DOI: 10.1093/nar/gkh405

Public web-based services from the European Bioinformatics Institute

Nicola Harte, Ville Silventoinen, Emmanuel Quevillon, Stephen Robinson, Kimmo Kallio, Xavier Fustero, Pravin Patel, Petteri Jokinen and Rodrigo Lopez*

European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received February 13, 2004; Revised and Accepted March 29, 2004

ABSTRACT

The mission of the European Bioinformatics Institute (EBI), an outstation of the European Molecular Biology Laboratory (EMBL) in Heidelberg, is to ensure that the growing body of information from molecular biology and genome research is placed in the public domain and is accessible freely to all parts of the scientific community in ways that promote scientific progress. To fulfil this mission, the EBI provides a wide variety of free, publicly available bioinformatics services. These can be divided into data submissions processing; access to query, analysis and retrieval systems and tools; ftp downloads of software and databases; training and education and user support. All of these services are available at the EBI website: http://www.ebi.ac.uk/services. This paper provides a detailed introduction to the interactive analysis systems that are available from the EBI and a brief introduction to other, related services.

INTRODUCTION

The European Bioinformatics Institute (EBI) is committed to providing free and unrestricted access to the biological databases that it produces and maintains, many in association with other international institutions. These databases contain data from a wide range of areas and include nucleic acids and protein sequences, and experimental data on their functions, structures and expression. The wealth of biological data available is accessible for browsing and retrieval via the WWW, ftp and email servers and programmatic interfaces using XML/SOAP-based web services. Many popular bioinformatics analysis tools are available. These include sequence similarity searches as well as structure and function prediction algorithms. A number of these services are provided in collaboration with academic and commercial software designers. In delivering these services, the EBI takes into account the following key factors: cost-effectiveness, scientific quality, level of service, speed and, ultimately, real value for the biological community. The EBI guarantees that usage of its services is fully confidential, secure and in accordance with internationally agreed standards and procedures.

EBI TOOLS

The EBI toolbox contains a wide range of services that allow the user to search, extract, manipulate and analyse data (see Table 1). The services are aimed primarily at analysts, but they can also be used to assist data submitters to enrich, complement and validate their data. The toolbox can be divided into the following five categories: (i) similarity searches; (ii) protein function analysis; (iii) sequence analysis; (iv) structural analysis and (v) miscellaneous.

Each of these services has a number of options that allow them to be used by a wide range of users, from laboratory technicians assessing the quality of their sequences to the most discerning structural biologists. Detailed help pages about the individual services and their various parameters are provided. In addition, tutorials are available for the majority of the tools in the '2can Bioinformatics' educational resource. This comprehensive documentation makes these services ideal for training and education purposes. The EBI encourages educators to contact the support help desk at http://www.ebi.ac.uk/support/ if they plan to use any of them during their courses.

Similarity search tools

There are five main similarity search algorithms available at the EBI: FASTA (1), WU-BLAST2 (2), NCBI-BLAST2 (3), MPsrch (4) and SCANPS (5). These can be used to identify similarities between novel query sequences and database sequences whose structure and function have been elucidated. These tools allow searches against a large number of databases including the EMBL Nucleotide Sequence Database (6); IMGT/LIGM (7); the UNIPROT protein sequence resource (8); IMGT/HLA; complete genomes and proteomes; patented

*To whom correspondence should be addressed. Tel: +44 1223 494423; Fax: +44 1223 494468; Email: rls@ebi.ac.uk

Table 1. An alphabetical list of the tools available from the EBI, their input and URL

Tool	Input	URL
Similarity searches		
BLAST2-NCBI	n, p	http://www.ebi.ac.uk/blastall/
BLAST2-WU	n, p	http://www.ebi.ac.uk/blast2/
BLAST2-WU parasite	n	http://www.ebi.ac.uk/blast2/parasites.html
FASTA	n, p	http://www.ebi.ac.uk/fasta33/nucleotide.html
FASTA-Genome/Proteome	n, p	http://www.ebi.ac.uk/fasta33/proteomes.html
FASTA-SNP	n	http://www.ebi.ac.uk/snpfasta3/
MPsrch	p	http://www.ebi.ac.uk/MPsrch/
Scanps2.3	p	http://www.ebi.ac.uk/scanps/
Protein Function Analysis		
CluSTr search	p	http://www.ebi.ac.uk/clustr/search.html
GeneQuiz	p	http://jura.ebi.ac.uk:8765/ext-genequiz/
InterProScan	n, p	http://www.ebi.ac.uk/InterProScan/
RADAR	p	http://www.ebi.ac.uk/Radar/index.html
Sequence Analysis		
Align	n, p	http://www.ebi.ac.uk/emboss/align/index.html
ClustalW	n, p	http://www.ebi.ac.uk/clustalw/index.html
CpGPlot/CpGReport/Isochore	n	http://www.ebi.ac.uk/emboss/cpgplot/
GeneMark	n	http://www.ebi.ac.uk/genemark/
GeneWise	p	http://www.ebi.ac.uk/Wise2/index.html
Pepinfo/Pepstats/Pepwindow	p	http://www.ebi.ac.uk/emboss/pepinfo/
PromoterWise	n	http://www.ebi.ac.uk/Wise2/promoterwise.html
Transeq	n	http://www.ebi.ac.uk/emboss/transeq/
Protein Structure Analysis		
DALI	c	http://www.ebi.ac.uk/dali/
MaxSprout	c	http://www.ebi.ac.uk/maxsprout/
MSD Services	p, c	http://www.ebi.ac.uk/msd/Services.html
Miscellaneous		
EMBL Computational Services	n, p	http://www.ebi.ac.uk/embl_services/index.html
Expression profiler	e	http://www.ebi.ac.uk/expressionprofiler/
QuickGO	t	http://www.ebi.ac.uk/ego/index.html
Readseq	n, p	http://www.ebi.ac.uk/cgi-bin/readseq.cgi
Data Retrieval		
Ensembl	t, n, p	http://www.ebi.ac.uk/ensembl/index.html
Dbfetch	a, i	http://www.ebi.ac.uk/cgi-bin/dbfetch
Web services	a, i	http://www.ebi.ac.uk/Tools/webservices/
SRS	t, n, p	http://srs.ebi.ac.uk
SRS3D	t, n, p	http://srs3d.ebi.ac.uk

Key: ‘n’ nucleotide, ‘p’ protein, ‘c’ structural coordinates, ‘e’ expression data, ‘t’ text, ‘a’ accession number, ‘i’ identifier.

sequences from the European Patent Office, the US Patent and Trademarks Office and the Japanese Patent Office, and the structures database PDB.

The EBI deploys this range of similarity tools because each of the algorithms provides a unique sequence search environment. FASTA, WU-BLAST2 and NCBI-BLAST2 can be used for both nucleotide and protein searches, while MPsrch and SCANPS are for protein searches only.

(i) FASTA is a set of programs that allow sequence similarity searching against both nucleotide and protein databases. FASTA has been found to be better than BLAST for conducting nucleotide searches since a smaller k-tuple than the minimum of seven for BLAST can be implemented. The program can be very specific when identifying long regions of low similarity, especially for highly diverged sequences. In addition to the searches against the primary sequence databases, FASTA searches can be conducted against a number of specialized databases, including complete genome and proteome databases and the European Molecular Biology Laboratory (EMBL) Whole Genome Shotgun (WGS) database. The SNP FASTA service conducts searches against the human genome variation database HGVBASE (9). FASTF and FASTS can be used to compare protein sequences obtained from peptide degradation and mass spectrometry experiments, respectively, against entire databases.

(ii) The EBI supports two versions of the BLAST package: WU-BLAST2 and NCBI-BLAST2. These share the same fundamental principles and have a common lineage for some portions of their code but offer different features. BLAST is the fastest of the sequence similarity search programs. WU-BLAST2 stands for Washington University Basic Local Alignment Search Tool version 2.0. The most important feature of this version at the EBI is the sensitivity parameter. Users can select a sensitivity setting, thus allowing them to control the overall performance of the program. Increasing the sensitivity means the search will take longer but be more selective. In the EBI WU-BLAST2 service, the gap open and extension penalties vary depending on the matrix and sensitivity settings selected. WU-BLAST2 can be used for both protein and nucleotide searches. The parasite BLAST service allows searches against the genomes of parasites. NCBI-BLAST2 offers the user the option of six different alignment outputs, adjustable gap open and extension penalties and a choice of filtering programs.

A very useful feature of NCBI-BLAST2 is the option of displaying the top predetermined number of hits in multiple sequence alignment form. This BLAST version is available for both protein and nucleotide searches. Sequences can be checked for vector contamination by searching against a special database known as EMVEC. This is an extraction of sequences from the SYNthetic and other divisions of EMBL.

(iii) MPsrch is a protein sequence comparison tool that implements the true Smith and Waterman algorithm. It is the most sensitive protein sequence similarity search tool available and the most reliable one for detecting more distantly related members. It also reports fewer false positive hits than other tools.

(iv) SCANPS (Scan Protein Sequence) is a protein sequence similarity search program. It also implements full Smith–Waterman-style searching and is capable of identifying multiple domain matches by using iterative profile searching, similar to the psi-blast approach.

Protein function analysis tools

The protein function analysis tools search for matches to the query sequence in the so-called secondary databases. These secondary databases contain the results of functional analysis of the sequences in the primary databases. Each one analyses the primary databases differently and as a result contains different information. Examples of secondary databases are Prosite (10), Pfam (11), PRINTS (12) and InterPro (13). When searches of these databases yield significant matches, these hits help to assign the query protein to a particular family. If the structure and function of the family are known, searches of the secondary databases offer a fast track to inferring biological function.

(i) InterProScan (14) is the most popular protein function analysis tool available from the EBI. It combines different protein function recognition methods into one resource. It scans a given sequence against the protein families of the InterPro member databases. The input to InterProScan can be either a nucleotide or a protein sequence. With nucleotides, the user is given an option of 15 translation tables and must choose a minimum open reading frame (ORF) size. InterProScan launches nine applications that search against specific databases (see Table 2) and have preconfigured cutoff thresholds. Users have the choice of selecting which applications to launch and in this way can customize their search. Each application returns a list of hits to the individual databases. The result is returned as an XML file. This file is processed to generate a graphical and a table view of the results (see Figures 1 and 2). There is also a perl-based version of InterProScan that can cope with bulk data processing. This version, the underlying applications, the databases used and all indexes of the data are freely available from the EBI's ftp server under the GNU licence agreement. InterproScan can be launched from three locations: the EBI tools page, within the Sequence Retrieval System (15) (protein sequence only) and by email.

Table 2. The core applications and databases of InterProScan

Program	Database
BlastProDom	ProDom
FingerPrintScan	PRINTS
HMMPIR	PIR
HMMPfam	Pfam
HMMSmart	SMART
HMMTigr	TIGRfams
ProfileScan	Prosite profiles
ScanReqExp	Prosite
SuperFamily	Superfamily

(ii) CluSTr search offers an automatic classification of UniProt Knowledgebase proteins into groups of related proteins by searching the CluSTr database (16) with free-text queries, accession numbers or identifiers.

(iii) GeneQuiz (17) is an integrated system for large-scale biological sequence analysis. It goes from a protein sequence to a biochemical function using a variety of search and analysis methods and up-to-date protein and DNA databases.

(iv) RADAR (18) stands for Rapid Automatic Detection and Alignment of Repeats in protein sequences. RADAR uses an automatic algorithm to segment a query sequence into repeats. It does this by identifying short compositional biases, gapped approximate repeats and complex repeat architectures.

Sequence analysis tools

This set of tools allows the user to carry out further, more detailed analysis of novel and existing sequences. including multiple sequence alignments, pairwise alignments and the determination of signalling regions, coding regions and ORF quality. The identification of these and other biological properties is a clue that aids the search to elucidate the specific function of a sequence. A number of the tools provided are from the EMBOSS package (19).

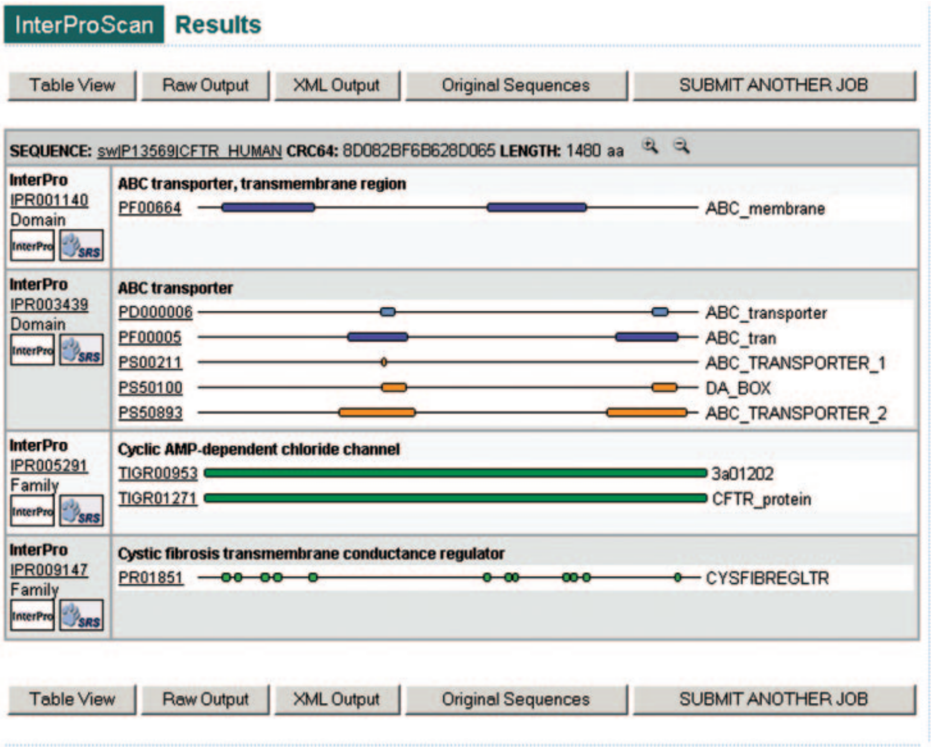
(i) ClustalW (20) is a general-purpose multiple sequence alignment tool for nucleotides and proteins. The ClustalW service (21) is one of the most popular of the EBI tools. It provides access to single-CPU and parallel versions of the software, thus catering for a wide range of uses, including the alignment of large sets of data such as complicated viral and bacterial genomes. The ClustalW service at the EBI is the only one offering the user ClustalW's ability to infer phylogenies.

(ii) Align compares two nucleotide or protein sequences. Needle is used to obtain an alignment that covers the length of both sequences, while Water is used to find the best region of similarity between two sequences.

(iii) Transeq translates a nucleic acid sequence to the corresponding peptide sequence. It can translate in all six reading frames.

(iv) Pepinfo, Pepwindow and Pepstats are tools that can be used to obtain general information about a protein sequence, including molecular weight, isoelectric point, charge, average residue weight, hydrophobicity values and physico-chemical properties.

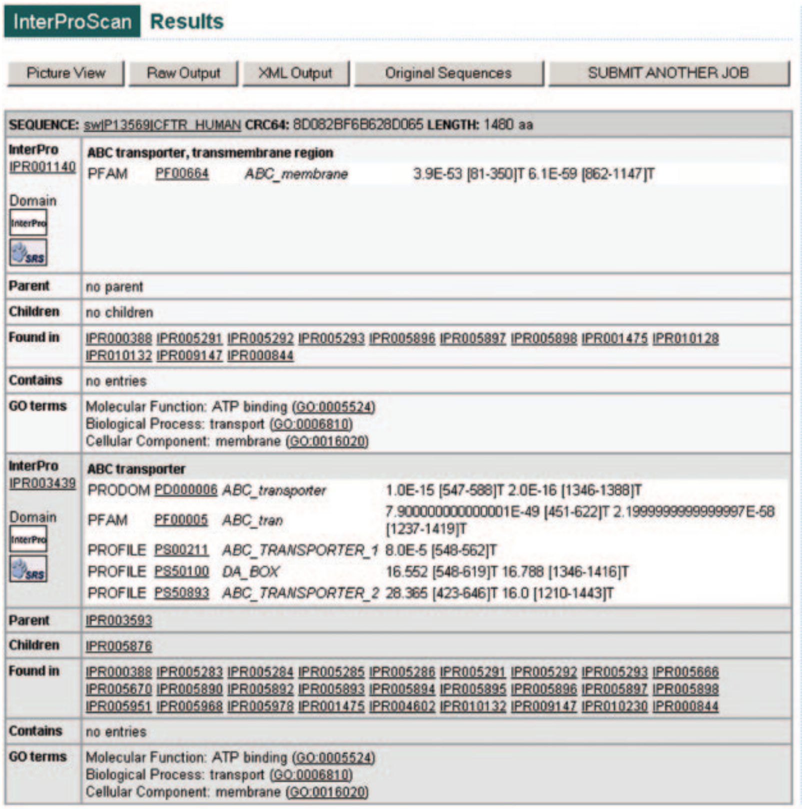
(v) CpGPlot, CpGReport and Isochore can be used to plot CpG-rich areas, report on CpG-rich regions and plot GC content over a sequence. CpG detection is very useful in



SEQUENCE: sw|P13569|CFTR_HUMAN **CRC64:** 8D082BF6B628D065 **LENGTH:** 1480 aa

InterPro IPR001140 Domain	**ABC transporter, transmembrane region** PF00664 ABC_membrane
InterPro IPR003439 Domain	**ABC transporter** PD000006 ABC_transporter PF00005 ABC_tran PS00211 ABC_TRANSPORTER_1 PS50100 DA_BOX PS50893 ABC_TRANSPORTER_2
InterPro IPR005291 Family	**Cyclic AMP-dependent chloride channel** TIGR00953 3a01202 TIGR01271 CFTR_protein
InterPro IPR009147 Family	**Cystic fibrosis transmembrane conductance regulator** PR01851 CYSFIBREGLTR

Figure 1. Graphical view of an InterProScan run on the human cystic fibrosis transmembrane regulator protein. Both this view and the table view (Figure 2) provide direct links to the InterPro entry in both the InterPro database and the SRS indexed version of it, from where the user can link to other classes of database. In the graphical view, matches to the InterPro databases and the relevant positions of these matches on the query sequence are displayed in cartoon-like form.



SEQUENCE: sw|P13569|CFTR_HUMAN **CRC64:** 8D082BF6B628D065 **LENGTH:** 1480 aa

InterPro IPR001140 Domain	**ABC transporter, transmembrane region** PFAM PF00664 *ABC_membrane* 3.9E-53 [81-350]T 6.1E-59 [862-1147]T
Parent	no parent
Children	no children
Found in	IPR000388 IPR005291 IPR005292 IPR005293 IPR005896 IPR005897 IPR005898 IPR001475 IPR010128 IPR010132 IPR009147 IPR000844
Contains	no entries
GO terms	Molecular Function: ATP binding (GO:0005524) Biological Process: transport (GO:0006810) Cellular Component: membrane (GO:0016020)
InterPro IPR003439 Domain	**ABC transporter** PRODOM PD000006 *ABC_transporter* 1.0E-15 [547-588]T 2.0E-16 [1346-1388]T PFAM PF00005 *ABC_tran* 7.900000000000001E-49 [451-622]T 2.1999999999999997E-58 [1237-1419]T PROFILE PS00211 *ABC_TRANSPORTER_1* 8.0E-5 [548-562]T PROFILE PS50100 *DA_BOX* 16.552 [548-619]T 16.788 [1346-1416]T PROFILE PS50893 *ABC_TRANSPORTER_2* 28.365 [423-646]T 16.0 [1210-1443]T
Parent	IPR003593
Children	IPR005876
Found in	IPR000388 IPR005283 IPR005284 IPR005285 IPR005286 IPR005291 IPR005292 IPR005293 IPR005666 IPR005670 IPR005890 IPR005892 IPR005893 IPR005894 IPR005895 IPR005896 IPR005897 IPR005898 IPR005951 IPR005968 IPR005978 IPR001475 IPR004602 IPR010132 IPR009147 IPR010230 IPR000844
Contains	no entries
GO terms	Molecular Function: ATP binding (GO:0005524) Biological Process: transport (GO:0006810) Cellular Component: membrane (GO:0016020)

Figure 2. A portion of the table view for the InterProScan run on the human cystic fibrosis transmembrane regulator protein shown in Figure 1. The table view contains information about the family relationships of the matches with links to parent and children entries, if they exist. If a match has been annotated by the Gene Ontology consortium, the assignments attributed to the object are also displayed here.

providing a method for the identification of housekeeping and widely expressed genes and their promoters, in particular in human sequences (22).

(vi) GeneWise compares a protein sequence or a protein profile to a DNA sequence allowing for introns and frameshifting errors.

(vii) PromoterWise compares two DNA sequences allowing for inversions and translocations and is thus ideal for promoter identification.

(viii) GeneMark (23) is used for the identification of possible genes. It assesses the coding potential of DNA sequences by using Markov models of coding and non-coding regions predetermined for many organisms.

Protein structure analysis tools

A number of public tools for the analysis of structural data are available from the EBI. The function of a protein is more directly a consequence of its structure than of its sequence, with structural homologues tending to share functions. The determination of a protein's 2D/3D structure is crucial in the study of its function.

(i) MSD Services (24) are a wide range of services for conducting structural studies that are provided by the Macromoleular Structure Database group at the EBI. MSDlite is a database interface that allows easy access to the PDB. MSDpro is a java-based tool that allows the user to construct complex relational queries of the PDB. MSDfold compares protein chains and structures and looks for similar ones in PDB and SCOP (25). PQS is a system for querying protein quaternary structures.

(ii) The DALI server (26) is a network service for comparing protein structures in three dimensions. The coordinates of a query protein structure are submitted and DALI compares them against those in the PDB. A multiple alignment of structural neighbours is mailed back to the user.

(iii) MaxSprout is a fast database algorithm for generating protein backbone and side chain coordinates from C (alpha) traces.

Miscellaneous tools

There are a number of useful tools developed at EMBL and the EBI that do not fall into the preceding categories.

(i) EMBL computational services is a collection of tools developed at the EMBL. The tools are divided into the same five categories as the EBI toolbox. These services have been developed by the biocomputing/research group in Heidelberg in close collaboration with wet-lab scientists.

(ii) Expression profiler (27) is a set of web-based tools for clustering, analysing and visualizing the data generated by high-throughput experiments with microrarray technologies.

(iii) QuickGO is a fast web-based viewer for the Gene Ontology (28) database.

(iv) Readseq can be used to change sequence formats and letter cases, remove gap characters, extract and remove the sequence of selected features and calculate a checksum for the sequence.

Database browsing and entry retrieval

The EBI provides a number of different interfaces, of varying complexity, that allow querying and retrieval of data from the large range of biological databases that are available.

(i) SRS stands for Sequence Retrieval System and it is the main information system used to provide access to data from the EBI. SRS has two functions, data retrieval and as an applications server. The system provides access to 175 data libraries and 154 applications, including many from the EMBOSS package. The system allows complex queries across large numbers of fields and/or databases at the same time. Operations that can be performed on the results of a query include linking the results to other databases, saving and viewing the results in different formats and launching applications on the results. Many of the tools from the EBI toolbox are also available from within SRS. Users have the option of uploading their own DNA or protein sequence data to the user-owned databases and in this way can use SRS purely for conducting sequence analysis.

(ii) The EBI maintains a second SRS installation, SRS3D. This system is ideal for the structural biologist since it provides an intuitive and interactive view of sequences, structures and feature data. The 3D viewer allows direct linking from parts of a 3D structure to respective entries in feature databases. Many specialized structural databases including HSSP (29), FSSP (30) and PSSH (31) are available in SRS3D.

(iii) Ensembl (32) is a joint project between the EBI and the Wellcome Trust Sanger Institute that annotates known genes and predicts new ones for large eukaryotic genomes. It is a comprehensive source of stable annotation with confirmed gene predictions that have been integrated from external data sources. There is a BLAST facility provided for searching against the nine currently available genomes.

(iv) Dbfetch can be used to retrieve up to 50 entries at a time from various up-to-date biological databases.

(v) The Web Services project aims to provide programmatic access to the various databases and retrieval and analysis services that the EBI provide. These services are implemented usiong SOAP over HTTP and WSDL.

Data submission

One of the most important tasks of the EBI is the submission processing and curation of biological data. The submission tools available are displayed in Table 3.

The EBI's FTP server

All the data in the EBI databases are publicly available and can be downloaded from the EBI FTP server. Software packages for UNIX, MAC, DOS (MS Windows) and VMS can also be obtained. Access to the server is by anonymous ftp.

Training and support

The EBI provides training and support to the scientific community through the educational resource '2can Bioinformatics' and through its industry programme.

Table 3. EBI submission tools

Data type	Tool	URL	Database(s)
Nucleotide	Webin	http://www.ebi.ac.uk/embl/Submission/webin.html	EMBL, IMGT/LIGM
Multiple alignment	Webin-align	http://www.ebi.ac.uk/embl/Submission/align_top.html	EMBLALIGN
Protein	Spin	http://www.ebi.ac.uk/swissprot/Submissions/spin/index.jsp	UniProt
Structure	AutoDep	http://www.ebi.ac.uk/msd/Deposition/Autodep.html	PDB
Electron microscopy	EMDep	http://www.ebi.ac.uk/msd-srv/emdep/index.html	EMDB
Expression	MiameExpress	http://www.ebi.ac.uk/miamexpress/	ArrayExpress

'2can Bioinformatics' was developed with the intention of targeting a wide audience including newcomers to biology, experienced biologists and computer programmers. The site is divided into five main sections: introduction to bioinformatics, basic biology, genes and disease, biological databases and tutorials. The first three sections provide concise introductions to basic concepts in molecular biology and bioinformatics while the main focus of the last two is on the tools and databases developed at the EBI and by its collaborators. Links to numerous external sites provide further information about the field of bioinformatics, and a searchable glossary of relevant terms is available.

The EBI also provides training to the European pharmaceutical, biotechnology, consumer goods, chemical and agricultural industries through its industry programme. Owing to the success of this programme, a second one has been launched for European small to medium-sized enterprises (SMEs).

Main EBI contacts and addresses

The following list contains a number of useful URLs and email contacts at the EBI. The main contact is http://www.ebi.ac.uk/support/, to which users should address all issues unless a specific alternative is mentioned below. When using the mail servers to submit BLAST, FASTA and InterProScan jobs, the user should send a mail to the respective address with the word 'help' in the message body to obtain instructions.

- http://www.ebi.ac.uk/services/—overview of EBI services
- http://www.ebi.ac.uk/2can/index.html—'2can Bioinformatics' education resource
- http://www.ebi.ac.uk/industry/index.html—EBI industry programme
- datasubs@ebi.ac.uk—data submission queries related to the EMBL database
- ebisrs@ebi.ac.uk—queries related to the EBI SRS server
- industrysupport@ebi.ac.uk—queries related to the industry programme and SME support
- interhelp@ebi.ac.uk—queries related to the InterPro project
- msdhelp@ebi.ac.uk—queries related to the EMSD project and its services
- blast@ebi.ac.uk—BLAST email submissions
- fasta@ebi.ac.uk—FASTA email submissions
- interproscan@ebi.ac.uk—InterProScan email submissions

ACKNOWLEDGEMENTS

We would like to thank our parent organization, the European Molecular Biology Laboratory; the Commission of the European Union, which provides funding through the TEMBLOR grant (QLRI-2001-00015) and the BioBabel grant (QLRT-2000-00981); the US National Institute of Health for the UniProt project funding through grant numbers 1 U01 HG02712-01 and 1R01HGO2273-01; and the Wellcome Trust for their generous funding and support. Further information about EBI funding can be obtained at http://www.ebi.ac.uk/Information/funding/.

REFERENCES

1. Pearson,W.R. (2000) Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol. Biol.*, **132**, 185–219.
2. Lopez,R., Silventoinen,V., Robinson,S., Kibria,A. and Gish,W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res.*, **31**, 3795–3798.
3. Altschul,S.F., Madden,T.L., Schäffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
4. Sturrock,S.S. and Collins,J.F. (1993) *MPsrch V1.3 User Guide*. Biocomputing Research Unit, University of Edinburgh, Edinburgh.
5. Barton,G.J. (1992) Computer speed and sequence comparison. *Science*, **257**, 1609–1610.
6. Kulikova,T., Aldebert,P., Althorpe,N., Baker,W., Bates,K., Browne,P., Van Den Broek,A., Cochrane,G., Duggan,K., Eberhardt,R. *et al.* (2004) The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.*, **32**, D27–D30.
7. Ruiz,M. and Lefranc,M. (2002) IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics*, **53**, 857–883.
8. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. and Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
9. Fredman,D., Siegfried,M., Yuan,Y.P., Bork,P., Lehväslaiho,H. and Brookes,A.J. (2002) HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. *Nucleic Acids Res.*, **30**, 387–391.
10. Hulo,N., Sigrist,C.J.A., Le Saux,V., Langendijk-Genevaux,P.S., Bordoli,L., Gattiker,A., De Castro,E., Bucher,P. and Bairoch,A. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.*, **32**, D134–D137.
11. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Erik,L.L. *et al.* (2004) The Pfam Protein Families Database. *Nucleic Acids Res.*, **32**, D138–D141.
12. Attwood,T.K., Bradley,P., Flower,D.R., Gaulton,A., Maudling,N., Mitchell,A.L., Moulton,G., Nordle,A., Paine,K., Taylor,P. *et al.* (2003) PRINTS and its automatic supplement prePRINTS. *Nucleic Acids Res.*, **31**, 400–402.
13. Mulder,N.J., Apweiler,R., Attwood,T.K., Bairoch,A., Barrell,D., Bateman,A., Binns,D., Biswas,M., Bradley,P., Bork,P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.*, **31**, 315–318.
14. Zdobnov,E.M. and Apweiler,R. (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics*, **17**, 847–848.

15. Etzold,T. and Argos,P. (1993) SRS—an indexing and retrieval tool for flat file data libraries. *Comput. Appl. Biosci.*, **9**, 49–57.
16. Kriventseva,E.V., Fleischmann,W., Zdobnov,E.M. and Apweiler, R. (2001) CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res.*, **29**, 33–36.
17. Hoersch,S., Leroy,C., Brown,N.P., Andrade,M.A. and Sander,C. (2000) The GeneQuiz web server: protein functional analysis through the Web. *Trends Biochem. Sci.*, **25**, 33–35.
18. Heger,A. and Holm,L. (2000) Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41**, 224–237.
19. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.
20. 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.
21. Chenna,R., Sugawara,H., Koike,T., Lopez,R., Gibson,T.J., Higgins,D.G. and Thompson,J.D. (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res.*, **31**, 3497–3500.
22. Larsen,F., Gundersen,G., Lopez,R. and Prydz,H. (1992) CpG islands as gene markers in the human genome. *Genomics*, **13**, 1095–1107.
23. Besemer,J., Lomsadze,A. and Borodovsky,M. (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res.*, **29**, 2607–2618.
24. Golovin,A., Oldfield,T.J., Tate,J.G., Velankar,S., Barton,G.J., Boutselakis,H., Dimitropoulos,D., Fillon,J., Hussain,A., Ionides,J.M.C. *et al*. (2004) E-MSD: an integrated data resource for Bioinformatics. *Nucleic Acids Res.*, **32**, D211–D216.
25. Andreeva,A., Howorth,D., Brenner,S.E., Hubbard,T.J.P., Chothia,C. and Murzin,A.G. (2004) SCOP database in 2004: refinements integrate structure and sequence family data. *Nucleic Acids Res.*, **32**, D226–D229.
26. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.
27. Vilo,J., Kapushesky,M., Kemmeren,P., Sarkans,U. and Brazma,A. (2003) Expression Profiler. In Parmigiani,G., Garrett,E.S., Irizarry,R. and Zeger,S.L. (eds), *The Analysis of Gene Expression Data: Methods and Software*. Springer Verlag, New York, NY.
28. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al*. (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
29. Sander,C. and Schneider,R. (1991) Database of homology derived protein structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
30. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–602.
31. Schafferhans,A., Meyer,J.E.W. and O'Donoghue,S.I. (2003) The PSSH database of alignments between protein sequences and tertiary structures. *Nucleic Acids Res.*, **31**, 494–498.
32. 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.