sequences derived from random clones, often have very little biological 'annotation', compared to the typical entry reported by a researcher who has carefully investigated a single gene. An entry produced by a genomic sequencing group, on the other hand, may be extensively annotated, but the features of the sequence may have been determined by similarity and thus be more or less putative. The EBI devotes considerable resources to ensuring that the biological information attached to nucleotide sequences is as complete as possible. Every effort is made to maintain consistency while preserving the (varied) richness of these data from their various sources.

The SWISS-PROT Protein Sequence database

The SWISS-PROT protein sequence database is maintained collaboratively by the EMBL Data Library and Amos Bairoch of the University of Geneva. It is distributed in the same file format

* To whom correspondence should be addressed

as the Nucleotide Sequence database, with which it is fully cross-referenced. SWISS-PROT entries are derived from various sources including translations of DNA sequences in the EMBL database, adapted from the Protein Identification Resource collection (8) (PIR, Washington, DC), extracted from the literature, and directly submitted by researchers. Its strengths are the quality and consistency of its annotation, non-redundancy, and the cross-references to other databases, especially to the EMBL nucleotide sequence database, PROSITE (9) and PDB (10). SWISS-PROT is distributed on CD-ROM every 3 months, and new entries can be retrieved between releases via the EBI network servers (see below).

The Radiation Hybrid mapping database

The Radiation Hybrid database (Rhdb) is a new development at the EBI (11). This database is an archive of raw data (i.e. PCR results on radiation hybrid panels) with links to other related databases. All cross-references known to the authors or the databases maintainers are included. The user is also able to directly query the relational database (on the World Wide Web) either by using a set of pre-compiled queries or by writing his own ad-hoc queries. The database is distributed in a similar file format as the EMBL database with which it is fully cross-referenced. It is distributed on CD-ROM twice a year and can also be retrieved between CD-ROM releases via the EBI network servers (see below).

The ImMunoGeneTics database

The ImMunoGeneTics database (IMGT) is a database (12) containing nucleotide sequence information of genes important in the function of the immune system. It collects and annotates sequences belonging to the immunoglobin superfamily which are involved in immune recognition. IMGT works in close collaboration with the EMBL database and with three other laboratories in Europe [LIGM (FR), ICRF (UK), Univ. Of Koln (DE)]. It is distributed on CD-ROM twice a year and can also be retrieved between CD-ROM releases via the EBI network servers (see below).

The Bio-Catalog

The Bio-Catalog is a list of software of general interest in molecular biology and genetics. First developed at CEPH/Généthon (13) it is now maintained and distributed by the EBI (14). In addition to this database the EBI maintains a repository of biology related software on its network servers. This software is also distributed once a year on CD-ROM.

Other databases

The EBI is a major distributor of molecular biological databases produced by other groups in Europe and world-wide. More than 50 databases are available via the EBI network servers (see Fig. 1 for the World Wide Web access to the EBI databases) and 30 of them are included on CD-ROM (see Table 1). The EBI also mirrors dbEST, a database of Expressed Sequences Tags developed at the NCBI (15), offering query and retrieval access through the World Wide Web.

DATA ACQUISITION

Today, approximately 95% of all nucleotide sequence data are directly submitted to one of the collaborating databases (EMBL, GenBank and DDBJ). This has reduced the delay between determination of a sequence and appearance of that sequence in the database compared to earlier years. The entries created by each group are exchanged on a daily basis. The remaining 5% are still extracted from the literature (especially patent documents), which is a time-consuming and error-prone task

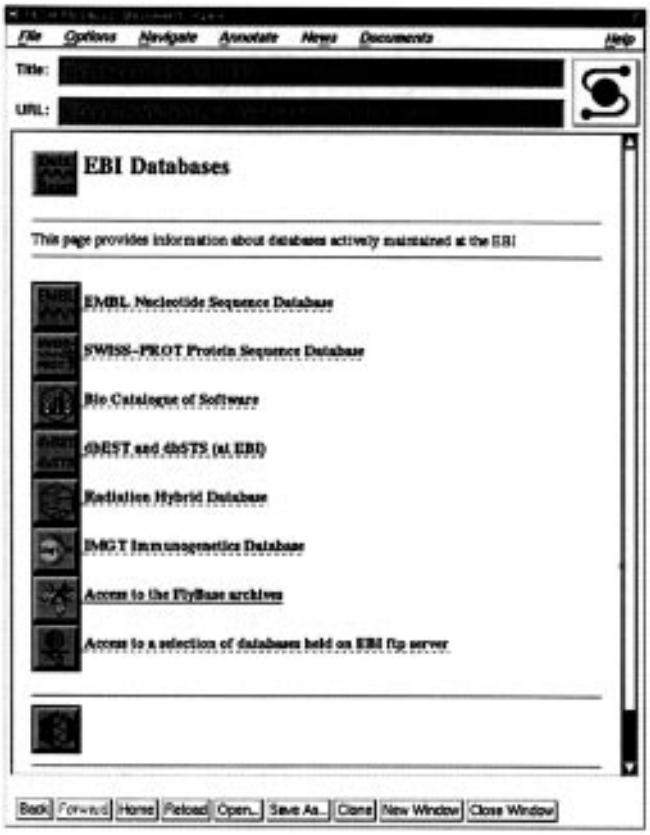

Figure 1. The EBI databases WWW page.

Direct submissions

The EBI provides a number of different mechanisms for the direct submission of data (see 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. Sequences submitted can be released either immediately after processing or upon publication, depending on the wishes of the submittor. In general, unless otherwise directed by the author, submitted sequences are available to the research community before the sequence appears in a journal. One of the direct submission mechanisms is via the Authorin program, which allows authors to 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 really automatically processed by the database annotation staff. The Authorin program can be obtained on diskettes from NCBI (GenBank/NCBI, NIH, Bldg 38A, Bethesda, MD 20894 USA; email: authorin@ncbi.nlm.nih.gov) or electronically from the EBI network server. 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

Table 1. Databases distributed by EBI and the mechanism of distribution in each case

Database	Description	Ref	CD-ROM	Server
3D ali	Structure-based sequence alignments	16		*
Alu	ALU sequences and alignments	17		*
Berlin RNA	5S rRNA sequences	18	*	*
Bio-Catalog	Directory of molecular biology and genetics software	14	*	*
Blocks	Protein Blocks Database	19	*	*
CpGisle	CpG islands database	20	*	*
Cutg	Codon usage tabulated from GenBank	21		*
dbEST	Expressed sequence tags	15		*
dbSTS	Sequence tagged sites	22		*
DSSP	Secondary structure assignments of pdb files	23		*
ECDC	Escherichia coli database collection	24	*	*
EMBL	Nucleotide sequence database	2	*	*
Enzyme	Database of EC nomenclature	25	*	*
EPD	Eukaryotic promoter database	5	*	*
FlyBase	Drosophila genetic map database	7		*
FSSP	Families of structurally similar proteins	26		*
HaemA	Haemophilia A database	27		*
HaemB	Haemophilia B database	28	*	*
HLA	HLA class I and II sequence database	29	*	*
HSSP	Protein structure-sequence alignments	30	*	*
Kabat	Proteins of immunological interest	31	*	*
LiMB	List of molecular biology databases	32	*	*
Lista	Yeast protein coding sequences	33	*	*
Methyl	Site-specific methylation	34		*
Misfolded	Deliberately misfolded protein models	35		*
NRL3D	Sequence-structure database	36		*
NRSub	Non-redundant Bacillus subtilis genome database	37		*
Nucleosomal DNA	Nucleosomal DNA sequences	38	*	*
P53	P53 mutations	39	*	*
PDB	Brookhaven protein structures database	10		*
PDB Select	Representative list of PDB chain identifiers	40		*
PIR	Protein sequence database	8		*
PKCDD	Protein kinase catalytic domain sequence database	41	*	*
Prints	Protein motif fingerprint database	42	*	*
Prodom	Protein sequence modules (recurring domains)	43		*
Prosite	Prosite pattern database	9	*	*
PUU	Database of structural domains	44		*
RDP	Ribosomal database project	45		*
REBASE	Restriction enzyme database	46	*	*
RELibrary	Comprehensive restriction enzyme lists	47	*	*
RepBase	Prototypic human repetitive DNA sequences	48	*	*
RHdb	Radiation hybrid database	11	*	*
RLDB	Reference library database	49	*	*
rRNA	Small subunit rRNA sequences	50	*	*
SBASE	Protein domain database	51	*	*
SeqAnalRef	Sequence analysis bibliography	52	*	*
SmallRNA	Compilation of small RNA sequences	53		*
SRP	Signal recognition particle database	54	*	*
SWISS-PROT	Protein sequence database	3	*	*
TFD	Transcription factor database	55		*
TransFac	Eukaryotic cis-acting regulatory DNA elements and trans-acting factors	6	*	*
Transterm	Translational termination signal database	56	*	*
tRNA	Database of tRNA sequences	57	*	*
Yeast	Yeast chromosome database	58		*

a copy is also published periodically in relevant journals (59,60). This submission form can either be sent to the EBI by post or by electronic mail. A new submission system has been developed at the EBI using the World Wide Web (WWW). There are many benefits to submitting sequences in this way. In particular, the EBI continually maintains and updates this system, ensuring the requested information is up to date.

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. 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

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 Data Library by request Electronic copies: 1. from the Data Library's 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* • *gopher://gopher.ebi.ac.uk/11/EMBL/releaseinfo/submission.form* 2. with each EMBL release 3. from the Data Library 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* Gopher: 1. *gopher://gopher.ebi.ac.uk/11/software/mac/authorin.hqx* 2. *gopher://gohper.ebi.ac.uk/11/software/dos/authorin.exe* Email, 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*

method or are expected to begin doing so in the near future is given below.

- European Drosophila Mapping Consortium
- French Arabidopsis cDNA project GDR
- Genexpress Généthon (FR)
- Généthon (FR)
- Genexpress Munich (DE)
- HIV project Amsterdam (NL)
- MHC project Tuebingen
- Mycoplasma capricolum NCHGR
- Sanger Centre (UK), *C.elegans* nematode project
- Sanger Centre (UK) Human genome project
- Sanger Centre (UK) *S.pombe*
- Sanger Centre (UK) Yeast Chromosome IV
- Sanger Centre (UK) Yeast Chromosome IX
- Sanger Centre (UK) Yeast Chromosome XIII
- Sanger Centre (UK) Yeast Chromosome XVI
- UK Human Genome Mapping Project
- Radiation Hybrid Mapping Consortium

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). All the 'backfile' documents have now been processed, with >25 000 protein and nucleotide sequences captured (with first priority outside the USA and Japan). It should be noted that only a portion of the patent entries are suitable for inclusion in the EMBL nucleotide sequence database; the others are made available in a separate file. The EBI and EPO are collaborating on new means of ensuring that patent sequences appear in the public databases with less delay in the future. Since September 1993, the EPO requires that protein and nucleotide sequences appearing in patent applications be submitted in an electronic form, which greatly facilitate the speedy incorporation of these sequences into the database as they become publicly available.

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 the 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 certain small percentage of published sequence data which has 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 DISTRIBUTION

CD-ROM

The data library no longer produces magnetic tape distributions replacing this operation by CD-ROM only, since it is inexpensive and can be used with a wide range of computer systems. It is distributed quarterly as a set of compact discs written in the international ISO 9660 standard format. Since release 44, there is a separate CD-ROM distribution for EMBL and SwissProt databases. The collaborative databases are distributed on a separate CD-ROM twice a year (see Table 1 for the list of databases included).

Software for data query and retrieval is also provided on the CD-ROM (61). The programs EMBL-Search for Macintosh and SRS for DOS (62) allow data access by entry name, accession number, keyword, citation, author name, taxonomic classifica-

tion, database cross-reference, free text, and date. EMBL-Search also provides access to the Prosite and Enzyme databases, and enables navigation between related entries via the cross-references built into these databases. It uses binary indices whose structure is documented and therefore available for other software systems. The SRS software is a DOS version (this is a port done by the EBI) of the sequence retrieval software used on the EBI network services. The sequence databases are also provided in NBRF format for use with software such as FASTA on Macintosh or MS-DOS systems.

EBI NETWORK SERVICES

In addition to archiving sequence and genome data, the EBI provides an ever-expanding number of free network services to external users. The EBI databases and software archives are currently accessible via electronic mail fileserver, FTP, gopher and World Wide Web (WWW). New and updated entries from all three collaborating nucleotide 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. The complete collection of additional specialist molecular biology databases is also available. Complementing these extensive data resources is a collection of molecular biology software for MS-DOS, Macintosh, VMS and UNIX. Documents such as subscription and submission forms, and the DDBJ/EMBL/GenBank Features Table Definition, can also be retrieved.

EBI network fileserver

The EBI network fileserver (63) 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.

EBI FTP server

This is the main route for retrieving databases or software from the EBI's archive. The EBI anonymous file transfer protocol (FTP) server enables navigation through the directory hierarchy for the anonymous user. 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 giving their e-mail address as the password.

EBI Gopher server

The EBI Gopher server simplifies the use of network services by hiding complexity behind a simple graphical user interface. The files are arranged in a hierarchy of directories like in the FTP server, but have more detailed titles. In addition to accessing the EBI molecular biology archives, links are provided to other information resources in Europe and world-wide. Gopher clients can access the server at gopher.ebi.ac.uk.

EBI World Wide Web server

The EBI WWW server provides the most advanced network access to a broad range of molecular biology information resources. In addition to the EBI molecular biology archives, sequence similarity search and database query/retrieval services are offered. Users can also directly submit their data using the direct submission entry page. Connect to the EBI WWW server using the URL: http://www.ebi.ac.uk which give access to the EBI home page and links to all EBI services.

Database query/retrieval

The EBI provides a query/retrieval system using SRS, the Sequence Retrieval System (64). This system allows entries to be retrieved based on a number of keywords. Specific query forms are accessible at the URL: http://www.ebi.ac.uk/srs/srsc

Sequence search facilities

The EBI provides a number of services that allow 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 Smith and Waterman (65) algorithm for sensitive searches of the protein and nucleotide sequence databases. It is implemented on a MasPar, a massively-parallel computer. 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 (66). 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. Both services can also be accessed interactively using the EBI World Wide Web server.

EMBnet

The European Molecular Biology network (EMBnet) 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, the EMBnet plays an increasingly 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.

HOW TO CONTACT THE EUROPEAN BIOINFORMATICS INSTITUTE

Network: Datalib@ebi.ac.uk (for general enquiries)
Datasubs@ebi.ac.uk (for data submissions to the EMBL and SwissProt databases)
Update@ebi.ac.uk (for corrections to nucleotide entries)
RHdb@ebi.ac.uk (for data submission to Rhdb)
Netserv@ebi.ac.uk (e-mail file server)

Table 3. Sites maintaining daily updated copies of EMBL Nucleotide Sequence Database (Sept 1995)

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

NetHelp@ebi.ac.uk (for network server enquiries)
ftp.ebi.ac.uk (anonymous FTP server)
gopher.ebi.ac.uk (Gopher server)
http://www.ebi.ac.uk (World Wide Web)
blitz@ebi.ac.uk (MPsrch protein sequence search server)
fasta@ebi.ac.uk (FastA sequence search server)

Postal address: EMBL Outstation-the EBI,
Hinxton Hall,
Hinxton,
Cambridge CB10 1RQ, UK.
Telephone: +44 (1223) 494400
Telefax: +44 (1223) 494468

REFERENCES

1 Rice, C.M., Fuchs, R., Higgins, D.G., Stoehr, P.J. and Cameron, G.N. (1993) *Nucleic Acids Res.*, **21**, 2967–2971.
2 Emmert, D.B, Stoehr, P.J. Stoesser, G. and Cameron, G.N. (1994) *Nucleic Acids Res.*, **22**, 3445–3449.
3 Bairoch, A. and Boeckmann, B. (1994) *Nucleic Acids Res.*, **22**, 3578–3580.
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 George, D.G., Barker, W.C., Mewes, H.-W., Pfeiffer, F. and Tsugita, A. (1994) *Nucleic Acids Res.*, **22**, 3569–3573.

9 Bairoch, A.and Bucher, P. (1994) *Nucleic Acids Res.*, **22**, 3583–3589.
10 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.
11 Rodriguez-Tomé, P. (1995) EMBL, Hinxton-EBI.
12 Lefranc, M.-P., Giudicelli, V., Busin, C., Malik, A., Mougenot, I., Delhais, P. and Chaume, D. (1995) *Ann. N.Y. Acad. Sci.*, in press.
13 Rodriguez-Tome, P. and Caterina, D. (1993) CEPH/Généthon.
14 Rodriguez-Tomé, P. (1995) EMBL-EBI.
15 Boguski, M.S., Lowe, T.M.J. and Tolstoshev, C.M. (1993) *Nature Genet.*, **4,** 332–333.
16 Pascarella, S. and Argos, P. (1992) *Protein Engng*, **5**, 121–137.
17 Jurka, J. and Smith, T. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 4775–4778.
18 Specht, T., Wolters, J. and Erdmann, V.A. (1991) *Nucleic Acids Res.*, **19**, 2189–2191.
19 Wallace, J.C. and Henikoff, S. (1992) *CABIOS*, **8**, 249–254.
20 Larsen, F., Gundersen, G., Lopez, R. and Prydz, H. (1992) *Genomics*, **13**, 1095–1107.
21 Wada, K., Wada, Y., Ishibashi, F., Gojobori, T. and Ikemura, T. (1992) *Nucleic Acids Res.*, **20**, 2111–2118.
22 Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) *Science*, **254**, 1434–1435.
23 Sander, C. (1993) EMBL, Heidelberg.
24 Wahl, R., Rice, P., Rice C.M. and Kröger, M. (1994) *Nucleic Acids Res.*, **22**, 3450–3455.
25 Bairoch, A. (1994) *Nucleic Acids Res.*, **22**, 3626–3627.
26 Holm, L., Ouzounis, C., Sander, C., Tuparev, G. and Vriend, G. (1992) *Protein Sci.*, **1**, 1691–1698.
27 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.
28 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.
29 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.
30 Sander, C. and Schneider, R. (1994) *Nucleic Acids Res.*, **22**, 3597–3599.
31 Kabat, E.A., Wu, T.T., Perry, H.M., Gottesman, K.S. and Foeller, C. (1992)
32 Keen, G., Redgrave, G., Lawton, J., Cinkosky, M., Mishra, S., Fickett, J. and Burks, C. (1992) *Mathl. Comput. Modelling*, **16**, 93–101.
33 Dölz, R., Mossé, M.-O., Bairoch, A., Slonimski, P.P. and Linder, P. (1994) *Nucleic Acids Res.*, **24**, 66–91.
34 McClelland, M, Nelson, M. and Raschke, E. (1994) *Nucleic Acids Res.*, **22**, 3640–3659.
35 Holm, L. and Sander, C. (1992) *J. Mol. Biol.*, **225**, 93–105.
36 Pattabiraman, N., Namboodiri, K., Lowrey, A. and Gaber, B.P. (1990) *Protein Seq. Data Anal.*, **3**, 387–405.
37 Perriere, G., Gouy, M. and Gojobori, T. (1994) *Nucleic Acids Res.*, **22**, 5525–5529.
38 Isohikhes, I. and Trifonov, E.N. (1993) *Nucleic Acids Res.*, **21**, 4857–4859.
39 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.
40 Hobohm, U., Scharf, M., Schneider, R. and Sander, C. (1992) *Protein Sci.*, **1**, 409–417.
41 Hanks, S.K. and Quinn, A.M. (1991) *Methods Enzymol.*, **200**, 38–62.
42 Attwood, T.K., Beck, M.E., Bleasby, A.J. and Parry-Smith, D.J. (1994) *Nucleic Acids Res.*, **22**, 3590–3596.
43 Sonnhammer, E.L.L. and Kahn, D. (1994) *Protein Sci.*, **3**, 482–492.
44 Holm, L. and Sander, C. (1994) *Proteins*, **19**, 256–268.
45 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.
46 Roberts, R.J. and Macelis, D. (1994) *Nucleic Acids Res.*, **22**, 3628–3639.
47 Raschke, E. (1993) *Genetic Analysis, Techniques and Applications*, **10**, 49–60.
48 Jurka, J., Walichiewicz, J. and Milosavljevic, A. (1992) *J. Mol. Evol.*, **35**, 286–291.
49 Lehrach, H. (1990) *Genome Analysis*, **1**, 39–81.
50 Neefs, J.M., Van de Peer, Y., De Rijk, P., Chapelle, S. and De Wachter, R. (1993) *Nucleic Acids Res.*, **21**, 3025–3049.
51 Pongor, S., Hátsági, Z., Degtyarenko, K., Fábián, P., Skerl, V., Hegyo, H., Myrvai, J. and Bevilacqua, V. (1994) *Nucleic Acids Res.*, **22**, 3610–3615.
52 Bairoch, A. (1994) University of Geneva.
53 Gu, J. and Reddy, R. (1994) *Nucleic Acids Res.*, **22**, 3481–3482.
54 Larsen, N. and Zwieb C. (1993) *Nucleic Acids Res.*, **21**, 3019–3020.
55 Ghosh, D. (1992) *Nucleic Acids Res.*, **20**, 2091–2093.
56 Brown, C.M., Stockwell, P.A., Dalphin, M.E. and Tate, W.P. (1994) *Nucleic Acids Res.*, **22**, 3620–3624.
57 Steinberg S., Misch, A. and Sprinzl M. (1993) *Nucleic Acids Res.*, **21**, 3011–3015.
58 Liebl, S. and Sonnhammer, E. (1994) MIPS, Germany and Sanger Centre, UK.
59 The EMBL Data Library (1993) *Nucleic Acids Res.*, **21**, i-vii.
60 The EMBL Data Library (1992) *Plant Mol. Biol.*,**18**, 1221–1224.
61 Fuchs, R. and Stoehr, P.J. (1993) *CABIOS* **9**, 71–77.
62 EMBL-EBI (1995).
63 Stoehr, P.J. and Omond, R.A. (1989) *Nucleic Acids Res.*, **17**, 6763–6764.
64 Etzold, T. and Argos, P. (1993) *CABIOS* **9**, 49–57.
65 Smith, T.F. and Waterman, M.S. (1981) *J. Mol. Biol.*, **147**, 195–197.
66 Pearson, W.R. and Lipman, D.J. (1988) *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.