Genome Databases

Introductory article

Marion L Carroll, *Louisiana State University Health Sciences Center, New Orleans, Louisiana, USA*

Son V Nguyen, *Louisiana State University Health Sciences Center, New Orleans, Louisiana, USA*

Mark A Batzer, *Louisiana State University Health Sciences Center, New Orleans, Louisiana, USA*

Genome databases are repositories of DNA sequences from many different species of plants and animals. They are linked electronically to supportive databases to aid in interpretation of the sequence data.

Article Contents

Introduction

A genome database can be described as a repository of DNA sequences from many different species of plants and animals. They are linked to supportive databases to aid in interpretation of the sequence data. Genome databases are designed and maintained in the electronic environment of one or several computers, using several operating systems, software applications, file transfer protocols and user interfaces. Genome databases contain sequence data generated by molecular biologists, geneticists and others using techniques in the laboratory that enable determination of the individual nucleotide order of a complete DNA sequence. DNA is the material within the cells of all living organisms that enable cellular processes to function efficiently, and is transmitted to future generations. Many publicly accessible databases can be viewed by web-browser from the GenBank Database server located at the National Center for Biotechnology Information (NCBI) (**Table 1**). GenBank is a database of nucleotide sequences from more than 58 000 organisms and is part of an international collaboration of sequence databases, which include the European Molecular Biology Laboratory (EMBL) and DNA Data Bank of Japan (DDBJ) (**Table 1**). GenBank receives and maintains DNA sequence submissions from multiple genome centres and university facilities participating in genome sequencing projects such as the Human Genome Project. EMBL and DDBJ warehouse data from various universities and genome centres in Europe and Japan and are collectively responsible for the collection, curating and dissemination of data generated from sequencing projects throughout the world.

The first genome database consisted primarily of small DNA viruses like the simian virus and phage ϕX174 containing about 5000 nucleotides of sequence. These successes in the late 1970s continued as 10-fold larger genomes of bacteriophage T7, lambda phage and larger genomes of bacteria and other single-cell organisms were sequenced. The sequence of the *Escherichia coli* genome was considered of great value as it is the most intensely studied bacterium. The *E. coli* genome contains 4.8×10^6 bases that required a concerted effort between many laboratories in order to be sequenced completely. In December of 1996, the sequencing of the *E. coli* genome was completed and warehoused at DDBJ (**Table 1**). Subsequently, the Human Genome Sequencing Project was initiated by a consortium of research laboratories around the world.

The human genome contains over 3 billion nucleotides that will be sequenced in multiple overlapping segments and stored electronically. These multiple overlapping segments are created using yeast, bacteria and viruses to harbour exogenous human DNA. A genomic library created using these organisms provides a source of fragmented genomic clones from which the DNA sequences of interest can be determined. There are several methods by which genome sequence data can be generated; however, the library sequencing method enables efficient reassembly of overlapping and contiguous DNA sequence fragments at completion. A well-constructed and curated filing system or database is required in order to maintain an accurate and accessible record of sequence information that will ultimately be used to piece together the contiguous DNA sequences representing a genome. The same type of information from the genomes of numerous mammalian species, plants, invertebrates and from single-cell microorganisms is also being generated and processed simultaneously. Here, we describe some genome databases, how they have been constructed to enable updating and editing and how they can be publicly accessed through the Internet.

What is a Database?

The volume of data created by genome mapping and sequencing projects is increasing; hence, reliable and efficient means to store these data are required. Manual collecting and depositing of data in a central repository such as a table with columns and rows had served as the initial organization of genome sequencing and mapping data. This is a tedious, time consuming and inefficient use of resources. The lack of synergy between large separate collections of these types of data was quickly recognized as a limitation. Informatics is the science that marries data

Table 1 Useful Website Resources

Resource	URL
Bacteria Database	[http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/5833.html] and [http://www.tigr.org/tdb/CMR/ghi/htmls/SplashPage.html]
C. elegans Database	[http://www.sanger.ac.uk/Software/Acedb/]
Celera	[http://www.celera.com/]
dbSNP	[http://www.ncbi.nlm.nih.gov/SNP/]
DDBJ	[http://www.ddbj.nig.ac.jp/]
ELSI	[http://www.nhgri.nih.gov/ELSI/]
Fly Database (FlyBase)	[http://fly.ebi.ac.uk:7081/]
GenBank	[http://www.ncbi.nlm.nih.gov/Genbank/index.html]
GenomeWeb	[http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-genome.html]
Human Database	[http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-genome.html]
Incyte	[http://www.incyte.com]
Jackson Laboratories	[http://www.jax.org/]
Legume Database (Soybase)	[http://genome.cornell.edu/cgi-bin/WebAce/webace?db = soybase]
Maize Database	[http://www.agron.missouri.edu/]
Mosquito Database	[http://klab.agsci.colostate.edu/index.html]
Myriad Genetics	[http://www.myriad.com/]
National Agricultural Library	[http://www.nal.usda.gov/]
National Human Genome Research Institute	[http://www.nhgri.nih.gov/About_NHGRI/]
NCBI	[http://www.ncbi.nlm.nih.gov/]
OMIM	[http://www3.ncbi.nlm.nih.gov/omim/]
Other animal genome databases (ArkDB)	[http://www.ri.bbsrc.ac.uk/arkdb/sites.html]
Puffer Fish Database	[http://fugu.hgmp.mrc.ac.uk/]
The Institute of Genome Research (TIGR):	[http://www.tigr.org/]
Tilapia Database	[http://www.ri.bbsrc.ac.uk/cgi-bin/arkdb/browsers/browser.sh?species = tilapia]
Yeast Database	[http://genome-www.stanford.edu/] and [http://www.bio.uva.nl/pombe/]
Zebrafish Database	[http://saturn.med.nyu.edu/zfish/pub/]

generated from laboratory experimentation with information technologies to enable the aggregation, interpretation and dissemination of useful combinations of relevant data. Informatics provides a semiautomated means to retrieve, filter and make comparisons and contrasts of data in an electronic format to minimize error, reduce redundancy between similar data sources and to make informed decisions about results. Tables or relational databases are frequently used as the framework on which many databases are developed. Alternatives include object-oriented methods that enable storage and retrieval of data as well as the flexibility to classify data and analyse results based on this data classification.

Relational databases

Relational systems are best described as being collections of joined tables as in an Excel or Lotus program. A table consists of a fixed number of columns or attributes and a variable number of rows containing cells for single data entry that is relevant to a particular attribute. Several tables of different attributes can be made more useful by linking or joining the cells in each table. In a relational database system the rows of data are presumably limited to the size and capacity of electronic storage with the columns or attributes expandable to include many different characteristics.

Object-oriented databases

The object-oriented method has proved highly successful in programming applications. An object-oriented database can classify data and use them as specific objects. These classes are arranged in a hierarchical structure that can inherit attributes from higher classes. In a biological system, we may have a class ‘experiment’ and a specialization of that class called ‘*in situ* hybridization experiment’. This gives object-oriented databases flexibility and extendibility – as genome mapping and sequencing evolves, classes are extended to take into account changing methods and techniques. Object-oriented databases are dynamic and can combine various classes to draw ‘virtual’ conclusions and generate an independent set of data. One can ask a 'gene object' to give its sequence. Then ask a marker object to draw itself on a genetic map. The modelling of classes controls the results of the application, placing relatively few limitations on the operations.

Commercial Databases

In addition to the publicly accessible databases there are a number of databases created by private companies for pharmaceutical research and development applications. A few of these companies include: Incyte Genomics [http://www.incyte.com/], Myriad Genetics [http://www.myriad.com/], and Celera [http://www.celera.com/].

Incyte Genomics

Incyte Genomics was one of the first biotech companies to engage in high-throughput computer-aided gene sequencing for the purpose of identifying genes and their corresponding proteins with potential therapeutic applications. Incyte's database discovery approach compares partial human genes or protein sequences to genes or proteins of known sequence in order to predict their biological or therapeutic function. Incyte used its approach to identify specific white blood cell proteins that might have pharmaceutical utility. Incyte provides a platform of genomic technologies designed to aid in the understanding of the molecular basis of disease. The sequence data is accessible through LifeSeq Public, which contains 1.4 million sequences in its public sequence domain plus 90 000 sequences from Incyte's proprietary database.

The Institute for Genomic Research

The Institute for Genomic Research (TIGR) in Rockville, Maryland inaugurated an era of molecular medicine by the identification of the majority of expressed human genes (ESTs, expressed sequence tags) (**Table 1**). TIGR has been at the forefront of the effort to sequence, analyse, and curate coding sequences using a strategy developed by founder and former president J. Craig Venter, now CEO of Celera. TIGR researchers have identified over half of the estimated 60 000 to 80 000 human genes. Nearly 40 percent of the genes found thus far in all species have no assigned role in cell metabolism or physiology. TIGR is devoting considerable effort to understanding the biological functions of these genes.

Myriad Genetics, Inc.

Myriad Genetics, Inc. is a biopharmaceutical company focused on the development of therapeutic and diagnostic products using genomic and proteomic technologies (**Table 1**). Myriad has developed a proprietary protein interaction network technology, ProNet®, which is used by the company to identify and analyse protein interactions. Because protein interactions mediate most cellular processes, identification of a protein's interacting complexes is critical in understanding its function and importance in disease processes. Understanding of protein function has a profound impact on defining drug targets and discovering novel therapeutics. Given an interesting biological pathway, Myriad will work with the pharmaceutical customer on an exclusive basis to characterize proteins in these pathways and determine their biological function.

Myriad Genetics currently markets two genetic tests, BRACAnalysis® and CardiRisk™. Inherited mutations in the *BRCA1* and *BRCA2* genes are responsible for approximately 7–10% of all breast and ovarian cancers. BRACAnalysis is a test for genetic predisposition to breast and ovarian cancer based on comprehensive DNA sequence analysis of the *BRCA1* and *BRCA2* genes. This test provides information on an individual's cancer risk helping physicians and their patients make better-informed health care decisions. Myriad also maintains a database of mutations identified in the *BRCA1* and *BRCA2* genes.

High blood pressure, cholesterol, smoking and lack of exercise can all lead to increased risk of heart disease. But half of all people who develop heart disease have none of these risk factors. Heredity must then be attributing to the increased risk in these patients. CardiaRisk™ analyses a specific part of a gene called *AGT*. This gene controls the production of angiotensinogen, a protein important for the regulation of blood pressure and heart function. CardiaRisk is the first clinically available service for evaluating genetic predisposition to cardiovascular disease providing a new and essential approach to risk assessment.

Celera

Celera provides access to comprehensive genomic databases, as well as proprietary software tools for viewing, browsing and analysing its data. The core sequences of Celera's Genome Reference Database are human, *Drosophila melanogaster* and mouse genomes sequenced by Celera. These sequences are integrated with more than 20 other reference databases and a suite of essential sequence analysis tools. Celera's database is a curated set of the genes, mRNAs, proteins and regulatory elements encoded in the human genome and in the genomes of other organisms important to normal human physiology and in disease mechanisms. This database serves to aid pharmaceutical companies by accelerating the understanding and use of genomic and related information. Celera has established the largest genomic production plant in the world, supported by one of the largest civilian supercomputing facilities.

The foundation of Celera's reputation is its public announcement of the complete sequencing of *D. melanogaster* and the near complete sequencing of the mouse and human genomes. The company published the assembled and annotated *D. melanogaster* genome on March 24, 2000. The nearly complete sequence of one human was announced in April 2000, which was followed by the announcement of the first assembly of a large percentage of the human genome in June 2000. To be included in the assembly of these various genomes are descriptions of gene function and expression, comparisons to other model organisms, data on protein structure and composition, and data on genetic variation. The resulting databases and informatics tools will be available by subscription to pharmaceutical, biotechnology and life science research organizations. These databases will facilitate efforts to identify novel genes for drug discovery and development, to improve agricultural products, and help develop therapies for individuals based on their unique genetic profiles.

Existing Genome Databases

Parasites and disease organisms

Over 700 species of bacterial, eukaryote and viral genomes have been completely or partially sequenced. Several of these species are responsible for diseases in many different plants and animals, including humans. An example of a disease with particular impact on world health is malaria caused by four known single cell parasites of which the most infective is *Plasmodium falciparum*. The Institute for Genomic Research (TIGR), the Naval Medical Research Center, the Sanger Centre and Stanford University are all taking part in sequencing the genome of *P. falciparum* and maintaining the sequence in a database that is accessible from NCBI [http://www.ncbi.nlm.nih.gov/]. This database contains sequence data for each of the 14 chromosomes of *P. falciparum* and information relevant to its genome. It also contains sequence alignments, genome maps, linkage markers and information about genetic studies. Links are provided to other websites and genetic data on related parasites. The sequences of many other pathological organisms can be found at NCBI simply by following the site map provided in the legend of its home page.

The genomic sequence of the common flu microbe *Haemophilus influenzae* Rd strain KW20^3 [http://www.ncbi.nlm.nih.gov/htbin-post/Taxonomy/wgetorg?-name = Haemophilus + influenzae] has also been completed. This is a small, nonmotile, Gram-negative bacterium whose only known natural host is humans. The common flu results from these commensal residents of the upper respiratory mucosa of children and adults, producing otitis media, respiratory infections and even meningitis by more virulent strains.

The most prevalent infectious disease within the human population worldwide is tuberculosis. The sequence of the bacteria *Mycobacterium tuberculosis* [http://www.sanger.-ac.uk/Projects/M_tuberculosis/] was completed in 1998. The genome is composed of 4 411 529 base pairs, and contains about 4000 genes. This knowledge will enable scientists to begin to understand the unusual biochemical properties of this microorganism. Within this database an updated list of predicted protein-coding genes can be found. There is also a separate list of non-protein coding genes. Interestingly, a very large portion of the *M. tuberculosis* coding capacity is devoted to the production of enzymes involved in lipid metabolism. Additional links to the TubercuList database at the Institut Pasteur provide a more complete analysis of protein structure and function. The MycDB contains information on all aspects of mycobacteria, relevant to the biology and pathobiology of the organism. This database contains data on physical mapping, nucleotide sequences, antigens, antibodies and relevant literature citations.

Caenorhabditis elegans is a member of the Rhabditidae, a large and diverse group of nematodes found in terrestrial habitats. Some rhabditids are pathogenic to or parasitic on animals. *C. elegans* has facilitated significant progress in the study of neurological function and development since the 1960s for several reasons. The adult soil nematode is a multicellular organism with a short life cycle. It can be easily cultivated and is small enough at 1 mm to be handled in large numbers like a microorganism. It has relatively few cells so that exhaustive studies of lineage and evolutionary patterns can be made and the entire organism is easily subject to genetic analysis. *C. elegans* is diploid and has five pairs of autosomal chromosomes and a pair of sex chromosomes (X), which occupy a genome of 97 Mb (97×10^6 bases).

Each chromosome being sequenced is warehoused in GenBank and also in a database designed as a result of the *C. elegans* project called ACeDB (for A *C. elegans* DataBase). ACeDB is a genome database system in development since 1992. It provides a custom database kernel, with a nonstandard data model designed specifically for handling scientific data flexibly, and a graphical user interface with many specific displays and tools for genomic analysis. ACeDB is used both for managing data within genome projects, and for making genomic data available to the scientific community. ACeDB sequence analysis tools have been generalized to be much more flexible and useful in the analysis of genomic databases from bacteria to humans.

Insects

Like *C. elegans*, the insect genome databases are extremely valuable to scientists in providing sources for modelling and understanding the complexity of the interaction of genes in multicellular organisms. A joint DNA sequencing effort between the University of California-Berkeley and European Drosophila Genome Projects has enabled the warehousing of DNA sequence data and a complete set of tools for the analysis of genome structure and function of the common fruitfly *Drosophila melanogaster*. This effort has resulted in the development of the genome database called FlyBase. The FlyBase project is carried out by a consortium of *Drosophila* researchers and computer scientists at Harvard University, University of Cambridge (UK), Indiana University, University of California and the European Bioinformatics Institute. The *D. melanogaster* genome contains four chromosomes consisting of about 180 000 000 base pairs that encode about 14 000 genes. Many of these genes have aided scientists in understanding some of the processes involved in human development at the molecular level. For example, the genes that determine which cells will become wings and which will become legs have been identified in *Drosophila*. Homologues or similar genes are involved in aspects of human limb bud and finger development.

Mosquitoes are another insect species whose genome is of interest. The order Diptera or the 'true flies' contain over 2700 varieties of mosquitoes. Some mosquitoes are capable of acting as vectors or carriers, transmitting diseases such as malaria and yellow fever to humans, encephalitis to humans and horses, and heartworm to dogs. The Mosquito Database (MsqDB) is devoted to both the genetic and physical chromosome mapping data of *Aedes aegypti*, *Anopheles gambiae*, *Culex pipiens* and other species. These studies help to elucidate the molecular biology and life cycle of these insects in order to control breeding, migration patterns and the spread of infectious diseases.

Plants

Curators and collaborators who are responsible for collecting, organizing and evaluating plant DNA sequence and expression data provide this information to the National Agricultural Library (NAL) who make this data available to the public. Of all the species of plants that have been studied, few have had as remarkable an influence on modern society as *Zea mays*. Maize or corn of the most common strain, teosinte (*Zea mays* subsp. *parviglumis*) is a wild grass occurring naturally in isolated patches currently restricted to elevations between 400 and 1700 m in the Mexican western Sierra Madre. The giant domesticated grass species (*Zea mays* subsp. *mays*) originates from tropical Mexico and is used to produce grain and fodder that are the basis of a number of food, feed, pharmaceutical and industrial products. Cultivation of maize and the manufacturing of its food products are associated with the rise of pre-Colombian MesoAmerican civilizations.

The maize genome database has been made available to cultivators and scientists as a source for studying the high genetic biodiversity in the Mexican maize pool, a factor of great importance for the breeding of current and future maize species. The maize genome consists of 10 chromosomes containing 3800 Mb of sequence. A yeast artificial chromosome library has made the analysis of this amount of data possible. This genome is fully searchable at the Maize Database (MaizeDB) (**Table 1**). Also available at this site are query/search forms that enable the retrieval of a number of categories including corn stocks, karyotype variations, gene maps, gene products, quantitative linkage analysis and Medline references.

Arabidopsis thaliana is a small flowering plant that is widely used by plant researchers as a model organism to study many aspects of plant biology. It is a member of the brassica family, which includes species such as cabbage and radish. *Arabidopsis* has several important advantages for the researchers in many areas of plant biology, especially genetics and molecular biology. It has a rapid life cycle, is easily cultivated and its genome is small, consisting on average of 135 Mb distributed among five chromosomes. The genome database of *Arabidopsis* contains approximately 112 Mb of sequence data. The Arabidopsis Information Resource (TAIR) is a collaborative project between the Carnegie Institution of Washington Department of Plant Biology and the National Center for Genome Resources. Each maintains genomic and literature information on the progress of this project.

The SoyBase project is part of the legume database project coordinated, developed and maintained at Iowa State University and funded by the USDA-ARS Plant Genome Program. This laboratory carries out projects of technology development and standards evaluation of soybean diversity and quantitative trait loci analyses. The primary objective of the laboratory is to develop and apply knowledge and technologies that will assist in the enhancement of soybeans for breeding. A bacterial artificial chromosome (BAC) library was created and consists of approximately 40 000 clones or four to five genome equivalents. A random sampling of 224 BAC clones yielded an average DNA insert size of 150 kb, giving a 98% probability of recovering any specific sequence. The SoyBase contains searchable catagories of data on plant genes, chromosome maps and DNA fingerprints that enable discrimination of plant variety and beneficial traits. There is also a collection of plant pathologies and proteins involved in metabolic pathways.

Fish genomes

The fish genome projects are as important to the understanding of aquaculture as plant genome projects are to the understanding of agriculture. In regions of the world where aquaculture is central to the economy, sustaining the vitality of various species of fish is imperative. In the past many techniques used were limited to the study of the developmental patterns, physiology and behaviour of fish. However, the availability of fish databases is making it possible to look at the genetics of these species, thereby facilitating the improvement of strains with respect to traits of commercial importance such as growth rate and flesh quality.

Tilapia (*Oreochromis* sp.) is a group of perch-like fish (family Cichlidae) native to the freshwaters of tropical Africa. They are one of the most important aquatic species in today's culture. The first relatively complete genetic map for a tilapia (*Oreochromis niloticus*) is currently available (**Table 1**). The maps are listed as linkage groups 1–30 (lg1–lg30), and efforts are underway to consolidate them into the 22 chromosomes that make up the genome of this species.

The puffer fish (*Fugu rubripes*) genome project is near completion. Information from this database (**Table 1**) reveals that this species of fish has essentially the same number of genes as the human genome, yet its genome size is only 400 Mb as compared to 3000 Mb in the human genome. Thus, identification of puffer fish homologues to human genomic regions will facilitate gene identification in humans.

The zebrafish has recently emerged as a premier organism for studies of vertebrate development and genetics. As a result, a comprehensive database (The Zebrafish Information Network) includes a broad range of data types generated by zebrafish research, including text, images and graphical information about mutations, gene expression patterns and the genetic maps. Data from the Stanford Zebrafish Genome Project has revealed extensive conservation of syntenic relationships among vertebrate genomes. The zebrafish genome is divided into 25 linkage groups (**Table 1**).

Yeast genomes

Saccharomyces cerevisiae, or baker's yeast, is an ideal eukaryotic microorganism for biological study. Yeast are easy to propagate and their genomes are easy to manipulate. This allows for the construction of an informative model of gene structure and function that can be compared with other eukaryotes, like humans. The complete sequence of the 16 chromosomes (13 Mb) of the *S. cerevisiae* genome was completed in 1996 by more than 100 laboratories from Europe, USA, Canada and Japan. *S. cerevisiae* as well as *Schizosaccharomyces pombe* (for which there are 14 Mb within three chromosomes) are very important as references to understanding the biology of humans and other higher eukaryotic organisms. The *Saccharomyces* Genome Database (SGD) provides mapping and sequence information as well as a comprehensive analysis of proteins and their functions. There is also a registry of yeast genes and file transfer protocol (FTP) site links for downloading tables of nucleotide sequence changes to some of the chromosomes. These tables contain summaries of the changes made to the systematic nucleotide sequence represented by SGD, the source of these changes, and a link to a table explaining the gene name changes.

Mammalian genomes

Many vertebrates are useful in developing models for biological processes. Mice have been viewed as the quintessential organism for modelling biological processes similar to those in humans. To this end, several databases have been constructed since 1991 to allow for warehousing and navigating the mouse genome. Facilities like those at the Jackson Laboratory in Bar Harbor, Maine, USA provide resources throughout the world for projects involving many normal, mutant and transgenic mouse and rat species. Informatic tools developed by the genome database groups at this facility enable access to databases that are used to capture, store, and manage high-quality data suitable for sequence and gene expression analyses and integration with other relevant biological and computer data.

In 1994 the genome mapping database called Ark DataBase (ArkDB) was adapted and developed to be linked to a variety of agricultural livestock. These species include pig, chicken, sheep, cattle and horse. The database (**Table 1**) consolidates all genome data for a particular species in a central resource. A complete data trail is present to allow tracing of data from the original maps, through experiments back to the original papers and primary data.

Humans

Though not as large as some plant genomes (e.g. maize) the 3000 Mb of human genome sequence divided into 22 autosomes, X and Y has been the source of intense investigation. The potential of a complete human genome sequence to science and medicine will be tremendous once the knowledge of all coding sequence for the complete genome (approximately 100 000) is determined. The National Human Genome Research Institute (NHGRI) (**Table 1**), originally established in 1989 and one of 24 institutes, centres or divisions that make up the National Institutes of Health (NIH) in Bethesda, Maryland, USA, is projecting to meet the human genome sequencing goal by 2003.

Part of the extramural research division at the NHGRI is the Ethical, Legal and Social Implication Research Program (ELSI) which is critical to understanding, maintaining and protecting the rights of individuals who may be affected by the human genome sequencing project.

Also, it is necessary to maintain control over those institutions participating in the generation of sequence data. ELSI focuses on five major concerns:

1. Issues surrounding sequence completion and the study of human genetic variation.
2. Issues raised by the integration of genetic with public and private health care.
3. Issues raised by the integration of genetics with environmental interactions.
4. Exploring genetics from a variety of philosophical, theological, and ethical perspectives.
5. Exploring how concepts of race and ethnicity are influence by genetic information.

The UK Human Genome Mapping Project Resource Centre (HGMP-RC) provides access to leading-edge tools for research in the fields of genetics and functional genomics. It is located on the Hinxton Genome Campus along with the Sanger Centre and the European Bioinformatics Institute. The GenomeWeb found at HGMP-RC is an up-to-date listing of all the URLs and site descriptions involved in characterizing human genome sequence data. Each day, URLs are extracted and retrieved through their proxy cache server so that documents are cached locally and anyone using the proxy server will get fast access to relevant documents. This process also verifies that the URL is valid and that the document has not disappeared or moved.

Among the sites at GenomeWeb is a link to the Online Mendelian Inheritance in Man (OMIM) database (**Table 1**). This is a database of human genes and genetic disorders curated by Johns Hopkins University and others and used by NCBI. The database contains textual information, pictures and reference information. It also contains NCBI's Entrez database of Medline articles and sequence information related to genetic diseases and inherited disorders.

The OMIM database is updated regularly with new discoveries of genes and disorders in humans. For example, mammalian neurexophilin was discovered as a 29-kDa neuronal glycoprotein that copurifies in a tight complex with neurexin I-alpha (600565). By searching sequence databases using the amino acid sequence of mammalian neurexophilin as the query, human expressed sequence tags (ESTs) corresponding to a family of proteins (neurexophilins-1 (NXPH1; 604639),-2 (NXPH2),-3 (NXPH3; 604636), and-4 (NXPH4; 604637)) were identified. The OMIM database facilitates the cataloguing as well as the ability to compare within and between species for characterizations that may suggest a function for newly identified genes.

Another database to facilitate the study of the human genome is the Human Genome Single Nucleotide Polymorphism Database (dbSNP). As of February 2000 there had been 26 397 submissions into the dbSNP. The human genome project has enabled genetic research to discover and associate very common single nucleotide variations or polymorphisms with heritable phenotypes. These types of variations occur approximately once in every 100–300 bases. In a collaboration between the National Human Genome Research Institute and the National Center for Biotechnology Information the dbSNP database (**Table 1**) was established to serve as a central repository for both single base nucleotide subsitutions and short deletion and insertion polymorphisms.

The human genome sequencing project is a worldwide cooperative effort to bring to the public the complete sequence of the human genome. The potential to science and medicine is enormous. An immediate value has been realized from this effort by computer science and information technologies. The tremendous data storage, assimilation, retrieval and communication requirements of genome sequence databases have created a challenge within these fields that few if any projects in the biological sciences have required. As a result, complex analytical tools will continue to be developed to analyse the genomes of new species of bacteria, plants and animals as they are sequenced to completion.

Acknowledgements

MAB was supported by award 1999-IJ-CX-K009 from the Office of Justice Programs, National Institute of Justice, Department of Justice. Points of view in this document are those of the authors and do not necessarily represent the official position of the US Department of Justice.

Further Reading

Benson DA, Boguski MS, Lipman DJ *et al.* (1999) GenBank. *Nucleic Acids Research* **27**: 12–17.

Blake JA, Richardson JE, Davisson MT, Eppig JT and the Mouse Genome Database Group (1999) The Mouse Genome Database (MGD): genetic and genomic information about the laboratory mouse. *Nucleic Acids Research* **27**: 95–98.

Cole ST, Brosch R, Parkhill J *et al.* (1998) Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537–544.

Durbin R and Thierry-Mieg J (1991) A *C. elegans* Database. Documentation, code and data available from anonymous FTP servers at: [lirmm.lirmm.fr], [cele.mrc-lmb.cam.ac.uk] and [ncbi.nlm.nih.gov].

Edwards KJ, Thompson H, Edwards D *et al.* (1992) Construction and characterization of a yeast artificial chromosome library containing three haploid maize genome equivalents. *Plant Molecular Biology* **19**: 299–308.

FlyBase (1999) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Research* **27**: 85–88.

McKusick VA (1998) *Mendelian Inheritance in Man. Catalogs of Human Genes and Genetic Disorders*, 12th edn. Baltimore: Johns Hopkins University Press.

Missler M and Sudhof TC (1998) Neurexophilins form a conserved family of neuropeptide-like glycoproteins. *Journal of Neuroscience* **18**: 3630–3638.

Shoemaker RC (1999) Soybean genomics from 1985–2002. *AgBiotech-Net* **1**: 1–4.

Watson JD (1990) The Human Genome Project: past, present and future. *Science* **248**: 44–51.

Wilcox KW and Smith HO (1975) Isolation and characterization of mutants of *Haemophilus influenzae* deficient in an adenosine 5′-triphosphate-dependent deoxyribonuclease activity. *Journal of Bacteriology* **122**: 443–453.

Bacteria Database, [http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/5833.html] and [http://www.tigr.org/tdb/CMR/ghi/htmls/SplashPage.html]

C. elegans Database, [http://www.sanger.ac.uk/Software/Acedb/]

Celera, [http://www.celera.com/]

dbSNP, [http://www.ncbi.nlm.nih.gov/SNP/]

DDBJ, [http://www.ddbj.nig.ac.jp/]

ELSI, [http://www.nhgri.nih.gov/ELSI/]

Fly Database (FlyBase), [http://fly.ebi.ac.uk:7081/]

GenBank, [http://www.ncbi.nlm.nih.gov/Genbank/index.html]

GenomeWeb, [http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-genome.html]

Human Database, [http://www.hgmp.mrc.ac.uk/GenomeWeb/human-gen-db-genome.html]

Incyte, [http://www.incyte.com]

Jackson Laboratories, [http://www.jax.org/]

Legume Database (Soybase), [http://genome.cornell.edu/cgi-bin/WebAce/webace?db = soybase]

Maize Database, [http://www.agron.missouri.edu/]

Mosquito Database, [http://klab.agsci.colostate.edu/index.html]

Myriad Genetics, [http://www.myriad.com/]

National Agricultural Library, [http://www.nal.usda.gov/]

National Human Genome Research Institute, [http://www.nhgri.nih.gov/About_NHGRI/]

NCBI, [http://www.ncbi.nlm.nih.gov/]

OMIM, [http://www3.ncbi.nlm.nih.gov/omim/]

Other animal genome databases (ArkDB), [http://www.ri.bbsrc.ac.uk/arkdb/sites.html]

Puffer Fish Database, [http://fugu.hgmp.mrc.ac.uk/]

The Institute of Genome Research (TIGR), [http://www.tigr.org/]

Tilapia Database, [http://www.ri.bbsrc.ac.uk/cgi-bin/arkdb/browsers/browser.sh?species = tilapia]

Yeast Database, [http://genome-www.stanford.edu/] and [http://www.bio.uva.nl/pombe/]

Zebrafish Database, [http://saturn.med.nyu.edu/zfish/pub/]