The European Bioinformatics Institute web site: a new view

Rodrigo Lopez*, Stephen Robinson, Asif Kibria, Nicola Harte, Gulam Patel, Rob Harper, Emmanuel Quevillon, Ville Silventoinen, Kimmo Kallio† and Petteri Jokinen†

EMBL Outstation the European Bioinformatics Institute, Hinxton Hall, Cambridge CB10 1SD, UK

Received on June 26, 2002; accepted on October 25, 2002

ABSTRACT

Summary: The European Bioinformatics Institute (EBI), and outstation of the European Molecular Biology laboratory, has revamped its web site for the second time since 1997 in order to address increased user demand as well as establishing better uniformity and easier accessibility for the ever growing number of users and services it offers to the community. A GRID-like hardware infrastructure has been put in place to provide round the clock services in a redundant and reliable fashion.

Availability: http://www.ebi.ac.uk/

Contact: rls@ebi.ac.uk

MOTIVATION

The main focus of bioinformatics today is the analysis of data produced from the so-called ‘genome era’. Technological advances in the fields of DNA sequencing, proteomics and structure determination have lead to an avalanche of genome, proteome, microarray and functional genomic data. These data are at the centre of all activities of the EBI and its availability demands the identification of standardized methods to access it in order to provide a comprehensive and uniform view to the user. The EBI widely recognizes this as it’s main goal and has thus embarked on a major effort to address these issues through a new web site.

One important goal of this new site is to encourage avoiding ‘parochialism’ in bioinformatics and use as much as possible the web real-estate, not only to promote EBI’s products and services but also those of its collaborators. The framework on which this new web site is built is based on the notion of three main channels of information.

THE CHANNELS

The ‘About the EBI’ channel is designed to provide information regarding the organizational structure, funding, staff, publications, job opportunities, Ph.D. studies programme and inform about events such as conferences and workshops.

The ‘Research’ channel is an introduction to current activities taking place at the EBI organized by group. This part of the web site captures the research spirit of the Institute and provides insight into advanced computational methodologies as well as into the data modelling and design aspects that drive these efforts.

The ‘Services’ channel introduces the user to what can be done and what to expect when using applications in the EBI’s toolbox. It also provides a network of links to the various databases and to information on how to obtain and use these. There are more than 100 services comprising more than 140 databases accessible in a free and unambiguous manner. Site maps are available throughout the site providing basic as well as in-detail links to EBI services. The main one can be found at: http://www.ebi.ac.uk/Information/sitemap.html.

THE TOOLBOX

The toolbox is the main access point to web based services. The EBI is committed to providing not only access to the data it produces but also to analytical tools that exploit it. The toolbox is available from http://www.ebi.ac.uk/Tools/ where the user will find many applications classified into the following categories:

Homology and similarity services

There are five main homology and similarity search algorithms available at the EBI: Fasta (Pearson and Lipman, 1988), WU-BLAST (Altschul et al., 1990), NCBI-BLAST2 (Altschul et al., 1997), MPSRCH (Sturrock and Collins, University of Edinburgh) and Scanps (Barton, 1993). These provide access to complete libraries such as the EMBL nucleotide sequence database, SWISS-PROT protein knowledge base, the non-redundant SWALL (SPTrEMBL) and PDB sequences as well as searches of completed genomes and proteomes.

Protein functional analysis

Protein function determination services are dominated by InterPro (Apweiler et al., 2001). Five methodologies are available to search the InterPro databases. InterProScan

*To whom correspondence should be addressed.
†EBI’s Systems group.
has rapidly become the de-facto standard for fast and reliable protein function characterization. All individual search methods are available as stand alone applications where parameter tweaking is allowed in order to fulfill specific requirements.

Structural analysis
For protein structure analysis and secondary structure prediction there are highly accurate algorithms such as DALI (Holm and Sander, 1993) and MaxSprout, SSM (Secondary Structure Matching, in press) and many more. There is a group at the EBI specially dedicated to structural biology—Macromolecular Structure Database (MSD) group that offers highly sophisticated services to the biostructures community. MSD can be found at: http://www.ebi.ac.uk/msd

Sequence analysis
This section contains links to sequence analysis services. By far the most popular service is the alignment of multiple nucleotide and protein sequences using the ubiquitous ClustalW package (Thompson et al., 1994). Various alignment viewers are available as back-ends to Clustalw as well as for Fasta and BLAST searches.

DATABASES SECTION
This part of the web site introduces the user to the databases maintained at the EBI and to what is their ultimate purpose and status. It is organized into groups according to molecule types and comprises: Nucleotide sequence databases such as the EMBL nucleotide sequence database (Stoesser et al., 2002) and activities that build around it; a section on protein sequence databases such as SWISS-PROT, TrEMBL and IPI; Protein function databases such as InterPro (Apweiler et al., 2001); and a section on 3D and secondary structure libraries such as EMSD, HSSP, FSSP, and DSSP. Finally, microarrays are represented by ArrayExpress, which today plays a central role in the archiving and analysis of array expression data and its distribution into the public domain.

DOWNLOADS SECTION
This section provides a map into the organization of EBI’s ftp server and links directly to the main databases as well as to the bioinformatics software archive. The main access point to this resource is: ftp://ftp.ebi.ac.uk/

SUBMISSIONS SECTION
A very important function of the EBI is to provide efficient and accurate means for users to submit their data into public domain databases. The Submissions section is a one-stop station to the submission tools and documents, which are a ‘must read’ for all scientists planning to make their data free and publicly available. These tools include WEBIN for entering data into the EMBL nucleotide sequence database, SWISS-PROT documents and text forms for protein sequences, AutoDep for 3D structural coordinates and MIAMExpress for microarray data.

INFRASTRUCTURE
The infrastructure on which EBI’s web services currently run is of in-house design. The hardware consists of one IP redirector, two static pages servers, 11 Linux mini farms and more than 256 CPU’s from various vendors, which act as the main computational backbone. This setup extensively uses vendor-specific cluster technology, SAN storage, NFS-caching and job scheduling and load balancing systems that effectively represent a GRID environment. The main aim in designing this system is to guarantee redundant, 24:7 service availability.

REFERENCES


