Published by Oxford University Press *Nucleic Acids Research, 2002, Vol. 30, No. 1* ***21–26***

The EMBL Nucleotide Sequence Database

Guenter Stoesser*, Wendy Baker, Alexandra van den Broek, Evelyn Camon, Maria Garcia-Pastor, Carola Kanz, Tamara Kulikova, Rasko Leinonen, Quan Lin, Vincent Lombard, Rodrigo Lopez, Nicole Redaschi, Peter Stoehr, Mary Ann Tuli, Katerina Tzouvara and Robert Vaughan

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

Received September 20, 2001; Accepted September 28, 2001

ABSTRACT

The EMBL Nucleotide Sequence Database (aka EMBL-Bank; http://www.ebi.ac.uk/embl/) incorporates, organises and distributes nucleotide sequences from all available public sources. EMBL-Bank is located and maintained at the European Bioinformatics Institute (EBI) near Cambridge, UK. In an international collaboration with DDBJ (Japan) and GenBank (USA), data are exchanged amongst the collaborating databases on a daily basis. Major contributors to the EMBL database are individual scientists and genome project groups. Webin is the preferred web-based submission system for individual submitters, whilst automatic procedures allow incorporation of sequence data from large-scale genome sequencing centres and from the European Patent Office (EPO). Database releases are produced quarterly. Network services allow free access to the most up-to-date data collection via FTP, email and World Wide Web interfaces. EBI's Sequence Retrieval System (SRS), a network browser for databanks in molecular biology, integrates and links the main nucleotide and protein databases plus many other specialized databases. For sequence similarity searching, a variety of tools (e.g. Blitz, Fasta, BLAST) are available which allow external users to compare their own sequences against the latest data in the EMBL Nucleotide Sequence Database and SWISS-PROT. All resources can be accessed via the EBI home page at http://www.ebi.ac.uk.

INTRODUCTION

The European Bioinformatics Institute (EBI) is an Outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, Germany. The EBI is located in the grounds of the Wellcome Trust Genome Campus near Cambridge, UK, next to the Sanger Centre and the UK Human Genome Mapping Project Resource Centre.

The main missions of the Service Programme of the EBI (1) centre on building, maintaining and providing biological databases and information services to support data deposition and exploitation. In this respect a number of databases are operated, namely the EMBL Nucleotide Sequence Database (EMBL-Bank), the Protein Databases (SWISS-PROT and TrEMBL), the Macromolecular Structure Database (MSD) and ArrayExpress for gene expression data plus several other databases many of which are produced in collaboration with external groups.

The EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/) is the European member of the tri-partide International Nucleotide Sequence Database Collaboration DDBJ/EMBL/GenBank. Main data sources are large-scale genome sequencing centres, individual scientists and the European Patent Office (EPO). Direct submissions to EMBL-Bank are complemented by daily data exchange with collaborating databases DDBJ (Japan) (2) and GenBank (USA) (3).

The EMBL database is growing rapidly as a result of major genome sequencing efforts. Within a 12 month period the database size has increased from about 6.7 million entries comprising 8255 million nucleotides (Release 63, June 2000) to over 12 million entries and 12 820 million nucleotides (Release 67, June 2001). During the same period the number of organisms represented in the database has risen by >30% to over 75 000 species.

SEQUENCE SUBMISSIONS

Why submit to EMBL-Bank?

Submission of new sequence data and update information to the public database is an essential prerequisite for building and maintaining a complete and up-to-date data set allowing the scientific community to perform similarity searches and analysis on the latest nucleotide and protein sequence data. This comprehensive repository of up-to-date primary sequences is essential for further computational analysis and genome research. Discovery of novel genes, identification of homologous genes, analysis of alternative splicing and detection of polymorphisms are only some of the uses of the database in the context of biomedical research, and this will only increase as large-scale sequencing efforts keep on depositing more high-throughput sequence data (HTG) and as more

*To whom correspondence should be addressed. Tel: +44 1223 494466; Fax: +44 1223 494472; Email: stoesser@ebi.ac.uk

complete genomes are being added to the database. Bioinformatics tools for database searching, sequence and homology searching, gene prediction, multiple sequence alignments, etc., are made available from the EBI allowing *in silico* analysis.

How to submit to EMBL-Bank

Direct submissions from individual scientists or sequencing groups is an important source of new nucleotide sequences and descriptive biological information. Often these submissions will include feature and function information based on experimental research, while genome project submissions typically include preliminary annotations based on prediction algorithms. Prior to submission, a World Wide Web-based interactive vector scanning service allows submitters to screen sequences for vector contamination.

Webin. Webin is EMBL's preferred web-based submission system for nucleotide sequences and biological annotation information. Webin is designed to allow fast submission of single, multiple or very large numbers of sequences (bulk submissions) and is available at http://www.ebi.ac.uk/embl/Submission/webin.html.

Sequin. Sequin is a stand-alone software tool developed by the NCBI for submitting and updating nucleotide sequences to the GenBank, EMBL or DDBJ databases. Sequin runs on Macintosh, PC/Windows and UNIX computers.

Genome project submissions. Large-scale sequencing projects have become the major source of new sequence data. EBI works closely with sequencing centers to ensure timely incorporation of these data into EMBL for public release. 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 genome data acquisition with the genome projects in the Sanger Centre, one of the most productive sequencing centres in Europe and world-wide. Database entries produced at the research sites are deposited and updated directly by the genome project groups using FTP or email. 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 represent 'unfinished' or 'finished' sequence data. New projects can be accommodated easily. Groups wishing to open accounts to submit genome sequence data should contact the database at datasubs@ebi.ac.uk. More information is available from http://www3.ebi.ac.uk/Services/GenomeSubm/.

Updating existing database entries. Over time an entry which was correct when created may become out of date: authors may make corrections to the sequence itself, or may discover new features which require annotation. Since such findings are often not published, it is important that authors communicate their new findings to the database. The preferred option is via the World Wide Web update form at http://www3.ebi.ac.uk/Services/webin/update/update.html.

Sequence alignment submissions. Webin-Align is EMBL's interactive web-based system for submission of alignment data from phylogenetic and population analysis of nucleotide sequences. Unique alignment numbers (e.g. DS32096) are assigned to each alignment submission and should be included in the published article. Currently accepted standard alignment formats include NEXUS, PHYLIP, CLUSTAL and GCG/MSF or SEQUIN/ASN.1 output. Nucleotide alignment data can be retrieved from the EBI's World Wide Web pages at http://www3.ebi.ac.uk/Services/align/listali.html or from the FTP server at ftp://ftp.ebi.ac.uk/pub/databases/embl/align. Submission information is available from http://www.ebi.ac.uk/embl/Submission/.

Sequences from patent literature. Patent Sequences are being captured in an ongoing collaboration with the EPO. The EPO's policy is to release data to the public (and to EMBL) 18 months after the patent application date, regardless of whether a patent has been granted or not. Immediately after release by the EPO the latest patent sequence data are integrated into the EMBL database and made available to the public. All entries derived from the EPO patent literature are available from ftp://ftp.ebi.ac.uk/pub/databases/embl/patent/. Additionally, these files include data from American and Japanese patent literature incorporated from NCBI (USA) and DDBJ (Japan).

Accession numbers and data confidentiality. Accession numbers are unique identifiers which permanently identify sequences in the database. Most journal editors require submission of sequence data to the DDBJ/EMBL/GenBank prior to journal publication. In response to a new submission, accession numbers are assigned and communicated to authors within two working days of receipt of submission. These accession numbers (e.g. X56734 and AL450380) are included in the manuscript/publication, thus allowing the community to retrieve the data upon reading the journal article.

During the submission process, submitters are prompted to specify whether their submitted data can be made available to the public immediately or whether the data should be withheld until an author-specified date. Data are never withheld after publication.

Table 1 provides EMBL-Bank web-based resources including detailed information on submissions, data access, genome data and database searching and analysis tools.

BIOLOGICAL ANNOTATION AND CURATION

The importance of careful curation of individual sequences submitted directly by individual researchers and discussed in the scientific literature is obvious. Such sequences have often been the subject of experimental research elucidating features and function, while genome project submissions in most cases will 'only' include preliminary gene annotations based on gene prediction programs. One of the most important features in EMBL entries is the protein coding sequence (CDS). All CDS features within EMBL entries are translated and subsequently added to the TrEMBL and SWISS-PROT protein sequence databases (4). Furthermore, coding regions in EMBL entries are cross-referenced (via /db_xref qualifier) to the protein databases, and where appropriate EMBL entries are linked to other specialised databases (e.g. species-specific databases). These cross-references allow access to additional information

Table 1. Summary of EMBL-Bank web-based resources including detailed information on submissions, data access, genome data as well as database searching and analysis tools

Title	URL
General	
EMBL-EBI Home Page	www.ebi.ac.uk/
EMBL Nucleotide Sequence Database	www.ebi.ac.uk/embl/
Documentation	
EMBL Database Documentation Page	www.ebi.ac.uk/embl/Documentation/
Database User manual	www.ebi.ac.uk/embl/Documentation/User_manual/usrman.html
Database Release Notes	www.ebi.ac.uk/embl/Documentation/Release_notes/current/relnotes.html
The DDBJ/EMBL/GenBank Feature Table Document	www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html
Taxonomy Database	www3.ncbi.nlm.nih.gov/Taxonomy/tax.html
Example Database Entry	www.ebi.ac.uk/cgi-bin/emblfetch?x64011
WEBALIGN: Sequence Alignment Submissions	www.ebi.ac.uk/embl/Submission/alignment.html
WebFeat: feature table keys/qualifiers definitions	www3.ebi.ac.uk/Services/WebFeat/
Annotation Examples: EMBL entry examples.	www3.ebi.ac.uk/Services/Standards/web/
DE line Standards: guidelines for entry definitions	www.ebi.ac.uk/embl/Documentation/de_line_standards.html
Sites maintaining daily updated copies of EMBL	www.ebi.ac.uk/embl/Access/other_sites.html
Submissions	
Submission of Nucleotide Sequence Data	www.ebi.ac.uk/embl/Submission/
Information for Submitters Document	www.ebi.ac.uk/embl/Documentation/information_for_submitters.html
Vector Scanning prior to submission	www2.ebi.ac.uk/blastall/vectors.html
WEBIN: web-based sequence submission system	www.ebi.ac.uk/embl/Submission/
SEQUIN: stand-alone sequence submission tool	www3.ebi.ac.uk/Services/Sequin/
Genome Project Submission Account guidelines	www3.ebi.ac.uk/Services/GenomeSubm/
WEBUP: sequence update form	www3.ebi.ac.uk/Services/webin/update/update.html
Access	
Access to servers, query tools and data archives	www.ebi.ac.uk/embl/Access/
Sequence Retrieval Service (SRS)	http://srs.ebi.ac.uk/
Current Database Release	ftp://ftp.ebi.ac.uk/pub/databases/embl/release/
Sequence Tagged Sites (STS) resources	www.ebi.ac.uk/embl/Access/sts.html
Expressed Sequence Tag (EST) resources	www.ebi.ac.uk/embl/Access/est.html
Sequences from the patent literature	ftp://ftp.ebi.ac.uk/pub/databases/embl/patent/
Genome data	
Completed Genomes Web Server	www.ebi.ac.uk/genomes/
Genome FTP server	ftp://ftp.ebi.ac.uk/pub/databases/embl/genomes/
Ensembl: automated analysis of genome data	http://ensembl.ebi.ac.uk/
Genome MOT: status of genome projects	www.ebi.ac.uk/Databases/Genome_MOT/genome_mot.html
Database searching, browsing and analysis tools	
Access to searching, browsing and analysis tools	www.ebi.ac.uk/Tools/

concerning the entry that is more appropriately stored in other dedicated databases.

Current database policy is to reject submissions for which no sequence annotation has been provided, unless these describe ESTs or unfinished HTGs. Both Webin and Sequin allow addition of sequence annotation—any number of relevant features can be easily added to the sequence feature table via the according feature forms. A team of biological curators review and check newly submitted data ensuring all mandatory information has been provided, that biological features are adequately described and that the conceptual translations of any coding regions follow genetic translation rules.

To assist submitters annotate their sequences, usage guides are available from the EMBL-EBI World Wide Web site and from within Webin.

WebFeat. A complete list of feature table key and qualifier definitions, providing full explanations of their use.

EMBL annotation examples. A selection of EMBL approved feature table annotations for some common biological sequences (e.g. ribosomal RNA, mitochondrial genome).

DE line standards. Guidelines and database conventions for creating suitable descriptions for submissions.

Automatic genome annotation

Ensembl (5) is a joint project between EMBL–EBI and the Sanger Centre to produce and maintain automatic annotation on eukaryotic genomes. Access is presented via a web-based genome browser at http://www.ensembl.org/.

DATA MANAGEMENT AND REPRESENTATION

Data is managed in a robust database management system (ORACLE), using a schema which facilitates integration and interoperability with other databases, especially protein sequences. Quarterly releases and daily updates for distribution and installation at remote sites are generated from this system.

Database entries are distributed in EMBL flat-file format which is supported by most sequence analysis software packages and also provides a structure that is easy to read. The EMBL flat file comprises of a series of strictly controlled line types presented in a tabular manner and consisting of four major blocks of data:

1. Descriptions and identifiers.
2. Citations: citation details of the associated publications and the name and contact details of the original submitter.
3. Features: detailed source information, biological features comprised of feature locations, feature qualifiers, etc.
4. Sequence: total sequence length, base composition (SQ) and sequence.

For details see the User Manual at http://www.ebi.ac.uk/embl/Documentation/.

Sequence identifiers

In addition to unique and stable accession numbers, EMBL database entries include sequence identifiers and version information which specify changes in sequences. The identifiers themselves remain stable within a given entry, whilst the version number increments with every sequence update. 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.

Nucleotide sequence identifier Example: SV AJ400848.1
Protein sequence identifier Example: protein_id="CAB88705.1"

Protein translations

Translations of protein coding regions represented by CDS features in EMBL entries are automatically added to the TrEMBL protein database. From these entries, SWISS-PROT curators subsequently create the SWISS-PROT database entries. EMBL nucleotide entries are cross-referenced (via the /db_xref qualifier) to the TrEMBL and SWISS-PROT databases.

Data integration

Links to external databases allow integration with specialised data collections, such as protein databases, species-specific databases, taxonomy databases, etc. The World Wide Web-based sequence retrieval system (SRS) enables users to easily navigate between cross-referenced database entries. Where appropriate, EMBL Database entries are cross-referenced to other databases like the Eukaryotic Promoter Database (6), TRANSFAC (7), IMGT (8), FlyBase (9), TrEMBL and SWISS-PROT. Cross-references to external databases are represented in the EMBL flat file line type 'DR' and, where appropriate, at the feature level via the feature qualifier /db_xref. For example, in the context of integrating novel unfinished high-throughput mouse cDNA sequences (HTC) about 20 000 cross-links to the Mouse Genome Database (MGD) have been created.

Database divisions

The EMBL Database currently consists of 19 divisions with each entry belonging to exactly one division. The division is indicated using three letter codes, e.g. PRO, Prokarotes; HUM, Human; PHG, Bacteriophages; PLN, Plants; etc. The grouping is mainly based on taxonomy with a few exceptions like the HTG, EST, sequence tagged sites (STS) and genome survey sequences (GSS) divisions and the recently created new database division 'HTC' representing unfinished high-throughput cDNA sequences. For these divisions, grouping is based on the specific nature of the underlying data.

GENOME REPRESENTATION

Completed genomes web server

Direct access to hundreds of completed genome sequences plus according protein translations is available at http://www.ebi.ac.uk/genomes/.

Recent additions include: *Buchnera* sp. APS, accession no. BA000003; *Pseudomonas aeruginosa*, AE004091; *Pasteurella multocida*, AE004439; *Lactococcus lactis*, AE005176; *Mycobacterium leprae*, AL450380; *Mesorhizobium loti*, BA000012; *Escherichia coli* O157, BA000007; *Agrobacterium tumefaciens*, AE007869; *Mycoplasma pulmonis*, AL445566; *Sinorhizobium meliloti* 1021, AL591688; *Caulobacter crescentus*, AE005673; *Aeropyrum pernix*, BA000002; *Pyrococcus abyssi*; AL096836; *Pyrococcus horikoshii*, BA000001; *Sulfolobus solfataricus*, AE006641.

HTGs

'Unfinished' DNA sequences generated by the high-throughput sequencing centers are represented in the HTG division and are rapidly made available to the scientific community for homology searches. Entries in this division all contain keywords to indicate the status of the sequencing (e.g. HTGS_PHASE1). 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. Once 'finished', HTG sequences are moved into the appropriate primary EMBL taxonomic division.

Base quality values

Quality scores (Phrap) from draft HTG data are available on the EBI FTP server at ftp://ftp.ebi.ac.uk/pub/databases/embl/quality_scores. The gzipped files in the directory contain base quality values for unfinished human sequences from Japanese, US and European sequencing centres. The FastA-type headers contain the EMBL sequence identifiers and versions of the corresponding database entries. Quality score files are updated on a daily basis.

ESTs

ESTs (single pass cDNA reads) constitute a major source of sequence records: of the 12 million entries in Release 67 (June 2001) more than 8 million entries represented EST data, the vast majority originating from human and mouse. In addition to the EST division files in the EMBL database release, EBI's ESTLIB provides further information about the libraries from which EST sequences were derived. The according EST division entries in EMBL are cross-referenced to ESTLIB with a /db_xref qualifier on the source feature, e.g. /db_xref="ESTLIB:863". An EBI mirror of NCBI's dbEST resources is available from ftp://ftp.ebi.ac.uk/pub/databases/dbEST/.

GSSs

GSSs are of similar nature to EST data, except that its sequences are genomic rather than cDNA (mRNA). The GSS division contains, for example, random 'single pass read' genome survey sequences, single pass reads from cosmid/BAC/YAC ends, exon trapped genomic sequences and Alu PCR sequences.

STS

The STS division contains sequence and mapping data on short genomic landmark sequences or STS.

Draft human genome

The completion of the human draft genome sequence was announced and published in February 2001 (10,11). The EMBL Database (together with GenBank and DDBJ) has been playing a key role in acquisition, storage and distribution of human genome sequence data.

EBI's genome web server. This server provides easy access to hundreds of completed genome sequences and is available at http://www.ebi.ac.uk/genomes/.

EMBL release HTG division. Since the beginning of the Human Genome Project, the international Human Genome Sequencing Consortium has been submitting human draft sequence data to the International Nucleotide Sequence Databases DDBJ/EMBL/GenBank. High-throughput human sequence data are incorporated into the EMBL Database HTG division and are made available to the public immediately from the EBI servers.

Genome Monitoring Table (Genome MOT). Unfinished and finished human data sorted by chromosome are available via EBI's Genome MOT (12) at http://www.ebi.ac.uk/genomes/mot/.

Ensembl. Automatic annotation, graphical views, web-searchable data sets including information on confirmed peptides, confirmed cDNAs, predicted peptides, repeat predictions along with integration of map information and SNPs are available from http://www.ensembl.org/.

Human proteome information. Human proteome information is available from SWISS-PROT at http://www.ebi.ac.uk/proteome/HUMAN/.

DATA DISTRIBUTION, SEARCHING AND SEQUENCE ANALYSIS

EBI network services

Database releases are produced quarterly and integrated into the EBI's SRS server. Databases and software can also be downloaded from the EBI's FTP server. EBI's network services allow access to the most up-to-date data collection via the Internet. Data access to EMBL nucleotide sequence data is also granted via email using the netserver or interactively via the World Wide Web where the main service comprises the SRS server.

SRS

The SRS server at the EBI integrates and links a comprehensive collection of specialised databanks along with the main nucleotide and protein databases. The SRS system (13) allows the databases to be searched using a number of fields including sequence annotations, keywords and author names. Complex querying and linking across all available databanks can also be executed and users should refer to the detailed instructions which are available online at http://srs.ebi.ac.uk/.

Sequence searching

The EBI provides a comprehensive set of sequence similarity algorithms that can be accessed both interactively from the EMBL–EBI World Wide Web site (http://www.ebi.ac.uk/Tools/) or by email. The EMBL Nucleotide Sequence Database can be searched as a whole or by individual taxonomic division. The most commonly used algorithms available are Fasta (14) and WU-Blast (15; see the WU-blast HELP page). Fasta will find a single high-scoring gapped alignment between the query nucleotide sequence and database sequences. Comparisons between a nucleotide sequence and the protein databases can be made using fastx/y, whilst tfastx/y allows comparisons between a protein sequence and the translated DNA databank. The EBI's Smith and Waterman (16) service comprises a comprehensive set of programs. These include today MPsrch (see help page), Edinburgh Biocomputing Systems (EBS) and Scanps (see help page). These facilitate more sensitive searches against protein sequence databases. In total, more than 200 databases are available for searching at the EBI. The new Fasta service for genomes and proteomes enables users to search on complete genomes and derived proteomes from public sequencing projects around the world. This service is produced in an collaborative effort involving the EMBL-Bank, SWISS-PROT/TrEMBL and Ensembl groups.

Sequence analysis

Specialised sequence analysis programs are also available from the EBI. Such services include multiple sequence alignment and inference of phylogenies using CLUSTALW (17), gene prediction using GeneMark (18), pattern searching and discovery using PRATT (19), motif identification using ppsearch (see help page) as well as applications which have been developed in-house for various other projects. EBI is in the process of adding more interactive sequence analysis resources based on the European Molecular Biology Open Software Suite (EMBOSS) (http://www.emboss.org/). The EBI is involved as an active collaborator and some of the applications, e.g. specialised tools for detecting CpG Islands, are already available.

EMBnet

The European Molecular Biology Network (http://www.embnet.org) was initiated in 1988 to link major European laboratories which provide bioinformatics to national scientific communities as well as being involved in active R&D in the fields of sequence analysis. One of the main tasks of the EMBnet network is the maintenance and updating of remote copies of the nucleotide and protein sequence databases which are updated daily. As bioinformatics grows, EMBnet plays an important role in providing a comprehensive program of bioinformatics training aimed specifically at both the wet lab researcher as well as programmers and systems administrators. A listing of mirror sites maintaining daily updated copies of the EMBL Database is available from the EBI at http://www.ebi.ac.uk/embl/Access.

CITING THE EMBL DATABASE

Please cite this article when referring to the EMBL Nucleotide Sequence Database.

CONTACTING THE EMBL DATABASE

Computer network: data submissions, datasubs@ebi.ac.uk; general inquiries, datalib@ebi.ac.uk; updates/publication notifications, update@ebi.ac.uk. Postal address: EMBL Nucleotide Sequence Submissions, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK. Tel: data submissions, +44 1223 494499; general, +44 1223 494444. Fax: data submissions, +44 1223 494472; general, +44 1223 494468.

REFERENCES

1. Emmert,D.B., Stoehr,P.J., Stoesser,G. and Cameron,G.N. (1994) The European Bioinformatics Institute (EBI) databases. *Nucleic Acids Res.*, **22**, 3445–3449.
2. Tateno,Y., Miyazaki,S., Ota,M., Sugawara,H. and Gojobori,T. (2000) DNA Data Bank of Japan (DDBJ) in collaboration with mass sequencing teams. *Nucleic Acids Res.*, **28**, 24–26. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 27–30.
3. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J., Rapp,B.A. and Wheeler,D.L. (2000)GenBank. *Nucleic Acids Res.*, **28**, 15–18. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 17–20.
4. Bairoch,A. and Apweiler,R. (2000) The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Res.*, **28**, 45–48.
5. Hubbard,T., Barker,D., Birney,E., Cameron,G., Chen,Y., Clark,L., Cox,T., Cuff,J., Curwen,V., Down,T. *et al.* (2002) The Ensembl genome database project. *Nucleic Acids Res.*, **30**, 38–41.
6. Périer,R.C., Praz,V., Junier,T., Bonnard,C. and Bucher,P. (2000) The Eukaryotic Promoter Database (EPD). *Nucleic Acids Res.*, **28**, 302–303. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 322–324.
7. Wingender,E., Chen,X., Hehl,R., Karas,H., Liebich,I., Matys,V., Meinhardt,T., Prüß,M., Reuter,I. and Schacherer,F. (2000) TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res.*, **28**, 316–319.
8. Ruiz,M., Giudicelli,V., Ginestoux,C., Stoehr,P., Robinson,J., Bodmer,J., Marsh,S.G.E., Bontrop,R., Lemaitre,M., Lefranc,G. *et al.* (2000) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Res.*, **28**, 219–221.
9. The FlyBase Consortium (1999) The FlyBase database of the *Drosophila* Genome Projects and community literature.*Nucleic Acids Res.*, **27**, 85–88. Updated article in this issue: *Nucleic Acids Res.* (2002), **30**, 106–108.
10. The Genome International Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
11. Venter,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A. *et al.* (2001) The sequence of the Human Genome. *Science*, **291**, 1304–1351.
12. Beck,S. and Sterk,P. (1998) Genome-scale DNA sequencing where are we? *Curr. Opin. Biotechnol.*, **9**, 116–121.
13. Etzold,T., Ulyanov,A. and Argos,P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.
14. Pearson,W.R. (1994) Using the FASTA program to search protein and DNA sequence databases. *Methods Mol. Biol.*, **24**, 307–331.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Smith,R.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Applied Math.*, **2**, 482–489.
17. Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
18. Borodovsky,M. and McIninch,J. (1993) GeneMark: parallel gene recognition for both DNA strands. *Comput. Chem.*, **17**, 123–133.
19. Jonassen,I., Collins,J.F. and Higgins,D.G. (1995) Finding flexible patterns in unaligned protein sequences. *Protein Sci.*, **4**, 1587–1595.