Biological Data Centres

Introductory article

Phillip Stafford, *Arizona State University, Tempe, Arizona, USA*

HoJoon Lee, *Arizona State University, Tempe, Arizona, USA*

Based in part on the previous version of this Encyclopedia of Life Sciences (ELS) article, Biological Data Centres by Phillip Stafford.

Article Contents

Online posting date: 15th January 2010

The age of genomics has pushed the cost of sequencing much lower than typical cost/ technology curves would predict. Sanger sequencing, next-gen sequencing (ABI SOLiD sequencing, Illumina GA, Roche 454) and even whole-chromosome imaging are providing sequence data faster than most laboratories can analyse or store. The biological data infrastructure that was established in the early 1990s is still in place, mostly because it was very well planned in terms of future needs. The three main biological data centres are NCBI (http://www.ncbi.nlm.nih.gov/), DDJB (http://www.ddbj.nig.ac.jp/) and EBI/EMBL (http://www.ebi.ac.uk/embl/). These centres will be discussed in the context of new types of high-density biological data, such as microarrays of various sorts. This article will discuss history, the tools that are provided to the public, other biological databases that support and integrate with sequencing databases and a projection of biology in the future.

History

EMBL/EBI

In 1980, scientists at the European Molecular Biology Laboratory (EMBL) recognized a need for a centralized computer database of deoxyribonucleic acid (DNA) sequences. Originally this database was used to collect, annotate and archive published sequences. However, the volume of data from direct electronic submission of sequences soon eclipsed the volume anticipated by the founders. The task of annotating and saving DNA sequence data quickly grew in scale as commercial sequencing projects began and the data became commercially relevant. European Bioinformatics Institute (EBI) took the lead in database structure for assembling genome sequences, but other centres soon followed. Today the Sanger Centre is leading the way in implementing next-gen sequencing projects.

To address the expansion of the original DNA database, the EMBL Council voted in 1992 to establish the EBI and to locate it at the Wellcome Trust Genome Campus in the United Kingdom, where it would be in proximity to the major sequencing efforts at Sanger Centre. From 1992 to 1995, a gradual transition occurred: the database moved from Heidelberg, where the EMBL is currently located, to the EBI on the Wellcome Trust Genome Campus. In addition to the Sanger Centre and EBI, the Wellcome Trust campus also houses the UK Medical Research Council Human Genome Mapping Project (HGMP) Resource Centre, and the United Kingdom: Life Science Organizations. Together, these institutes provide one of the world's largest concentrations of expertise in genomics and bioinformatics. The mission of the EBI 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 facets of the scientific community in ways that promote scientific progress. One of the fastest growing disciplines that emerged from this initial influx of sequence data was comparative genomics (see **Table 1**). The insight into conservation of DNA sequence has led to the discovery of sequences in organism's genomes that are poison (i.e. some sequences cannot be tolerated) to the identification of highly conserved noncoding regions that may play a role in disease and disregulation of gene expression. The biostatistics for sequence analysis have emerged from a need to analyse EBI's data, and other data centres have joined their efforts. **See also**: Medical Research Council (MRC); Sequencing the Human Genome: Novel Insights into its Structure and Function; The Wellcome Trust

CIB/DDBJ

The DDBJ came online in 1986. It was founded at the National Institute of Genetics (NIG) in Mishima, Japan, which is governed by the Japanese Ministry of Education, Science, Sport and Culture. It was designed from the beginning to be one of the international DNA sequence databases and was well equipped to mirror its sister sites. DDBJ is the sole DNA data bank in Japan, and is officially certified to collect DNA sequences from researchers and to

ELS subject area: Structural Biology

How to cite:
Stafford, Phillip; and Lee, HoJoon (January 2010) Biological Data Centres. In: Encyclopedia of Life Sciences (ELS). John Wiley & Sons, Ltd: Chichester.
DOI: 10.1002/9780470015902.a0003026.pub2

Table 1 Sizes of completed vertebrates genomes

Mammal	*Homo sapiense* (human)	3107 million bases	Approximately 24 119 genes	KEGG
Mammal	*Equus caballus* (horse)	2484 million bases	Approximately 17 619	KEGG
Mammal	Ornithorhynchidae (Platypus)	1996 million bases	Approximately 16 387 genes	KEGG
Mammal	*Felis catus* (cat)	4045 million bases	Approximately 20 000	
Mammal	*Rattus norvegicus* (rat)	2834 million bases	Approximately 26 142 genes	KEGG
Mammal	*Macaca mulatta* (Rhesus macaque)	2864 million bases	Approximately 23 965	KEGG
Mammal	*Pan troglodytes* (chimpanzee)	3350 million bases	Approximately 25 184 genes	KEGG
Mammal	*Callithrix jacchus* (marmoset)	3030 million bases	No estimation yet	
Mammal	*Pongo pygmaeus abelii* (orangutan)	3446 million bases	Approximately 12 728	Ensembl
Mammal	*Mus musculus* (mouse)	2725 million bases	Approximately 29 452 genes	KEGG
Mammal	*Cavia porcellus* (guinea pig)	2723 million bases	Approximately 14 143 genes	Ensembl
Mammal	*Canis lupus familiaris* (dog)	2531 million bases	Approximately 19 807 gene	KEGG
Mammal	*Bos taurus* (cow)	2917 million bases	Approximately 22 334 genes	KEGG
Mammal	*Monodelphis domestica* (opossum)	3605 million bases	Approximately 19 114	KEGG
Vertebrate	*Petromyzon marinus* (lamprey)	1027 million bases	No estimation yet	
Vertebrate	*Gallus gallus* (chicken)	1100 million bases	Approximately 18 118	KEGG
Vertebrate	*Anolis carolinensis* (lizard)	1781 million bases	Approximately 12 043	Ensembl
Vertebrate	*Gasterosteus aculeatus* (stickleback)	463 million bases	Approximately 14 881	Ensembl
Vertebrate	*Taeniopygia guttata* (zebra finch)	1233 million bases	1706	Ensembl
Vertebrate	*Xenopus tropicalis* (*X. tropicalis*)	1513 million bases	8540	KEGG
Vertebrate	*Danio rerio* (zebrafish)	1440 million bases	27 485	KEGG
Vertebrate	*Tetraodon nigroviridis* (tetraodon)	402 million bases	27 918	KEGG
Vertebrate	*Takifugu rubripes* (fugu)	400 million bases	22 041	KEGG
Vertebrate	*Oryzias latipes* (medaka)	869 million bases	25 084	KEGG

issue the internationally recognized accession number to data submitters. Data are collected mainly from Japanese researchers, although accession numbers are granted to researchers in any other country. As data are exchanged between EMBL/EBI and GenBank/NCBI (National Center for Biotechnology Information) on a daily basis, the three data banks share virtually the same data at any given time.

In 1995, a new centre was established at the NIG known as the Centre for Information Biology (CIB), which allows the DDBJ to expand its activities. The CIB is composed of four distinct laboratories that devote themselves not only to ongoing database activities but also to projects in information biology and molecular evolution. The primary mission of the DDBJ is to provide a geographically convenient location for scientists in the Pacific Rim to submit their sequence data rapidly. Equally important, the DDBJ continues basic research on molecular evolution and developing and improving bioinformatics software, processing existing data and rapid data dissemination. Data are accepted by the DDBJ by either Sequin (developed by GenBank) or by a DDBJ-specific electronic submission known as Sakura. These new submission methods are flexible and rapid, and replace the older Authorin, a now defunct method of submission via floppy disk.

GenBank/NCBI

On 4 November 1988, legislation was passed that established the NCBI as a division of the US National Library of Medicine (NLM) at the National Institutes of Health (NIH). NLM was chosen for its experience in creating and maintaining biomedical databases and, as part of the NIH, it could establish an intramural research programme in computational biology. The collective research components of the NIH currently make up the largest biomedical research facility in the world. **See also**: National Institutes of Health (NIH)

In 1990, the NCBI offered a programme that could be used to find similar DNA sequences and provide users a statistical measure of DNA sequence similarity. BLAST (Basic Local Alignment Search Tool) soon became the tool of choice among molecular biologists and currently supports over 100 000 searches per day. In October 1992, the NCBI assumed formal responsibility for GenBank, the actual DNA sequence database that had previously been distributed on CD-ROM.

The consortia

International sequencing consortium

The three components (DDBJ, EBI and NCBI) of the international DNA database consortium are distinct entities, but all have a common purpose: the accurate and rapid distribution of DNA sequence data. Each database has its own objectives, but all are now part of an electronic community that provides access to new and archived DNA sequence data. The mirroring of sequence data is not only an important safeguard but also allows scientists from various geographical parts of the world to have rapid access to the DNA database closest to them. **See also**: Genome Databases

SNP Consortium and International HapMap Project. Although these data centres provide sequence data to the public via the Internet, they are not always perfectly synchronous. For example, a chromosome position at NCBI may not be identical to a chromosome position at UCSC. Given the enormous amount of variability in the human genome, the differences in alignment tools, the details of error correction and telomere and centromere handling, it is actually quite remarkable that these centres are as close to one another as they are. These data centres provide many tools for browsing, analysing and submitting genomic data. NCBI and UCSC tools are the class leaders, and set the standard by which other software tools are measured.

Other biological data centres and visualization tools

Today, biologists have at their disposal a rich collection of molecular data in the form of DNA and RNA (ribonucleic acid) sequences, polymorphism data (SNPs, single nucleotide polymorphism), methylation data (CpG islands), siRNA (small interfering RNA), ncRNA (non-coding RNA), mtDNA (mitochondrial DNA), and a wealth of other high-density information about genes and gene function. A brief list is included below, see also **Table 2**: **See also**: Human Genetics: Online Resources

RefSeq	Reference sequence	http://www.ncbi.nlm.nih.gov/RefSeq/
DbEST	Expressed sequence tag	http://www.ncbi.nlm.nih.gov/dbEST/
dbSTS	Sequence tagged sites	http://www.ncbi.nlm.nih.gov/dbSTS/
DbSNP	Single nucleotide polymorphism	http://www.ncbi.nlm.nih.gov/SNP/index.html
dbMHC	Major histocompatibility complex	http://www.ncbi.nlm.nih.gov/gv/mhc/main.cgi?cmd = init
dbSKY	Spectral karyotyping	http://www.ncbi.nlm.nih.gov/sky/
dbFISH	Fluorescence *in situ* hybridization	
GEO	Gene Expression Omnibus	http://www.ncbi.nlm.nih.gov/geo/
Haplotype	HapMap project	http://www.ncbi.nlm.nih.gov/geo/
GeneCard	Gene annotations	http://www.genecards.org/
KEGG	Metabolism pathways	http://www.genome.jp/kegg/
BioRag	Expression pathways	http://www.biorag.org/

These databases and online software tools have helped in the assembly and interpretation of massive biological data that have been collected over the past two decades. The tools provided by each site provide the public with a straightforward way of visualizing and retrieving these data. As sequence data grows increasingly dense, it becomes difficult to make useful biological interpretations. This problem appeared during the development of expression microarrays and it took years for the bioinformatics and statistical methods to catch up. Microarrays

Table 2 Websites of interest

Description	Site	URL
Gene	Gene Ontology	http://www.geneontology.org/
	Gene Card	http://www.genecards.org/
	ACE	http://www.ncbi.nlm.nih.gov/IEB/Research/Acembly/index.html
Centralized repositories	EBI/EMBL	http://www.ebi.ac.uk/
	NCBI	http://www.ncbi.nlm.nih.gov/
	DDBJ	http://www.ddbj.nig.ac.jp/
	UCSC genome browser	http://genome.ucsc.edu/
SNP	HapMap project	http://www.hapmap.org/
RNA	ENCODE project	http://www.genome.gov/10005107
	GEO	http://www.ncbi.nlm.nih.gov/geo/
	EST	http://www.ncbi.nlm.nih.gov/dbEST/
Alternative splicing	EC gene	http://genome.ewha.ac.kr/ECgene/
	ASTB	http://www.ebi.ac.uk/asd/
Cancer	Sanger Institute	http://www.sanger.ac.uk/
	CGAP	http://cgap.nci.nih.gov/
Protein	Swiss Prot	http://ca.expasy.org/sprot/
	PDB	http://www.rcsb.org/pdb/home/home.do
	PIR	http://pir.georgetown.edu/
Protein interaction	iHOP	http://www.ihop-net.org/UniPub/iHOP/
Pathway	KEGG pathway	http://www.genome.ad.jp/kegg/pathway.html

contain thousands or even millions of ‘probes’ that detect individual biological molecules simultaneously. GEO, the Gene Expression Omnibus, contains a great deal of high-density array data containing CGH, SNP, gene expression, exon, splicing, methylation and even protein data. It is possible to integrate all of these data using only the chromosome position as a guidepost. Each of the databases listed earlier can provide their data in the context of physical chromosome position. With this as a primary ‘key’, most other databases can be searched for compatible information, and the scientist can assemble a molecular picture of the state of an organism.

Software tools

For years, NCBI has been the main repository for sequencing and microarray data – they have set the standard for biological data storage and dissemination throughout the world. Their database schemas, analysis tools, even their personnel management structures are freely available on request and they even offer technical training for their many software tools. Many local repositories have modelled themselves after NCBI; this trend helps centralize data even faster when these local databases merge with NCBI. Certain segments within NCBI have become even more important as microarray and next-gen sequencing data become more common. GEO is a fairly recent department at NCBI but they have quickly become the *de facto* standard for uploading, storing and retrieving array data. Their simplified data structures, fast FTP servers and the almost universal publication requirement for making array data public has led to explosive growth at GEO. They are charged with retaining and distributing microarray data, but data and image files have grown far faster than the natural growth and cost reduction of online storage (Kryder’s Law). Since this data facility has become so important to non- and for-profit groups alike, funding resources for data storage has taken several new turns and twists. A new cost-sharing paradigm is likely to emerge in the next few years as the inherent value of sequence and microarray data increases. The development of high-throughput analysis technologies, new database software and new storage hardware such as RAID (redundant array of inexpensive disks; striped with interleave parity) 5 and SSD (solid-state disks) will increase storage and speed. Data warehousing, database federation and search algorithms like Wolfram Alpha and Google will enhance our continual need for speed and accuracy in the foreseeable future.

The online access to visualization tools has been quite useful to casual browsers. Below we show several online tools that visitors might run across during their searches of genomic data. The first is an interactive genome map at NCBI, the second an interactive map at UCSC, the third is an analysis of a sample dataset at GEO. These tools make it possible to do much of the necessary analysis online, without the need for dedicated user-supplied tools. **See also**: Genome Sequence Analysis; Mining Biological Databases

1. Online UCSC genome map viewer: http://genome.ucsc.edu/cgi-bin/hgGateway.

Table 3 Bioinformatics tools websites

Description	Site	URL
Sequence alignment	BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi
	ClustalW2	http://www.ebi.ac.uk/Tools/clustalw2/index.html
Sequence alignment with genome	BLAT	http://genome.ucsc.edu/cgi-bin/hgBlat
Motif search	MEME	http://meme.sdsc.edu/meme4_1
	Prosite	http://ca.expasy.org/prosite/
Primer design	Primer3	http://frodo.wi.mit.edu/
Phylogenetics	Phylogenetic program	http://evolution.genetics.washington.edu/phylip/software.html
MHC binding	Rankpep	http://bio.dfci.harvard.edu/Tools/rankpep.html
	Syfpeithi	http://www.syfpeithi.de
Genome browser	UCSC genome browser	http://genome.ucsc.edu/cgi-bin/hgGateway

2. Online NCBI genome map viewer: http://www.ncbi.nlm.nih.gov/mapview/.
3. Online GEO analysis tools: http://www.ncbi.nlm.nih.gov/sites/GDSbrowser?acc = GDS10.

Tools and Software

The UCSC Genome Browser Database (GBD, http://genome.ucsc.edu) is a collection of genome sequence data and integrated annotations for a large number of organisms, including extensive genome comparison resources. In the past year, 13 new genome assemblies have been added, including two important primate species, orangutan and marmoset, bringing the total to 46 assemblies for 24 different vertebrates and 39 assemblies for 22 different invertebrate animals. UCSC sequences may be viewed graphically with the Genome Browser giving the user a reference point (chromosome position) with the choice of adding or subtracting other sequence-based and chromosome position-based data. Zooming in and out of the chromosome is rapid, finding homologies across species is very intuitive, and visualizing SNPs, mRNA (messenger RNA), DNA, gene predictions, regulatory elements, repeats, pairwise multiple genome alignments and even transcriptome data simultaneously is quite simple. A variety of other bioinformatics tools are provided, including BLAT, the Table Browser, the Gene Sorter, the Proteome Browser, VisiGene and Genome Graphs.

The NCBI genome map viewer is a data-driven visualization tool that allows many logs of zooming range to move from a single nucleotide out to see the local chromosome arm all the while showing information that is appropriate for the scale. The most striking feature of this sort of visualization is the sparce nature of actual coding genes in the genome. Intervening sequences (introns) are enormous, and coding regions make up only 1.5% of the human genome. The other striking feature is the complexity of alternative splicing. The genome map viewer shows the alternate splicing forms of a gene, so while the human genome may contain very few actual genes (even less than the roundworm or the ancient fern) the complexity of the human proteome is due to the diversity of splicing forms from a single multiexon gene.

GEO has arguably one of the best free online analysis suites for biological data. Heat maps, clustering, classification, gene-by-gene queries and biological search tools make the task of finding relevant data quite simple. Although microarray datafiles are very large, GEO has a browser and a series of query tools that enable one to navigate data very simply and quickly. This is not inherently easy – GEO made a requirement that all data that is uploaded must contain a minimal amount of information about the microarray experiment (aka MIAME). Unlike ArrayExpress at EBI, GEO decided to require only part of the official MIAME guidelines for their metadata, so submission of experimental data to GEO is not as onerous as submission to ArrayExpress.

Other tools and software components are available from a variety of sources, and all presented here are free (see **Table 3**).

Other databases: UniProt (Swiss-Prot/TrEMBL)

UniProt and its accessory databases and software collections (ExPASY) provides protein sequence and function data. UniProt is composed of UniProtKB (Knowledge-Base), the source for curated protein information, including function, classification and cross-references. Within UniProtKB is Swiss-Prot (manual annotation and review) and TrEMBL (automatic annotation, no review). UniRef provides clustered sets of sequences from the UniProtKB and some UniProt Archive records to obtain coverage of sequence space at high and low resolution while hiding redundant sequences. UniParc is a simple but comprehensive repository, to keep track of sequences and their identifiers. UniProt Metagenomic and Environmental Sequences (UniMES) database is a repository specifically developed for metagenomic and environmental data. These databases provide an excellent and orthogonal cross-reference when examining the coding potential inherent in genomic and EST sequence databases. Protein databases are an excellent cross-reference to other more structure/function databases like the Protein Databank (pdb, www.rcsb.org) and the Biomolecular Interaction Network Database (www.bind.ca).

The HapMap database

The goal of the International HapMap Project is to develop a haplotype map of the human genome, the HapMap, which will describe the common patterns of human DNA sequence variation. The HapMap is expected to be a key resource for researchers to use to find genes affecting health, disease and responses to drugs and environmental factors. Although any two unrelated people are the same at approximately 99.9% of their DNA sequences, the remaining 0.1% is important because it contains the genetic variants that influence how people differ in their risk of disease or their response to drugs. Discovering the DNA sequence variants that contribute to common disease risk offers one of the best opportunities for understanding the complex causes of disease in humans (http://www.hapmap.org/abouthapmap.html).

The gene ontology consortium

The Gene Ontology (GO) project is a collaborative effort to assign gene function, metabolic roles and cellular location of gene products. The GO project created three vocabularies that describe gene products in terms of their associated biological processes, cellular components and molecular functions. The original development of the vocabulary was the primary effort that got the GO project started. Once a standardized but dynamic list of functional terms was created, genes had to be placed within the Ontology. Since gene products can be multifunctinal the GO has redundant citizens (genes that show up in more than one location). Lastly, a volunteer group much continuously maintain and update the software that accesses these terms and makes the interface between the public and the knowledge as transparent as possible. Analysts today would have a hard time believing that at one time most genes had no defined function, even in simple bacteria. Today one can quickly examine expression or sequence data and find out the enzymatic/metabolic functions and cellular localization of almost any gene, even those from newly sequenced organisms (http://www.geneontology.org/GO.doc.shtml).

Biology in the future

Computer programs that predict protein structure based on primary amino acid sequence are becoming much better at learning protein structure and protein:protein interactions. The Brookhaven Protein Data Bank (PDB) is a three-dimensional structural database managed by the RCSB (Research Collaboratory for Structural Bioinformatics) and maintains structural data generated by researchers using X-ray crystallography and other methods. In the future, it may be possible to access the DNA database, retrieve a DNA sequence, translate that sequence into amino acids and find the three-dimensional structure of that protein. Even more likely would be the analysis of genomic sequence of cancer patients, discovery of a mutant cancer gene, followed by a search for small molecules that incapacitate that mutant gene product. As biology and computer technology advance, and more genomes are sequenced, there will be a renaissance of biological discovery based not necessarily on new observations, but on understanding existing sequence data and integrating many forms of that data. Biomedical research can utilize completed genomes to identify disease-associated genes and predict potential genetic problems as a diagnostic and prognostic tool. The central repositories of biological data will provide the central location for biomedical research. Interpreting the enormous amount of sequence data by bioinformatics tools will play a pivotal role in the future of drug discovery and healthcare improvements. Another key issue is the relationships between genotypes and disease-associated mutations versus real-world disease susceptibility and drug response. Many projects like the Cancer Genome Atlas Project (http://cancergenome.nih.gov/) and 1000 Genomes (http://www.1000genomes.org) project allow access to all of the (statistically relevant) sequence information one might desire. **See also**: Primary Protein and Nucleic Acid Three-dimensional Structure Databases; Protein Family Databases; Protein Sequence Databases; Protein Structure Prediction; Protein Tertiary Structures: Prediction from Amino Acid Sequences

Indeed, at the heart of the future of biomedical research, management and distribution of biological data will become increasingly important.

Further Reading

NCBI Handbook. Available at http://www.ncbi.nlm.nih.gov/books/bv.fcgi?rid=handbook.

Pevsner J (2009) *Bioinformatics and Functional Genomics*, 2nd edn. Hoboken, NJ: Wiley-Blackwell Publishers.

Tateno Y, Fukami-Kobayashi K, Miyazaki S, Sugawara H and Gojobori T (1998) DNA Data Bank of Japan at work on genome sequence data. *Nucleic Acids Research* **26**: 16–20.