Biological Data Resources at the EMBL-EBI

Herramientas para el manejo de datos biológicos en el EMBL-EBI

Rodrigo Lopez[1]

Recibido: septiembre 10 de 2008 Aprobado: octubre 23 de 2008

Introduction

The European Bioinformatics Institute (EBI) is an Outstation of the European Molecular Biology Laboratory (EMBL). These are Europe's flagships in bioinformatics and basic research in molecular biology. The EBI has been maintaining core data resources in molecular biology for 15 years and is notionally custodian to the largest collection of databases and services in Life Sciences in Europe. EBI provides access in a free and unrestricted fashion to these resources to the international research community. The data resources at the EBI are divided into thematic categories. Each represents a special knowledge domain where one or several databases are maintained. The aims of this note are to introduce the reader to these resources and briefly outline training and education activities which may be of interest to students as well as academic staff in general. The web portal for the EBI can be found at http://www.ebi.ac.uk/ and represents a single entry point for all data resources and activities described below.

The Resources

Nucleotide and Genomes

Nucleotide and genomes resources are at the core of most of the work being carried out by the bioinformatics research community at present. The resources in this category contain vast amounts of data (on the order of 10TB at present), which are submitted to the EBI's central nucleotide repository EMBL-Bank [1] and the Trace archive. This database is part of INSDC (International Nucleotide Sequence Database Collaboration (see: www.insdc.org), which is composed of the National Centre of Biotechnology Information (NCBI) in the United States, the DNA Database of Japan

1 EMBL-EBI. Wellcome Trust Genome Campus, Hinxton, CB10 1SD, Cambridge, UK. rls@ebi.ac.uk

(DDBJ) and EBI. EMBL-Bank is the oldest nucleotide repository in the world and synchronises content with the other partners of the collaboration on a daily basis.

Genome specific resources include the ENSEMBL project [2], which is run between the Wellcome Trust Sanger Institute and the EBI. ENSEMBL currently provides access to more than 30 metazoan genomes. These include primates, such as human, macaque and orangutan; rodents, such as mouse, rat and rabbits; laurasiatherians, such as horses, dogs and cats; afrotheria, such as elephants; and selected species of marsupials, birds, reptiles and fish.

GenomeReviews [3] is a parallel project with ENSEMBL and provides access to the genomes of archaea, bacteria, bacteriophages and selected eukaryota. In order to bridge the information gaps and improve the understanding of genomes, there is a project known as Integr8. This one provides easy access to integrated information about deciphered genomes and their corresponding proteomes. Available data includes DNA sequences (from databases including the EMBL-Bank, Genome Reviews, and ENSEMBL); protein sequences (from databases including the UniProt Knowledgebase and IPI); statistical genome and proteome analysis (performed using InterPro, CluSTr, and GOA); and information about orthology, paralogy, and synteny.

Other resources, which are maintained in collaboration with the EBI, include the IMGT, the International IMmunoGeneTics Database [4]. This is a high-quality integrated knowledge resource specialized in the immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC), immunoglobulin superfamily (IgSF), major histocompatibility complex superfamily (MhcSF) and related proteins of the immune system (RPI) of human and other vertebrate species.

The Human Genome Organisation (HUGO) Gene Name Nomenclature Committee (HGNC) [5] is hosted at the EBI since 2007. It aims to approve and assign unique gene symbol and name to every gene in the human genome.

Splicing events, that give rise to much of the variation and diversity within single species are catalogues and maintained in the Alternative Splicing and Transcript Diversity database project, or ASTD [6], which creates a database of alternative splice events and transcripts of genes from human, mouse and rat.

Proteins

The Protein category contains a large set of databases and international collaborations that encompass the domain of functional and structural knowledge available to date. The most important resource is the UniProt Knowledgebase [7], which is an international collaboration between the Swiss Institute of Bioinformatics (SIB), The Protein Identification Resource (PIR) and the EBI. UniProt is the central hub for the collection of functional information on proteins, with accurate, consistent, and rich annotation, derived from direct submissions and translations from the INSDC databases. UniProtKB is composed of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL, mainly. There exist databases which have been specially designed to search the knowledgebase and that comprise clusters of proteins at 100%, 90% and 50% similarity in the UniRef [8] datasets. In addition to these, there are special protein sequence repositories that specialise on novel types of data derived from state-of-

the-art experiments. The most important of these is the UniMes database that was specifically developed for metagenomic and environmental data. There are also data collection services that provide information on the history and their relationship to other proteins. UniParc [9] is the UniProt Archive, which contains a non-redundant collection of protein sequences extracted from public databases and contains protein sequences a wide range of sources, including patents. UniSave , the UniProtKB Sequence/Annotation Version Archive [10], is a repository of UniProtKB/Swiss-Prot and UniProtKB/TrEMBL entry versions dating back to 1984. This is an extremely important resource at it describes chronologically each change that has been made to each protein sequence currently available.

As in the case of genes and their naming (see HGNC above), EBI is engaged in maintaining a portal that provides information on the standardisation of the naming of proteins, which is known as the IntEnz [11] data resource. This one specialises on enzymes and contains the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) on the nomenclature and classification of enzyme-catalysed reactions. IntEnz represents a centralised repository for the naming, classification and assigned function of enzymes contained in the protein sequence databases.

Structures

The EBI has developed and maintains a set of molecular structure databases throughout the years. The most important of these is the international protein structures databases known as the PDB (Protein Database) [12]. The deposition, curation and structure validation effort in Europe is coordinated by the EBI, in collaboration with the RCSB Protein Database in the United States. In addition the the 3D structure information held within the PDB, there are secondary structure and fold classification resources and specialised services that provide access to small molecules such as ligands and monomers, residue modifications, RESID [13], and catalytic sites and residues identified in enzymes using structural data as in the Catalytic Site Atlas, CSA [14].

The domain of small molecular entities is represented by the CHEBI, Chemical Entities of Biological Interest database [16]. This resource focuses on small chemical compounds involved in a wide variety of reactions. The definition of an entity in this database comprises constitutionally or isotopically distinct atom, molecule, ion, ion pair, radical, radical ion, complex, conformer, etc., identifiable as a separately distinguishable entity.

Gene Expression

Microarrays represent the most important breakthrough in experimental Life Sciences that allow for the measurement in time of gene expression events and their level. Public resources in gene expression are concentrated mainly on two databases at the moment. The NCBI maintains GEO, Gene Expression Omnibus [15], and the EBI runs ArrayExpress [17], which is a public repository of transcriptomics data and gene-indexed expression profiles. Public data in ArrayExpress can be accessed using a wide range of criteria, such as gene, sample, experimental properties, species, submitter, etc.

Molecular Interactions and Pathways

At the EBI, molecular interactions data resources are comprised by IntAct [18]. This describes how molecules interact with each other forming strong (covalent) and weak (ionic and hydrogen) bonds. In this database, these are represented as binary-interactions. Inter-molecular contacts give rise to reaction cascades commonly known as pathways. In biology and bioinformatics these are referred to as metabolic pathways. These are often represented as graphical maps that described how particular enzymes catalyze reactions and the products of these yield molecules that regulate homeostasis within an organism. The number of data point in these graphs is growing by the day at present and it is important that data repositories for all data-types derived from metabolomic studies get established and linked to resources such as protein sequences, structures and nucleotides mentioned earlier. There are two distinct types of resource at present taking care of metabolomic knowledge: BioModels [19] is a database of annotated biological models that allows biologists to store, search and retrieve published mathematical models of biological interests. Reactome [20], on the other hand, is a curated database of biological processes in humans. Reactome will not only be useful to general biologists as an online textbook of biology, but also to bioinformaticians for making new discoveries about biological pathways.

Protein Families

A protein family is a classification or grouping of related proteins that share a common sequence, structure or function. There are many classification systems at present that classify proteins according to specific criteria. Some resources use domains identified in sequences using regular expressions or profiles, while others use 3D-structure relationships, inferred by advanced sequence or 3D structural alignments. The InterPro Consortium [21] at the EBI, is in charge of coordinating and maintaining a central resource for these classifications and currently counts with input from the resource outlined in Figure 1. InterPro provides an integrated view of the commonly used signature databases, and provides tools such as InterProScan [22], to analyse and predict protein architectures, domains and functions.

Literature and Text

Almost all scientific record is stored in books and journals. Access to this knowledge is crucial for the maintenance of high quality standards in the annotations contained in the EBI databases. Currently, the EBI is engaged with the National Library of Medicine in the US, The British Library and various universities in Europe in the implementation of systems that provide free and unrestricted access to citations and publications in the biological domains. At present the EBI offers the means to access more than 21 million distinct citations from resources that include Agricola, Chinese Biological Abstracts, Citeseer, Patents and PubMed/Medline. This is all done within a service called CiteXplore (see: http://www.ebi.ac.uk/citexplore/), which combines literature search with text mining tools. Some of these tools are invaluable for determining biological by using novel techniques that allow for the disambiguation of semantic types, e.g. proteins, genes, species, drugs, etc.

Figure 1: InterPro Consortium member databases.

New Resources

Very large amounts of genomic information are currently being generated by the latest sequencing instruments. The include 454 and Solexa sequences capable of producing the same amount of sequence stored in EMBL-Bank in one week. This is also equivalent to two complete human genomes every 24 hours. International collaborations, such as the projects known as '1000 genomes' (see: www.1000genomes.org) are developing resources to bring the analysis of these data to the biomedical research community at large. It aims to create the most detailed and medically useful picture to date of human genetic variation. Drawing on the expertise of multidisciplinary research teams, the 1000 Genomes Project will develop a new map of the human genome that will provide a view of biomedically relevant DNA variations at a resolution unmatched by current resources. As with other major human genome reference projects, data from the 1000 Genomes Project will be made swiftly available to the worldwide scientific community through freely accessible public databases.

In parallel with the 1000 Genomes projects, EBI is engaged in EGA, the European Genotype Archive. This database project is designed to be a repository for all types of genotype experiments, including case control, population, and family studies. It will include SNP and CNV genotypes from microarray based methods and genotyping done with re-sequencing methods.

EBI is custodian to large amounts of scientific information. In order to make these data more accessible and provide easier navigation between the resources, EBI has invested heavily in the development of systems that provide browser-based and programmatic access to all its resources. This resource is known as the 'EB-eye' and is available from the top of all web pages on the EBI portal (see Figure 2). The EB-eye is a novel approach to discovering and exploring the relationships that exist between the different databases describe above. The system is very fast, always up-to-date and provides and uncompromised view of all the data types available at present.

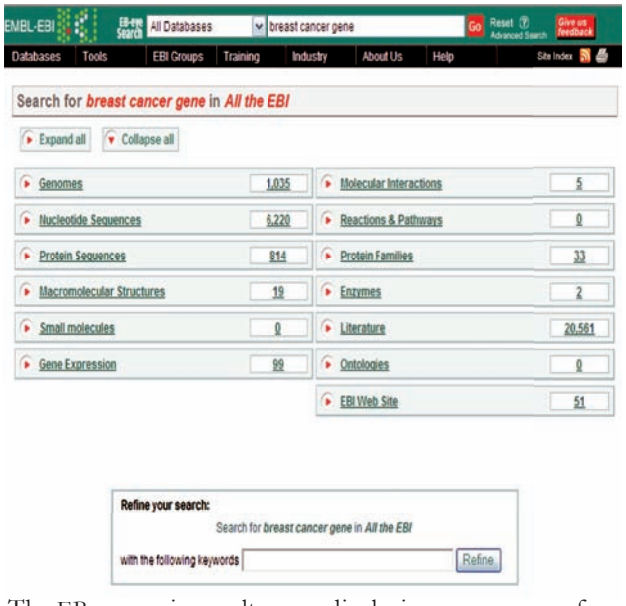

Figure 2: The EB-eye main results page displaying a summary of results across all EBI databases.

Users can carry out complex queries using Web Services (see: http://www.ebi.ac.uk/Tools/webservices) and allows for a high-degree of customisation with local data resources. In the near future, it will be possible to launch applications directly from the system. This will allow for easier and more comprehensive analysis of data, without the burden of requiring special know-how.

Training and Education

The EMBL-EBI training programme is organised under the umbrella of EICAT (see: http://www.embl.org/training/eicat.html), the EMBL **I**nternational **C**entre for **A**dvanced **T**raining. EICAT coordinates training activities for scientists at different levels. Under this programme there are opportunities for Master and PhD students from around the world. The programme comprises hand-on training at basic, intermediate and advanced levels.

Part of the outreach and training effort EBI is currently engaged in, involves coming closer with the scientist users in order to understand and better serve their needs. In this activity the EBI organises road-shows where selected members from the EBI staff will travel and run specialised workshops, specially tailored to local needs. Roadshows are typically held over two days and can cover four main 'maxi' modules (each lasting half a day and presented in more detail here) or a combination of maxi and 'mini' modules. Sessions can be run in parallel to cater to the differing interests and requirements of the audience. Please contact the roadshow coordinator Janet Copeland, email: copeland@ebi.ac.uk, Tel: +44 (0)1223 492510, for further details.

References

1. Cochrane G., Akhtar R., Aldebert P., Althorpe N., Baldwin A., Bates K., Bhattacharyya S., Bonfield J., Bower L., Browne P., Castro M., Cox A., Demiralp F., Eberhardt R., Faruque N., Hoad G., Jang M., Kulikova T., Labarga A., Leinonen R., Leonard S., Lin Q., Lopez R., Lorenc D., McWilliam H., Mukherjee G., Nardone F., Plaister S., Robinson S., Sobhany S., Vaughan R., Wu D., Zhu W., Apweiler R., Hubbard T., Birney E. (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Research Database Issue Jan 2008 [epub ahead of print]*

2. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Gräf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kähäri A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S. Ensembl 2008. *(Jan-2008) Nucleic acids research, 36 (Database issue) :D707-14*

3. Sterk P., Kersey P.J., Apweiler R. (2006) Genome Reviews: Standardizing Content and Representation of Information about Complete Genomes. *OMICS 10(2): 114-118.*

4. Lefranc MP, Giudicelli V, Regnier L, Duroux P. IMGT, a system and an ontology that bridge biological and computational spheres in bioinformatics. *(Jul-2008) Briefings in bioinformatics, 9 (4):263-75*

5. Bruford EA, Lush MJ, Wright MW, Sneddon TP, Povey S, Birney E. The HGNC Database in 2008: a resource for the human genome. *(Jan-2008) Nucleic acids research, 36 (Database issue) :D445-8*

6. Stamm S., Riethoven J.J., Le Texier V., Gopalakrishnan C., Kumanduri V., Tang Y., Barbosa-Morais N.L., Thanaraj T.A. (2006) ASD: a bioinformatics resource on alternative splicing. *Nucleic Acids Research 34: D46-D55 (2006).*

7. The UniProt consortium. The Universal Protein Resource (UniProt) Nucleic Acids Res. 36:D190-D195(2008).

8. Suzek B.E., Huang H., McGarvey P., Mazumder R., Wu C.H. UniRef: Comprehensive and Non-Redundant UniProt Reference Clusters *Bioinformatics 23:1282-1288(2007).*

9. Leinonen R., Diez F.G., Binns D., Fleischmann W., Lopez R., Apweiler R. UniProt Archive *Bioinformatics 20:3236-3237(2004).*

10. Leinonen R., Nardone F., Zhu W., Apweiler R. UniSave: the UniProtKB Sequence/Annotation Version database *Bioinformatics 22:1284-1285(2006).*

11. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R. IntEnz, the integrated relational enzyme database. *(1-Jan-2004) Nucleic acids research, 32 (Database issue) :D434-7*

12. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *(Jan-2007) Nucleic acids research, 35 (Database issue):D301-3*

13. Garavelli JS. The RESID Database of Protein Modifications as a resource and annotation tool. *(Jun-2004) Proteomics, 4 (6) :1527-33*

14. Using a Library of Structural Templates to Recognise Catalytic Sites and Explore their Evolution in Homologous Families. James W. Torrance, Gail J. Bartlett, Craig T. Porter, Janet M. Thornton (2005) J Mol Biol. 347:565-81

15. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R. NCBI GEO**:** mining tens of millions of expression profiles--database and tools update. *(Jan-2007) Nucleic acids research, 35 (Database issue) :D760-5*

16. Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008) ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res. 36*, D344–D350.

17. Ivliev AE, PA, Villerius MP, den Dunnen JT, Brandt BW. Microarray retriever: a web-based tool for searching and large scale retrieval of public microarray data. *(1-Jul-2008) Nucleic acids research, 36 (Web Server issue) :W327-31*

18. Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R, Kohler C, Khadake J, Leroy C, Liban A, Lieftink C, Montecchi-Palazzi L, Orchard S, Risse J, Robbe K, Roechert B, Thorneycroft D, Zhang Y, Apweiler R, Hermjakob H. IntAct--open source resource for molecular interaction data. *(Jan-2007) Nucleic acids research, 35 (Database issue) :D561-5*

19. Le Novère N, Bornstein B, Broicher A, Courtot M, Donizelli M, Dharuri H, Li L, Sauro H, Schilstra M, Shapiro B, Snoep JL, Hucka M. BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *(1-Jan-2006) Nucleic acids research, 34 (Database issue):D689-91*

20. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Reactome: a knowledge base of biologic pathways and processes. *(2007) Genome biology, 8 (3) :R39*

21. Nicola J. Mulder, Rolf Apweiler, Teresa K. Attwood, Amos Bairoch, Alex Bateman, David Binns, Peer Bork, Virginie Buillard, Lorenzo Cerutti, Richard Copley, Emmanuel Courcelle, Ujjwal Das, Louise Daugherty, Mark Dibley, Robert Finn, Wolfgang Fleischmann, Julian Gough, Daniel Haft, Nicolas Hulo, Sarah Hunter, Daniel Kahn, Alexander Kanapin, Anish Kejariwal, Alberto Labarga, Petra S. Langendijk-Genevaux, David Lonsdale, Rodrigo Lopez, Ivica Letunic, Martin Madera, John Maslen, Craig McAnulla, Jennifer McDowall, Jaina Mistry, Alex Mitchell,, Anastasia N. Nikolskaya, Sandra Orchard, Christine Orengo, Robert Petryszak, Jeremy D. Selengut, Christian J. A. Sigrist, Paul D. Thomas, Franck Valentin, Derek Wilson, Cathy H. Wu and Corin Yeats (2007) New developments in the InterPro database. *Nucleic Acids Res. 35 (Database Issue):D224-228*

22. Quevillon E., Silventoinen V., Pillai S., Harte N., Mulder N., Apweiler R., Lopez R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Research 33: W116-120*