D46–D53 *Nucleic Acids Research, 2005, Vol. 33, Database issue*
doi:10.1093/nar/gki026

The European Bioinformatics Institute's data resources: towards systems biology

Catherine Brooksbank*, Graham Cameron and Janet Thornton

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

Received September 15, 2004; Revised and Accepted September 21, 2004

ABSTRACT

Genomic and post-genomic biological research has provided fine-grain insights into the molecular processes of life, but also threatens to drown biomedical researchers in data. Moreover, as new high-throughput technologies are developed, the types of data that are gathered *en masse* are diversifying. The need to collect, store and curate all this information in ways that allow its efficient retrieval and exploitation is greater than ever. The European Bioinformatics Institute's (EBI's) databases and tools have evolved to meet the changing needs of molecular biologists: since we last wrote about our services in the 2003 issue of *Nucleic Acids Research*, we have launched new databases covering protein–protein interactions (IntAct), pathways (Reactome) and small molecules (ChEBI). Our existing core databases have continued to evolve to meet the changing needs of biomedical researchers, and we have developed new data-access tools that help biologists to move intuitively through the different data types, thereby helping them to put the parts together to understand biology at the systems level. The EBI's data resources are all available on our website at http://www.ebi.ac.uk.

INTRODUCTION

Since the European Molecular Biology Laboratory (EMBL) launched the world's first nucleotide sequence database in 1982, we have witnessed a revolution in high-throughput molecular biology. As well as easy access to individual nucleotide and protein sequences, biomedical researchers now need to search and analyse whole genomes, transcriptomes, proteomes, interactomes and metabolomes. EMBL's European Bioinformatics Institute (EBI) serves as the European node for globally coordinated efforts to collect and disseminate molecular biological data. Through collaboration with its partners around the world it maintains the world's most comprehensive range of molecular databases, covering almost every molecular domain necessary to life, from small molecules and ions through to the genomic sequences of higher organisms (Table 1). But although our resources are diverse, they all uphold six basic principles:

Accessibility: All our resources are freely available to the research community, without restriction.

Compatibility: The EBI is committed to and proactive in the development of standards in bioinformatics and promoting their adoption; the development of these standards greatly facilitates data sharing.

Comprehensive data sets: Where several publicly available repositories exist, we have negotiated data-sharing agreements to ensure that our resources contain comprehensive and up-to-date datasets. We also negotiate with publishers to ensure that, wherever practicable, biological data are placed in a public repository as part of the publication process and cross-referenced in the relevant publication.

Portability: All our datasets are available for download from the EBI website. In many cases the entire software system can be downloaded and installed locally.

Quality: Our databases are enhanced through annotation: features of the genes or proteins stored in them are extracted from other sources, defined and interpreted. Much of our annotation is performed by highly qualified biologists, and our automated annotation is subjected to rigorous quality control.

Navigability: Wherever possible our datasets are cross-linked, allowing researchers to move seamlessly among different molecular domains.

Altogether, the EBI hosts around 160 databases and a comprehensive range of tools for data access, submission and analysis. All of these can be accessed from our services site map at www.ebi.ac.uk/services. Here, we provide a taste of some of our most widely used and recently launched resources. We hope that this article will provide a starting point from which to explore individual resources, many of which are covered in more detail elsewhere in this issue.

GENES TO GENOMES

EMBL-Bank (www.ebi.ac.uk/embl) (1) is Europe's primary resource for DNA and RNA sequence information. It is

*To whom correspondence should be addressed. Tel: +44 0 1223 492525; Fax: +44 0 1223 494468; Email: cath@ebi.ac.uk

Table 1. Access, submissions and user support for the EBI's main data resources and integration tools

Topic	Database	Access	Submissions	Downloads	User support
Data resources					
General	All	www.ebi.ac.uk/services/	www.ebi.ac.uk/Submissions/	www.ebi.ac.uk/FTP/	www.ebi.ac.uk/help/ www.ebi.ac.uk/support/
DNA and RNA sequences	EMBL-Bank	www.ebi.ac.uk/embl/	www.ebi.ac.uk/embl/Submission/	ftp.ebi.ac.uk/pub/databases/embl/	www.ebi.ac.uk/embl/Documentation/ datalib@ebi.ac.uk
Vertebrate genomes	Ensembl	www.ensembl.org	Ensembl does not accept direct submissions; genome sequences should be submitted to EMBL-Bank (see above)	www.ensembl.org/Download/	www.ensembl.org/Docs/ www.ensembl.org/helpdesk
Prokaryotic genomes	Genome Reviews	www.ebi.ac.uk/GenomeReviews/	Genome Reviews does not accept direct submissions; genome sequences should be submitted to EMBL-Bank (see above)	ftp.ebi.ac.uk/pub/databases/genome_reviews/	www.ebi.ac.uk/GenomeReviews/userman.html genome_reviews@ebi.ac.uk
Transcriptomes	ArrayExpress	www.ebi.ac.uk/arrayexpress/	www.ebi.ac.uk/arrayexpress/Submissions	ftp.ebi.ac.uk/pub/databases/arrayexpress/	www.ebi.ac.uk/arrayexpress/Help arrayexpress@ebi.ac.uk arraysubs@ebi.ac.uk (submission enquiries)
Protein sequences	UniProt	www.uniprot.org	Directly sequenced protein sequences can be submitted to UniProt/SwissProt at www.ebi.ac.uk/swissprot/Submissions UniProt does not provide accession numbers, in advance, for protein sequences that are the result of translation of nucleic acid sequences.	www.ebi.uniprot.org/database/download.shtml	www.ebi.uniprot.org/support/ help@uniprot.orgdatalib@ebi.ac.uk (submission enquiries)
Protein structures	MSD	www.ebi.ac.uk/msd/	www.ebi.ac.uk/msd/Deposition.html	www.ebi.ac.uk/msd/services/FTPdownload.html	www.ebi.ac.uk/msd/Documentation.html msd@ebi.ac.uk
Protein–protein interactions	IntAct	www.ebi.ac.uk/intact/	www.ebi.ac.uk/intact/index.html#submission	www.ebi.ac.uk/intact/index.html#soft (for software) ftp.ebi.ac.uk/pub/databases/intact/ (for data)	www.ebi.ac.uk/intact/doc/html/documentation.html Intact-help@ebi.ac.uk datasubs@ebi.ac.uk (submission enquiries)
Pathways	Reactome	www.reactome.org	Contact the Reactome Development team at dev@reactome.org	www.reactome.org/download	www.reactome.org/userguide/userguide.html

Table 1. *Continued*

Topic	Database	Access	Submissions	Downloads	User support
Small molecules	ChEBI	www.ebi.ac.uk/chebi/	ChEBI does not accept direct submissions	ftp.ebi.ac.uk/pub/databases/chebi/	www.ebi.ac.uk/chebi/pages/html/help.html www.ebi.ac.uk/chebi/emailChebi.do?toolBarForward=true
Controlled vocabularies describing functional attributes of gene products	Gene Ontology	www.geneontology.org	GO does not accept direct submissions but you can submit suggestions for new terms by clicking on the Curator Requests Tracker link at www.geneontology.org	www.geneontology.org/index.shtml#downloads	www.geneontology.org/GO.contents.doc.html go@geneontology.org
Annotations of gene products with GO terms	Gene Ontology Annotation Database	www.ebi.ac.uk/GOA/	GOA does not accept direct submissions, but users are encouraged to send updates and corrections to goa@ebi.ac.uk	ftp.ebi.ac.uk/pub/databases/GO/goa/	www.ebi.ac.uk/GOA/goaHelp.html goa@ebi.ac.uk
Integrated access to data resources					
Genomes and proteomes	Integr8	www.ebi.ac.uk/integr8	Integr8 does not accept direct submissions	First use QuickSearch to find the relevant species, then click on the downloads link	www.ebi.ac.uk/integr8/ContentsHelpPage.do or click on local help in the left-hand menu bar for context-dependent help. www.ebi.ac.uk/support/
Complex queries across several databases	BioMart	www.ebi.ac.uk/biomart	Not applicable	www.ebi.ac.uk/biomart/install.html	www.ebi.ac.uk/biomart/install.html www.ebi.ac.uk/biomart/contact.html
Retrieve data from several databases	SRS	srs.ebi.ac.uk	Not applicable	Not applicable	srs.ebi.ac.uk/doc www.ebi.ac.uk/support
	Web services	www.ebi.ac.uk/Tools/webservices	Not applicable	www.ebi.ac.uk/Tools/webservices/download.html	www.ebi.ac.uk/Tools/webservices/FAQ.html sharmila@ebi.ac.uk

produced in an international collaboration with two other nucleotide sequence databases, GenBank (USA) (2) and the DNA Databank of Japan (DDBJ) (3). Each of the three groups collects a proportion of the total sequence data reported worldwide, and all new and updated database entries are exchanged between the groups on a daily basis, so users can access sequences submitted to any of the three collaborating databases through EMBL-Bank. They can also submit new nucleotide sequences to EMBL-Bank using a web-based submission tool and be confident that they will appear in all three databases.

EMBL-Bank is now 22 years old but it continues to evolve to meet the needs of its users. In 2002 it began to accept third-party annotations (TPA)—submissions that update or improve entries previously submitted by others. TPA must be supported by publication in a peer-reviewed journal, and contain a cross-reference to the original submission. As of September 2004, the TPA subsection of EMBL-Bank contained over 4400 entries.

Computational gene prediction methods have proved to be an invaluable way of annotating newly sequenced genomes, but ultimately the only way of confirming that a gene exists and checking that its start site has been predicted correctly is to sequence the 5′ end of every transcript. There are several high-throughput projects in progress to do this, and one of EMBL-Bank's current challenges is to develop a practicable way of storing these large numbers of short sequences, typically 18–22 bp long. Another emerging challenge is handling the data from metagenomics projects, in which DNA from entire ecosystems, including soil, sea-water and the gastrointestinal tract, is sequenced (4).

EMBL-Bank provides access to completed and partially completed genome sequence data through its Genomes Pages (www.ebi.ac.uk/genomes/). It is now nearly 10 years since the first prokaryotic genome, that of *Haemophilus influenzae* (5), was deposited in EMBL-Bank. Since then, the body of knowledge about protein function has increased significantly, and methods for inferring function from sequence have also improved. Researchers need to be able to superimpose current knowledge onto the sequence, but editorial control of the data in EMBL-Bank belongs to their authors, who cannot always update the original entries themselves. The EBI has addressed this issue by developing two related resources:

Genome Reviews (www.ebi.ac.uk/GenomeReviews/) (6) is a comprehensive and standardized resource for completely sequenced prokaryotic genomes. It takes completed genome sequences from EMBL-Bank and adds detailed and standardized annotation, including additional cross-references to other databases, providing information on coding information, domains, protein processing and function, for example. In total, the number of cross-references has increased from roughly 650 000 in the corresponding portion of EMBL-Bank to 2.5 million in Genome Reviews (as of September 2004). Genome Reviews is updated every two weeks. The data in Genome Reviews have been incorporated into Integr8 (www.ebi.ac.uk/integr8) (6), a new portal for completely sequenced genomes and their predicted proteomes that utilizes cross-links between coding sequences, protein sequences and protein structures. Integr8 offers comprehensive statistical analyses of genomes and proteomes, and allows biologists to fully exploit the wealth of information available.

Genome Reviews currently covers prokaryotes with completely sequenced genomes, and coverage will be expanded to include some lower eukaryotes in the future. Ensembl (www.ensembl.org) (7), produced jointly by the EBI and the Wellcome Trust Sanger Institute, provides access to animal genomes, with a focus on vertebrates. As of September 2004, 13 animal genomes are available in Ensembl; recent additions include those of the dog, chimpanzee and chicken. For each organism, Ensembl is based on assemblies generated by the relevant genome centre or consortium. For example, for the human sequence, Ensembl collaborates with the US National Centre for Biotechnology Information (NCBI) and University of California Santa Cruz (UCSC) to assess the human assemblies, and all three genome-browsing sites (Ensembl, UCSC and NCBI) use the same data.

Ensembl uses computational evidence (typically database comparisons) to annotate the genomic sequence and predict the positions of genes, and it documents the supporting evidence for these predictions. Its highly customizable displays allow users to zoom around from the level of whole chromosomes down to the sequence, find single nucleotide polymorphisms and other features, and compare two genomes. Complex queries can be performed using Ensmart (8), Ensembl's data-mining tool, and users can add their own annotations using DAS (Distributed Annotation System) (9). Another new feature is pre-Ensembl, a pre-build site that provides displays of genomes in the process of being annotated. Genomes are made available through pre-Ensembl when the Ensembl team has done the initial BLAST analysis on a new assembly but has not completed the gene build. It provides early access to new assemblies of a genome, with limited views of the data.

TRANSCRIPTOMES

Microarrays allow snapshots to be made of gene expression levels on a genomic scale, and are revolutionizing all areas of the molecular life sciences, from basic biology to drug discovery. ArrayExpress (www.ebi.ac.uk/arrayexpress) (10) is the EBI's database for collecting information about microarray experiments. It is the world's first database to store microarray information in a way that conforms to agreed community standards, the MIAME (Minimum Information About a Microarray Experiment) standards devised by the Microarray Gene Expression Data (MGED) Society (www.mged.org) (11). Communities that have not traditionally used transcriptome data, such as drug approval agencies and nutritional researchers, are now adopting these standards, and ArrayExpress facilitates the sharing of vast amounts of microarray-based data that previously would have been difficult to access or analyse.

Since its launch in December 2002, almost 10 000 hybridizations have been submitted to ArrayExpress. Initially most of its submissions came from a relatively small number of high-throughput transcriptomics laboratories through their own pipelines. It is testimony to the widespread adoption of MIAME standards that now over half of its submissions come from smaller-scale experiments submitted using MIAMExpress, often as part of the publication of a paper. Many journals require or recommend authors of microarray-data-based papers to submit their data to a MIAME-compliant database, and the

MGED Society has recently published an open letter (12) urging publishers to embrace the MIAME standards consistently, by requiring authors of microarray-based papers to submit their data to one of the three recognized MIAME-compliant data resources: ArrayExpress (10), the Gene Expression Omnibus (GEO) (13) or CIBEX (14). One particularly useful feature of ArrayExpress in this regard is a secure system that provides password-protected access to journal editors and reviewers. Once a paper is published, the data can then be moved to the public part of the database. The ArrayExpress team also produces Expression Profiler (15), a modular suite of microarray data-analysis tools that can be chained together to perform complex routines.

PROTEINS TO PROTEOMES

Protein sequence databases have become a crucial resource for analysing the proteomes of newly sequenced organisms, making intelligent predictions about the functions of newly identified proteins, and moving towards understanding how proteins interact to create pathways, networks and systems. Until recently there were two major efforts to make information about proteins publicly available. One was a collaboration between the EBI and the Swiss Institute of Bioinformatics that resulted in two complementary databases, Swiss-Prot (16), renowned for providing a great depth of information on proteins through high-quality manual curation, and TrEMBL (16), a much larger database in which information on protein function is derived computationally by comparison with other proteins. The other was the PIR International Protein Sequence Database (PIR-PSD) (17), the world's first database of classified and functionally annotated proteins. These databases held different, but overlapping, subsets of proteins. Swiss-Prot, TrEMBL and PIR-PSD have now been combined to create UniProt (www.uniprot.org) (18), a new universal protein resource that is the world's most comprehensive catalogue of information on proteins. UniProt is produced and maintained by the UniProt Consortium, a collaboration among the EBI, the Swiss Institute of Bioinformatics (SIB) and the Protein Information Resource (PIR). UniProt provides a 'one-stop shop', allowing easy access to all the publicly available information on proteins. It is split into three parts, each optimized for a different use:

The UniProt Archive, UniParc, is the most comprehensive publicly accessible non-redundant protein sequence database available, providing links to all underlying sources and versions of these sequences.

The UniProt knowledgebase provides access to a wealth of functional information on proteins. Every UniProt knowledgebase entry contains the amino acid sequence, protein name or description, taxonomic data and citation information. In addition to this, UniProt's curators, in consultation with external experts, add as much annotation as possible: information on protein function, post-translational modifications, functional domains and active sites, subunit structure, subcellular location, diseases associated with mutations in the protein's gene, and sequence features including conflicts and variants are added, along with clear indications of the quality of annotation in the form of evidence attribution to experimental and computational data.

The UniRef databases provide clustered sets of sequences from UniProt to provide complete coverage of sequence space at several resolutions. Each UniRef database is built using the UniProt Knowledgebase entries as a set of masters. All the records that are 100, 90 or 50% identical are then merged to create UniRef100, UniRef90 and UniRef50. This yields a database size reduction of ~40% for UniRef90 and 65% for UniRef50, providing for significantly faster sequence searches. All the sequences in each cluster are ranked to facilitate the selection of a representative sequence.

The EBI provides a range of services to simplify proteome analysis. These can be accessed through Integr8 (6). Using UniProt's proteome sets [or the International Protein Index (19) for species whose proteomes are still accumulating annotations at a rapid rate], Integr8 computes the following information for each proteome: (i) ranked lists of the most common families, domains and repeats [using InterPro (20), which combines information from several source databases to produce an integrated documentation resource for protein families, domains and functional sites, thus providing deeper coverage than any of the contributing databases]; (ii) comparisons with other proteomes; (iii) hierarchical clusters (21) of proteins based on sequence similarity [useful for identifying orthologues, paralogues and singletons—proteins whose sequence is unrelated to any others in the organism]; (iv) functional classifications using the Gene Ontology (GO) (www.geneontology.org) (22) and (v) links to structural information including secondary (HSSP) (23) and tertiary (PDB) (24) structures. Full proteome sets for all those organisms whose genomes have been fully sequenced can also be downloaded. Other features of Integr8 include the ability to select your own 'basket' of organisms to search, perform customized comparative analyses and extract information from several databases in a single query using BioMart (www.ebi.ac.uk/biomart), a new EBI tool for complex queries (see later).

It is also possible to add your own annotations to the UniProt Knowledgebase, UniParc and IPI datasets using Protein DAS, a feature annotation package for local installation. The UniProt DAS server can be accessed via www.ebi.ac.uk/uniprot-das/.

STRUCTURES

Solving the three-dimensional (3D) structures of proteins can reveal a great deal of functional and mechanistic information, as well as facilitating drug design. The Macromolecular Structure Database (MSD) (25) is the European project for the collection, management and distribution of data about macromolecular structures. MSD is a partner in the Worldwide Protein Databank (wwPDB) (26): it collaborates closely with the Research Collaboratory for Structural Bioinformatics (RCSB) (24) and the Protein Databank Japan (PDBj) (www.pdbj.org) to make structures freely available to the research community. MSD is a relational database that presents the PDB's data in a consistent way. It has a range of powerful tools that allow you to visualize and compare structures, search for ligands and find proteins that are the targets of structural genomics efforts. It includes structures determined by X-ray crystallography, NMR and 3D electron microscopy; 3D electron microscopy particularly important for the study of

larger proteins and their complexes, and will help to bridge the gap between molecules and cellular architecture.

MSD's data-access tools are tailored to three different levels of user: MSDbar caters for those who are new to structural biology. It is a toolbar application that takes a few seconds to install and can be used to search the most widely used structural databases (MSD, RCSB, PDBj and OCA) (http://bip.weizmann.ac.il/oca-bin/ocamain) directly from the user's browser, using a general text search or searching on specific fields such as author name, keyword or bound ligand. MSDlite caters for more experienced structural biologists; users can tailor their searches using a wide range of different search filters and customize the results page. MSDPro is a powerful search tool for the expert user; it has a drag-and-drop interface that allows users to build, save and combine complex queries.

PROTEIN–PROTEIN INTERACTIONS

Cataloguing the molecular ingredients of life—genes, transcripts, proteins and structures—is the first step towards understanding biological systems. The next one is collecting all the known protein–protein interactions. The EBI has been working closely with the Human Proteomics Organization (HUPO) (www.hupo.org) to develop standards for describing protein–protein interactions. The PSI-MI format (Proteomics Standards Initiative Molecular Interaction format) (27) is a community standard data model for representing and exchanging protein-interaction data. It has been jointly developed by members of the Proteomics Standards Initiative (PSI), a work group of HUPO, and is supported by major protein-interaction data providers. One result of this is the IntAct database (www.ebi.ac.uk/intact) (28)—a central, public repository for storing and accessing protein–protein interaction information. IntAct is being populated both with experimental data and with curated literature data. As of September 2004 it contained 45 000 binary interactions involving around 30 000 proteins. IntAct enables users to find all the proteins that interact with a protein of interest, probe more deeply into individual experiments to gain both a degree of confidence in the specific interaction and its functional consequence, graphically display interaction networks, analyse interaction networks using GO terms, visualize minimal connecting networks for protein sets, download data in PSI-MI format and install the complete IntAct system locally. Once installed, you can add your own data and choose whether or not to make them publicly available through the IntAct website.

IntAct is a member of the International Molecular-Interaction Exchange (IMEx) consortium, a collaborative group of providers of molecular interaction data [other members are BIND (29), DIP (30), MINT (31) and MIPS (32)] that will begin to exchange data on a regular basis as of January 1. The IMEx Consortium is also campaigning for published molecular interaction data to be placed in the public domain in a standardized format.

PATHWAYS

Step three along the route from parts to systems is to document how interactions combine to create pathways. Reactome (www.reactome.org) (33), previously known as the Genome Knowledgebase, is a collaboration among Cold Spring Harbor Laboratory (CSHL), The European Bioinformatics Institute and The Gene Ontology Consortium to develop a curated resource of core pathways and reactions in human biology.

Reactome covers biological pathways ranging from the basic processes of metabolism to complex regulatory pathways such as hormonal signalling. While Reactome is targeted at human pathways, it also includes many individual biochemical reactions from non-human systems such as rat, mouse, pufferfish and zebrafish.

Bench biologists who are experts on a particular pathway provide the basic information in Reactome. It is then checked and managed by curators at CSHL and EBI, peer-reviewed by other researchers and published on the Web. Each pathway is represented as a series of events (which may be broken down into sub-events) with defined inputs and outputs. Each event has a list of components, complete with cross-links to UniProt (18), Ensembl (7) and LocusLink (34); multimeric components have a second-level page that lists all the proteins in a complex with their cross-links to other databases. All the information in Reactome is backed up by its provenance: either a literature citation or an electronic inference based on sequence similarity.

Reactome has a unique user interface in which pathways are represented as a series of constellations in a 'starry sky', which can be used to navigate through the universe of human reactions and visualize connections between pathways. Reactome entries are cross-referenced to PubMed (34), UniProt (18), LocusLink (34), Ensembl (7) and GO (22).

SMALL MOLECULAR ENTITIES

Molecular biologists tend to focus on genes and proteins, but small molecules are equally important to life. The EBI's most recently launched database bridges the gap between the world of proteins and that of small molecules. Called ChEBI (Chemical Entities of Biological Interest, www.ebi.ac.uk/chebi), it catalogues small molecules, atoms, ions, ion pairs, radicals and other small chemical entities.

ChEBI combines information on small molecular entities from three main sources to create a non-redundant resource: small molecules from the EBI's IntEnz database of enzymes (35), the COMPOUND database from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (36) and the Chemical Ontology (http://cvs.sourceforge.net/viewcvs.py/obo/obo/ontology/biochemical/). The Chemical Ontology makes ChEBI uniquely powerful because it allows relationships between molecular entities or classes of entities to be recorded in a defined way. Each entity in the database is described in terms of its chemistry and, where known, its broad biological function. For example, FAD is described as a flavin adenine dinucleotide (chemistry) and as a cofactor (function). Synonyms for each entity are listed and searchable. ChEBI also defines the relationships between macromolecules and small molecular entities: there are cross-links to every protein in the UniProt protein knowledgebase that is documented to interact with each entity.

INTEGRATED DATA RETRIEVAL

The shifting focus from parts to systems affects not only the types of data that biologists require, but also how they need to

access the data; it is no longer sufficient to be able to search individual databases for specific data types. One of the greatest challenges in bioinformatics is integrating these data types, often from databases with very different structures, in a meaningful and user-friendly way.

As well as asking gene-centric or species-centric questions such as 'give me all the DNA-binding proteins in *Escherichia coli*', biologists also need to ask function-oriented questions, such as 'find me all the mouse proteins involved in cytokine signalling'. This requires that objects in the databases are described consistently. We do this via controlled vocabularies, the most widely used of which is GO (22). The GO Consortium has created a defined vocabulary to describe the functions, processes and cellular components that are associated with gene products. Annotating other databases with GO terms facilitates uniform queries across them. For example, the GO Annotation (GOA) (37) project at the EBI assigns GO terms to proteins in UniProt (18). GOA is the largest contributor of annotations to the GO Consortium effort: the current release has almost 5 million annotations to over a million proteins. Terms are assigned by a combination of electronic mappings and curator judgements.

Documenting the relationships between objects in the different databases (e.g. this gene encodes this protein, which has this structure) is also crucial for adding biological meaning to the data and allowing users to navigate logically around the databases. We have created over 20 million cross-references in our databases. One result of this work is Integr8 (6), discussed above, which provides a single point of access for information about genomes, the proteins they encode, and the structures and functions of those proteins. Another area of considerable recent progress is the addition of around 3.3 million full-text cross-links to the scientific literature; users can now access full-text articles directly via the EBI's implementation of MEDLINE and from literature citations in some of its databases.

But perhaps the ultimate challenge to data integration is extracting meaningful data from all the databases in a single query. One new development at the EBI that aims to address this need is BioMart. Put simply, by ignoring the content of the data and focusing instead on their structure, BioMart allows users to query different databases simultaneously, and build up complex queries by chaining simpler queries together. To date, BioMart can be used to query genomes (Ensembl) (7), manual annotation of genomes (VEGA) (38), proteomes (UniProt) (18), macromolecular structures (MSD) (25) and single nucleotide polymorphisms (dbSNP) (34). BioMart can be accessed over the Web using MartView (www.ebi.ac.uk/biomart/martview), or by installing one of two standalone interfaces for local use: MartShell uses the command line whereas MartExplorer is a graphical interface that resembles the web-based interface. Users can build their own BioMart databases as well as download the 'ready-made' databases detailed above. Other ways of facilitating data retrieval from several datasets include the EBI's SRS server (srs.ebi.ac.uk), web services (www.ebi.ac.uk/Tools/webservices) and the distribution of several of our databases in XML format.

CONCLUSIONS

As new technological developments in biological research have allowed the collection of larger and more diverse data types, the EBI has endeavoured to stay at the forefront of this research, providing new and increasingly sophisticated resources to meet the needs of its users. We cannot and do not develop our services in isolation; our service teams are an integral part of the communities that they serve. If you need help with any of our services, we provide help documentation at www.ebi.ac.uk/help and our help desk can be contacted from www.ebi.ac.uk/support. We welcome queries and feedback from all our users; with your help we will be able to continue to serve the needs of biological researchers, and fulfil our aim of accelerating scientific progress through the provision of high-quality, freely accessible services.

ACKNOWLEDGEMENTS

We thank the following persons for their help in putting this article together: Rolf Apweiler, Ewan Birney, Alvis Brazma, Evelyn Camon, Kirill Degtyarenko, Kim Henrick, Henning Hermjakob, Midori Harris, Maria Jesus-Martin, Arek Kasprzyk, Paul Kersey, Tamara Kulikova, Rodrigo Lopez, Claire O'Donovan, Peter Stoehr and Jawahar Swaminathan. Projects at the EBI are supported by the European Molecular Biology Laboratory (EMBL), the European Commission, the Wellcome Trust, the US National Institutes of Health, the UK Biotechnology and Biosciences Research Council, UK Medical Research Council, the UK Engineering and Physical Sciences Research Council, the UK Department of Trade and Industry and the EBI Industry Programme.

REFERENCES

1. Kanz,C., Aldebert,P., Althorpe,N., Baker,W., Baldwin,A., Bates,K., Browne,P., van den Broek,A., Castro,M., Cochrane,G. *et al.* (2005) The EMBL nucleotide sequence database. *Nucleic Acids Res.*, **33**, D29–D33.
2. Benson,D.A., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Wheeler,D.J. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
3. Tateno,Y., Saitou1,N., Okubo,K., Sugawara,H. and Gojobori,T. (2005) DDBJ in collaboration with mass-sequencing teams on annotation. *Nucleic Acids Res.* **33**, D25–D28.
4. Toussaint,A., Ghigo,J-M. and Salmond,G.P.C. (2003) A new evaluation of our life-support system: bacterial benefactors—and other prokaryotic pursuits *EMBO Rep.*, **4**, 820–824.
5. Fleischmann,R.D., Adams,M.D., White,O., Clayton,R.A., Kirkness,E.F., Kerlavage,A.R., Bult,C.J., Tomb,J.F., Dougherty,B.A., Merrick,J.M. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, 496–512.
6. Kersey,P., Bower,L., Morris,L., Horne,A., Petryszak,R., Kanz,C., Kanapin,A., Das,U., Michoud,K., Phan,I. *et al.* (2005) Integr8 and Genome Reviews: integrated views of complete genomes and proteomes. *Nucleic Acids Res.* **33**, D297–D302.
7. Birney,E., Andrews,D., Bevan,P., Caccamo,M., Cameron,G., Chen,Y., Clarke,L., Coates,G., Cox,T., Cuff,J. *et al.* (2004) Ensembl 2004. *Nucleic Acids Res.*, **32**, D468–D470.
8. Kasprzyk,A., Keefe,D., Smedley,D., London,D., Spooner,W., Melsopp,C., Hammond,M., Rocca-Serra,P., Cox,T. and Birney,E. (2004) EnsMart: a generic system for fast and flexible access to biological data. *Genome Res.*, **14**, 160–169.
9. Dowell,R.D., Jokerst,R.M., Da,A., Eddy,S.R. and Stein,L. (2001) The Distributed Annotation System. *BMC Bioinformatics*, **2**, 7.
10. Parkinson,H., Sarkans,U., Shojatalab,M., Abeygunawardena,N., Contrino,S., Coulson,R., Farne,A., Garcia Lara,G., Holloway,E., Kapushesky,M. *et al.* (2005) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **33**, D553–D555.

11. Brazma,A., Hingamp,P., Quackenbush,J., Sherlock,G., Spellman,P., Stoeckert,C., Aach,J., Ansorge,W., Ball,C.A., Causton,H.C. *et al*. (2001) Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature Genet*., **29**, 365–371.
12. Ball,C.A., Brazma,A., Causton,H., Chervitz,S., Edgar,R., Hingamp,P., Matese,J.C., Parkinson,H., Quackenbush,J., Ringwald,M. *et al*. (2004) Submission of microarray data to public repositories. *PLoS Biol*., **31**, E317.
13. Barrett,T., Suzek,T.O., Troup,D.B., Wilhite,S.E., Ngau,W.-C., Ledoux,P., Rudnev,D., Lash,A.E., Fujibuchi,W. and Edgar,R. (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res*., **33**, D562–D566.
14. Ikeo,K., Ishi-i,J., Tamura,T., Gojobori,T. and Tateno,Y. (2003) CIBEX: center for information biology gene expression database. *C. R. Biol*., **326**, 1079–1082.
15. Kapushesky,M., Kemmeren,P., Culhane,A.C., Durinck,S., Ihmels,J., Korner,C., Kull,M., Torrente,A., Sarkans,U., Vilo,J. and Brazma,A. (2004) Expression Profiler: next generation—an online platform for analysis of microarray data. *Nucleic Acids Res*., **32**, W465–W470.
16. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al*. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*., **31**, 365–370.
17. Wu,C.H., Yeh,L.S., Huang,H., Arminski,L., Castro-Alvear,J., Chen,Y., Hu,Z., Kourtesis,P., Ledley,R.S., Suzek,B.E. *et al*. (2003) The Protein Information Resource. *Nucleic Acids Res*., **31**, 345–347.
18. Bairoch,A., Apweiler,R., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R., Magrane,M. *et al*. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
19. Kersey,P., Duarte,J., Williams,A., Karavidopulou,Y., Birney,E. and Apweiler,R. (2004) The International Protein Index—an integrated database for proteomics experiments. *Proteomics*, **4**, 1985–1988.
20. 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.
21. Kriventseva,E.V., Servant,F. and Apweiler,R. (2003) Improvements to CluSTr: the database of SWISS-PROT+TrEMBL protein clusters. *Nucleic Acids Res*., **31**, 388–389.
22. 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.
23. Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure–sequence alignments. *Nucleic Acids Res*., **25**, 226–230.
24. Deshpande,N., Addess,K.J., Bluhm,W.F., Merino-Ott,J.C., Townsend-Merino,W., Zhang,Q., Knezevich,C., Chen,L., Feng,Z., Green,R.K. *et al*. (2005) The RCSB Protein Data Bank: a redesigned query system and relational database based on the mmCIF schema. *Nucleic Acids Res*., **33**, D233–D237.
25. Golovin,A., Oldfield,T.J., Tate,J.G., Velankar,S., Barton,G.J., Boutselakis,H., Dimitropoulos,D., Fillon,J., Hussain,A., Ionides,J.M. *et al*. (2004) E-MSD: an integrated data resource for bioinformatics. *Nucleic Acids Res*., **32**, D211–D216.
26. Berman,H., Henrick,K. and Nakamura,H. (2003) Announcing the worldwide Protein Data Bank. *Nature Struct. Biol*., **10**, 980.
27. Hermjakob,H., Montecchi-Palazzi,L., Bader,G., Wojcik,J., Salwinski,L., Ceol,A., Moore,S., Orchard,S., Sarkans,U., von Mering,C. *et al*. (2004) The HUPO PSI's molecular interaction format—a community standard for the representation of protein interaction data. *Nat. Biotechnol*., **22**, 177–183.
28. Hermjakob,H., Montecchi-Palazzi,L., Lewington,C., Mudali,S., Kerrien,S., Orchard,S., Vingron,M., Roechert,B., Roepstorff,P., Valencia,A. *et al*. (2004) IntAct: an open source molecular interaction database. *Nucleic Acids Res*., **32**, D452–D455.
29. Bader,G.D., Betel,D. and Hogue,C.W.V. (2003) BIND, the Biomolecular Interaction Network Database. *Nucleic Acids Res*., **31**, 248–250
30. Salwinski,L., Miller,C.S., Smith,A.J., Pettit,F.K., Bowie,J.U. and Eisenberg,D. (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res*., **32**, D449–D451.
31. Zanzoni,A., Montecchi-Palazzi,L., Quondam,M., Ausiello,G., Helmer-Citterich,M. and Cesareni,G. (2002) MINT: a Molecular INTeraction database. *FEBS Lett*., **513**, 135–140.
32. Mewes,H.W., Amid,C., Arnold,R., Frishman,D., Guldener,U., Mannhaupt,G., Munsterkotter,M., Pagel,P., Strack,N., Stumpflen,V. *et al*. (2004) MIPS: analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*., **32**, D41–D44.
33. Joshi-Tope,G., Gillespie,M., Vastrik,I., D'Eustachio,P., Schmidt,E., de Bono,B., Jassal,B., Gopinath,G., Wu,G., Matthews,L. *et al*. (2005) Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res*., **33**, D428–D432.
34. Wheeler,D.L., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S., Helmberg,W. *et al*. (2005) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*., **33**, D39–D45.
35. Fleischmann,A., Darsow,M., Degtyarenko,K., Fleischmann,W., Boyce,S., Axelsen,K.B., Bairoch,A., Schomburg,D., Tipton,K.F. and Apweiler,R. (2004) IntEnz, the integrated relational enzyme database. *Nucleic Acids Res*., **32**, D434–D437.
36. Kanehisa,M., Goto,S., Kawashima,S., Okuno,Y. and Hattori,M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*., **32**, D277–D280.
37. Camon,E., Magrane,M., Barrell,D., Lee,V., Dimmer,E., Maslen,J., Binns,D., Harte,N., Lopez,R. and Apweiler,R. (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res*., **32**, D262–D266.
38. Ashurst,J.L., Chen,C.-K., Gilbert,J.G.R., Jekosch,K., Keenan,S., Meidl,P., Searle,S.M., Stalker,J., Storey,R., Trevanion,S. *et al.* (2005) The Vertebrate Genome Annotation (Vega) database. *Nucleic Acids Res*., **33**, D459–D465.