Protein Databases

Vivienne Baillie Gerritsen, *Swiss Institute of Bioinformatics, Geneva, Switzerland*
Amos Bairoch, *Swiss Institute of Bioinformatics, Geneva, Switzerland*

An abundance of protein databases are available, dealing with fields as diverse as protein sequences, protein domains, posttranslational modifications and protein–protein interactions. Such resources are crucial to proteomics research.

Advanced article

Article contents

doi: 10.1038/npg.els.0005251

Introduction

Protein databases have been around for the best part of half a century. One of the very first protein databases was the Atlas of Protein Sequence and Structure developed by the late Margaret Dayhoff who founded the Protein Information Resource (PIR). A series of books were published from 1965 to 1978 until the quantity of data grew so much that an electronic form was made available to the scientific community, known as the PIR-International Protein Sequence Database. Swiss-Prot, the protein sequence knowledgebase founded in 1986 by Amos Bairoch, took its inspiration from PIR but strove to develop a database that was nonredundant and extremely well documented. In the past four decades, a great many diverse databases have sprouted: protein sequence databases, two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) databases, three-dimensional (3D) structure databases, posttranslational modification databases and metabolic databases among others. It is now common knowledge that information on proteins is crucial to all biological research because of the functions these molecules carry out in cells and the roles they have in disease processes.

Improvements in deoxyribonucleic acid (DNA) sequencing technology are generating information on hundreds of thousands of novel proteins and the number is growing exponentially with ongoing genomic projects. Today, with this increasing volume and variety of information on proteins, databases of databases are budding, and it is toward these central resources that scientists active in modern biological research should turn, particularly in the field of proteomics. This is why it is impossible to list every database that exists, especially as their stories often may be compared with the birth and death of stars: some are long-lived, while others peter out within months of existence. The range of protein databases, however, can be divided into categories, which we have listed below, bearing in mind those relevant to research in the field of human genomics. A source of over a thousand databases of interest to proteomic research is available (see Web Links).

Protein Sequence Databases

The most comprehensive source of information on proteins is found in protein sequence databases, of which there are two types:

- universal databases whose aim is to collect biological information on the most varied amount of species and
- specialized databases that cater for specific groups or families of proteins or specific organisms.

A universal protein knowledgebase: United Protein Databases or UniProt

The availability of an ever-increasing volume and variety of protein sequences and functional information has seen the creation of a number of protein sequence knowledgebases, namely Swiss-Prot and TrEMBL, operated by researchers from Switzerland and the European Molecular Biology Laboratory (EMBL), and the PIR at Georgetown University Medical Centre and the National Biomedical Research Foundation (US). These groups combined their strengths into a central public resource: the United Protein Databases or UniProt. The UniProt nonredundant database is built around Swiss-Prot and TrEMBL, and each UniProt entry is a central hub for data regarding a specific protein.

Swiss-Prot and TrEMBL (see Web Links)

Swiss-Prot (Boeckmann *et al*., 2003) is a nonredundant curated protein knowledge resource that provides a high level of annotation. Besides the stark protein sequence, a Swiss-Prot entry offers a wide variety of information including the description of the function of a protein, its domain structure, posttranslational modifications, variants, numerous links to all sorts of other databases, and so on.

Early on it became apparent that Swiss-Prot alone could not cope with the flow of information on

hundreds of thousands of new proteins resulting from the never-ending improvements in DNA sequencing technology. In response, a supplementary database, Translation of EMBL nucleotide sequence database or TrEMBL, was created. TrEMBL (Boeckmann *et al*., 2003) consists of computer-annotated entries in Swiss-Prot format derived from the translation of coding sequences (CDs) in the EMBL nucleotide sequence database. Hence TrEMBL entries are preliminary Swiss-Prot entries that have not yet been manually annotated.

Specialized Protein Sequence Databases

There are a wide variety of databases dedicated to groups of proteins, or families of proteins. Some have no more than a handful of entries, while others are less modest and provide a far wider scope. It would be quite pointless and near impossible to give a list of all existing specialized protein sequence databases. What is more, the development of many is stunted after a short existence, while new ones sprout almost on a daily basis. A positive side to all this is the appearance of information systems that attempt to collect specific data into a central resource. In a feeble attempt to show the diversity, three specialized protein sequence databases are briefly described.

G-protein-coupled receptor database: GPCRDB (see Web Links)

GPCRDB (Horn *et al*., 1998) is an information system that collects and disseminates GPCR-related data. It holds sequences, mutant data and ligand-binding constants as primary (experimental) data. Computationally derived data such as multiple sequence alignments, 3D models, phylogenetic trees and 2D visualization tools are added to enhance the database's usefulness.

A protease database: MEROPS (see Web Links)

MEROPS (Rawlings *et al*., 2002) provides a wealth of information on proteases. There are data on individual proteases, protease families and also clans into which the families are grouped. Hundreds of proteases can be found by name, identifier or the organism in which they occur.

International ImMunoGeneTics database: IMGT (see Web Links)

IMGT (Lefranc, 2001) is a high-quality integrated information system that specializes in immunoglobulins, T cell receptors and major histocompatibility complex molecules of all vertebrate species. It includes sequence databases, Web resources and interactive tools, and the IMGT server provides common access to data relative to the field of immunogenetics.

Protein Domain and Family Databases

The sequence of a new protein can be so distantly related to any other that the detection of any resemblance by similarity searches is obsolete. However, proteins do have their fingerprints, otherwise known as patterns, motifs or signatures. These are particular clusters of residue types in the sequence, which reflect conserved regions important to the function of the protein. A popular way to identify such motifs between proteins is to perform a pairwise alignment. When the identity is higher than 40%, this method gives good results. However, the weakness of the pairwise alignment is that no distinction is made between an amino acid at a crucial position (like an active site) and an amino acid with no critical role. A multiple sequence alignment gives a more general view of a conserved region by giving a better picture of the most conserved residues, which are also usually those essential for the protein's function. Several databases have developed their own methods based on multiple sequence alignment in order to identify conserved regions (**Table 1**). A search performed on these databases is very often more sensitive than a pairwise alignment and can help to identify very remote homology (less than 20%).

InterPro (see Web Links)

InterPro (Mulder *et al*., 2002) is an integrated documentation resource for protein families, domains and functional sites that was developed to rationalize the complementary efforts of the individual protein signature database projects. PRINTS, PROSITE, Pfam, ProDom, SMART and TIGRFAMs form the InterPro core. Each InterPro entry includes a unique accession number, functional descriptions and literature references, and links are made back to the relevant member databases.

InterPro is a useful resource for whole-genome analysis and has already been used for the proteome analysis of a number of completely sequenced organisms, including preliminary analyses of the human genome. **Table 1** gives a list of the InterPro database members as well as a brief description and their URL addresses.

Table 1 InterPro database members

Database (Ref)	Description	Method	URL
PROSITE (Falquet *et al*., 2002)	A well-documented database of protein families and domains	Regular expressions and generalized profiles	http://www.expasy.org/prosite/
Pfam (Bateman *et al*., 2002)	A large collection of multiple sequence alignments and hidden Markov models covering protein domains and families	Profiles based on hidden Markov models	http://www.sanger.ac.uk/Software/Pfam/index.shtml
SMART (Letunic *et al*., 2002)	A collection of protein families and domains	Profiles based on hidden Markov models	http://smart.embl-heidelberg.de/
TIGRFAMs (Haft *et al*., 2001)	A collection of protein families with an emphasis on microbial proteins	Profiles based on hidden Markov models	http://www.tigr.org/TIGRFAMs/index.shtml
PRINTS (Attwood *et al*., 2002)	A well-documented database of conserved motifs used to characterize protein families	Fingerprints	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/
ProDom (Corpet *et al*., 2000)	Protein domain database	Automated clustering of homologous domains based on iterative PSI-BLAST searches	http://prodes.toulouse.inra.fr/prodom/doc/prodom.html

Three-dimensional Structure Databases

The primary structure of a protein dictates its spatial or tertiary structure. The knowledge of the 3D structure of a protein is crucial for the understanding of its precise function, drug design and other biotechnological applications. Today still, the elucidation of a protein's 3D structure is a labor-intensive and technically demanding process. However, the combination of fast X-ray detectors, nuclear magnetic resonance (NMR) methods, modern biotechnological methods and access to synchrotron X-ray sources has accelerated the process from months to days.

PDB: Protein Data Bank (see Web Links)

PDB (Berman *et al*., 2000) is a collection of 3D structures of proteins but also nucleic acids and other biological macromolecules. It is in fact the sole worldwide archive of structural data of biological macromolecules currently holding almost 20 000 entries. Naturally, this number is expected to grow rapidly once the genomics findings are combined with the structural findings. PDB remains a resource of tremendous and critical importance in the discovery of new pharmacological agents, new catalysts, new biomaterials and possibly even nanodevices.

Derived structural databases

All proteins tend to have structural similarities, with others echoing a common evolutionary origin. There are a number of derived databases that merge structural and sequence information directly from PDB. Two can be mentioned: the SCOP database and the CATH domain database (see Web Links). The SCOP database (Murzin *et al*., 1995) provides a detailed and comprehensive description of the structural and evolutionary relationships between proteins whose structure is known. CATH (Orengo *et al*., 1997) is a hierarchical domain classification of protein structures derived from PDB.

Other Types of Database

There is a never ending variety of databases, which simply reflects the infinite variety of research that goes on worldwide. As for all databases, some die fast while others develop over the years to offer a wider and better range of tools for the biological sciences. Besides protein sequence databases, protein domain databases and 3D structure databases, the scientific community offers databases that collect information on posttranslational modifications, 2D-PAGE databases and protein–protein interaction databases, to name only three.

Posttranslational modification databases

Once sequenced, most proteins are the target of posttranslational modifications (PTMs). Indeed, without such modifications, such proteins cannot function. It is therefore of prime importance to characterize PTMs. There are few databases with information on PTMs because the vast majority of this kind of information is already held with the protein knowledgebase Swiss-Prot. However, the following is worth mentioning.

GlycoSuiteDB: a database of glycoprotein glycan structures (see Web Links)

GlycoSuiteDB (Cooper *et al*., 2001) is based on information derived from the scientific literature and provides detailed information on the glycan structure. Regarding proteins, when the glycan structures are known to be attached to a specific protein, direct links are made to Swiss-Prot and TrEMBL.

Two-dimensional polacrylamide gel electrophoresis databases

A growing number of databases dedicated to proteins identified on 2D-PAGE have appeared in the recent years and their impact on proteome research is already quite significant. 2D-PAGE databases contain two separate components: image data and text information. The former consists of one or more reference gel maps for a given biological sample (tissue, physiological fluid, or a cell in the event of free-living organisms); the latter gives detailed information on each of the spots on the maps, such as their apparent molecular weight and isoelectric point on the map, the name of the protein, the method of identification and cross-references to relevant databases.

Swiss-2Dpage: two-dimensional polyacrylamide gel electrophoresis database (see Web Links)

Swiss-2Dpage (Hoogland *et al*., 2000) was created and is maintained collaboratively by the Central Clinical Chemistry Laboratory of the Geneva University Hospital and the Swiss Institute of Bioinformatics (SIB). It contains 2D-PAGE and sodium dodecyl sulfate (SDS) PAGE reference maps and information on identified proteins from a variety of human biological samples, such as the liver, plasma, colon and platelets. The proteins in Swiss-2Dpage have been identified by the methods of microsequencing, immunoblotting, gel comparison, amino acid composition, peptide mass fingerprinting and/or tandem mass spectrometry.

World-2Dpage (see Web Links)

There is, as for most types of database, an increasing number of 2D gel databases. World-2Dpage (Hoogland *et al*., 1999) is a complete index of 2D-PAGE databases and services. It not only offers the list of databases and their web addresses, but also gives information on the organism and the tissue or fluid involved, as well as sites for laboratory services, 2D-PAGE training, image analysis and links to related meetings and societies.

Protein–protein interaction databases

Protein–protein interactions (PPI) lie at the heart of most biological processes: signal transduction, metabolic pathways and immune response. PPI data are of crucial scientific and medical relevance; indeed understanding the interactions between encoded proteins of a given genome is a critical step in functional genomic analysis. Several PPI databases have been compiled to document and describe protein–protein interactions.

DIP: Database of Interacting Proteins (see Web Links)

The DIP database (Xenarios *et al*., 2002) is a collection of PPIs that are determined experimentally. A consistent set of PPIs is the result of information gathered from a variety of sources. Data within DIP is curated both manually by expert curators and automatically. The database provides a comprehensive and integrated tool for browsing and extracting information on PPIs. Additional information can be found on the position of an interaction within a biological pathway and on specific post-translational modifications.

BIND: Biomolecular Interaction Network Database

The BIND database (Bader *et al*., 2001) stores full descriptions of interactions, molecular complexes and pathways, among which are PPIs. BIND presents protein interactions from the molecular level to the pathway level and can be used, for instance, to study networks of interactions.

MINT: a Molecular INTeractions database

The MINT database (Zanzoni, 2002) stores functional interactions between biological molecules, i.e. proteins, RNA and DNA. Of interest to the protein specialist, MINT now focuses on experimentally verified protein–protein interactions. The database consists of entries extracted from the scientific literature where interaction information is found.

Protein Information in Other Types of Database

Many databases are not directly dedicated strictly to the world of proteins; however, important and varied information on proteins can be found within a great number of them, among which are the genomic databases and metabolic databases.

Genomic and genetic variation databases

Genomic databases offer a very wide scope of data resources that refer to a specific organism. The aim of a genomic database is to provide a maximum of

information on the genetic organization of a given species. Information includes gene names, gene localization (i.e. the position on a chromosome) as well as numerous cross-references to nucleotide and protein sequence databases. A number of these databases hold information particularly useful for proteome studies by describing specific gene mutations and their effect on an organism's phenotype.

OMIM: Online Mendelian Inheritance in Man (see Web Links)

OMIM (McKusick, 1998) is a collection of human genes and genetic disorders maintained by the McKusick–Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and the National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). The database offers a wealth of textual information provided in each entry, some of which can be useful in the context of protein studies.

HGVS: Human Genome Variation Society (see Web Links)

Formerly the HUGO Mutation Database Initiative, the Human Genome Variation Society (HGVS, Auerbach, 2000) was created to promote the discovery and free publication of information on the variations in human genes by fostering a central repository for such variations. There are hundreds of genomic databases or databases that revolve around specific mutations. As for all types of database, each research group tends to create its own database and its own nomenclature system. The various databases are listed by category, their contents briefly described and their URL address given.

HGMD: Human Gene Mutation Database (see Web Links)

HGMD (Krawczak and Cooper, 1997) is a comprehensive database of gene lesions underlying human inherited disease. All submissions are linked to mutation sites that have been experimentally defined.

dbSNP: Single Nucleotide Polymorphism database (see Web Links)

Signal nucleotide polymorphisms are the most common genetic variations. The SNP database (Sherry *et al*., 2001) is a repository of all the genetic variations that are being discovered as the human genome is being deciphered. Hence, a large number of artificial variations still have to be confirmed. Submissions to the database may be done directly, so information that has not yet been published in a peer-reviewed medium can be found here.

Metabolic and enzyme nomenclature databases

Metabolic databases are particularly heterogeneous data resources whose aim is to be comprehensive in the description of enzymes, biochemical reactions and metabolic pathways. Such resources can prove to be particularly helpful in the light of proteome research. A number of these databases provide detailed descriptions of all known enzymatic reactions catalyzed by a specific organism, while others tend to specialize in a subset of biochemical pathways expressed in a variety of organisms.

BRENDA: A Comprehensive Enzyme Information System

BRENDA (Schomburg *et al*., 2002) is the main collection of enzyme functional data available to the scientific community. It is maintained and developed at the Institute of Biochemistry at the University of Cologne.

KEGG: Kyoto Encyclopedia of Genes and Genomes (see Web Links)

KEGG (Kanehisa *et al*., 2002) strives to computerize the current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting molecules or genes. It consists of four types of data: pathway maps, molecule tables, gene tables and genome maps. By its scope, KEGG is in fact more than just a metabolic database.

ENZYME: an enzyme nomenclature database (see Web Links)

The ENZYME database (Bairoch, 2000) is not a metabolic database in the strictest sense. It is in fact a repository of information relative to the nomenclature of enzymes and is based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB). Data include the recommended name of a characterized enzyme, its alternative names (if any), its catalytic activity, cofactors (if any) as well as cross-references to a number of databases.

Conclusions

It is a particularly difficult task to give a comprehensive list of the databases that exist in the field of proteomics. Furthermore, what is valuable within one database for one scientist may be of no value for another. A list of the major databases, hence perhaps the most frequently updated, used and relevant to the

field of proteomics, is given here. One must keep in mind that the essence of databases is to grow and develop constantly, and the only way to get a good idea of the worth of one or the other is to visit a given database and browse through it.

See also

Genetic Databases
Genome Databases
Protein Sequence Databases

References

Attwood TK, Blythe M, Flower DR, *et al*. (2002) PRINTS and PRINTS-S shed light on protein ancestry. *Nucleic Acids Research* **30**: 239–241.

Auerbach AD (2000) Eighth International HUGO-Mutation Database Initiative Meeting, April 9, Vancouver, Canada. *Human Mutation* **16**: 265–268.

Bader GD, Donaldson I, Wolting C, *et al*. (2001) BIND – The Biomolecular Interaction Network Database. *Nucleic Acids Research* **29**: 242–245.

Bairoch A (2000) The ENZYME database in 2000. *Nucleic Acids Research* **28**: 304–305.

Bateman A, Birney E, Cerruti L, *et al*. (2002) The Pfam Protein Families Database. *Nucleic Acids Research* **30**: 276–280.

Berman HM, Westbrook J, Feng Z, *et al*. (2000) The Protein Data Bank. *Nucleic Acids Research* **28**: 235–242.

Boeckmann B, Bairoch A, Apweiler R, *et al*. (2003) The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Research* **31**(1): 365–370.

Cooper CA, Harrison MJ, Wilkins MR and Packer NH (2001) GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Research* **29**: 332–335.

Corpet F, Servant F, Gouzy J and Kahn D (2000) ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research* **28**: 267–269.

Falquet L, Pagni M, Bucher P, *et al*. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Research* **30**: 235–238.

Haft DH, Loftus BJ, Richardson DL, *et al*. (2001) TIGRFAMSs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research* **29**: 41–43.

Hoogland C, Sanchez J-C, Tonella L, *et al*. (2000) The 1999 SWISS-2DPAGE database update. *Nucleic Acids Research* **28**: 286–288.

Hoogland C, Sanchez J-C, Walther D, *et al*. (1999) Two-dimensional electrophoresis resources available from ExPASy. *Electrophoresis* **20**: 3568–3571.

Horn F, Weare J, Beukers MW, *et al*. (1998) GPCRDB: an information system for G protein-coupled receptors. *Nucleic Acids Research* **26**: 277–281.

Kanehisa M, Goto S, Kawashima S and Nakaya A (2002) The KEGG databases at GenomeNet. *Nucleic Acids Research* **30**: 42–46.

Krawczak M and Cooper DN (1997) The Human Gene Mutation Database. *Trends in Genetics* **13**: 121–122.

Lefranc M-P (2001) IMGT, the international ImMunoGeneTics database. *Nucleic Acids Research* **29**: 207–209.

Letunic I, Goodstadt L, Dickens NJ, *et al*. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Research* **30**: 242–244.

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

Mulder NJ, Apweiler R, Attwood TK, *et al*. (2002) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Briefings in Bioinformatics* **3**: 225–235.

Murzin AG, Brenner SE, Hubbard T and Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology* **247**: 536–540.

Orengo CA, Michie AD, Jones S, *et al*. (1997) CATH – a hierarchic classification of protein domain structures. *Structure* **5**: 1093–1108.

Rawlings ND, O'Brien EA and Barrett AJ (2002) MEROPS: the protease database. *Nucleic Acids Research* **30**: 343–346.

Schomburg I, Chang A, Hofmann O, *et al*. (2002) BRENDA, a resource for enzyme data and metabolic information. *Trends in Biochemical Sciences* **27**: 54–56.

Sherry ST, Ward MH, Kholodov M, *et al*. (2002) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Research* **29**: 308–311.

Xenarios I, Salwinski L, Duan XJ, *et al*. (2002) DIP: the Database of Interacting Proteins. A research tool for studying cellular networks of protein interactions. *Nucleic Acids Research* **30**: 303–305.

Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M and Cesareni G (2002) MINT: a Molecular INTeraction database. *FEBS Letters* **513**(1): 135–140.

Web Links

The ExPASy list of Biomolecular servers. This site lists the major (over one thousand!) databases of interest relative to proteomic research
http://www.expasy.org/alinks.html

BIND.The Biomolecular Interaction Network Database stores full descriptions of interactions, molecular complexes and pathways, among which are protein–protein interactions
http://bind.mshri.on.ca/

BRENDA. The main collection of enzyme functional data available to the scientific community, maintained and developed at the Institute of Biochemistry at the University of Cologne
http://www.brenda.uni-koeln.de/

CATH. A hierarchical domain classification of protein structures derived from PDB (see below)
http://www.biochem.ucl.ac.uk/bsm/cath_new/index.html

DbSNP. The Single Nucleotide Polymorphism database is a repository of all the genetic variations which are discovered as the human genome is being deciphered
http://www.ncbi.nlm.nih.gov/SNP

DIP.The Database of Interacting Proteins is curated both manually by expert curators and automatically. The DIP database provides a comprehensive and integrated tool for browsing and extracting information on protein–protein interactions
http://dip.doe-mbi.ucla.edu/

ENZYME. An enzyme nomenclature database based on the recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (IUBMB)
http://us.expasy.org/enzyme/

GlycoSuiteDB. A database of glycoprotein glycan structures derived from the scientific literature. Regarding proteins, when the glycan structures are known to be attached to a specific protein, direct links are made to Swiss-Prot and TrEMBL databases (see below)
http://www.glycosuite.com/

GPCRDB. An information system, which collects and disseminates data related to the G-protein-coupled receptor
http://www.gpcr.org/7tm

HGMD. The Human Gene Mutation Database is a comprehensive database of gene lesions underlying human inherited disease
http://www.hgmd.org/

HGVS. The Human Genome Variation Society was created to promote the discovery and free publication of information on the variations in human genes by fostering a central repository for such variations
http://www.hgvs.org/

IMGT. The international ImMunoGeneTics database is a high-quality integrated information system that specializes in immunoglobulins, T cell receptors and major histocompatibility complex molecules of all vertebrate species
http://imgt.cines.fr:8104/

InterPro. An integrated documentation resource for protein families, domains and functional sites, which was developed to rationalize the complementary efforts of the individual protein signature database projects that form the InterPro core (see Table 1)
http://www.ebi.ac.uk/interpro/

KEGG.The Kyoto Encyclopedia of Genes and Genomes strives to computerize the current knowledge of molecular and cellular biology in terms of the information pathways that consist of interacting molecules or genes
http://www.genome.ad.jp/kegg/

MEROPS. Provides data on individual proteases, protease families and also clans into which the families are grouped
http://merops.iapc.bbsrc.ac.uk/

MINT. The Molecular Interactions database. Stores functional interactions between biological molecules.
http://cbm.bio.uni.oma2.it/mint/

OMIM. The Online Mendelian Inheritance in Man database is a collection of human genes and genetic disorders maintained by the McKusick–Nathans Institute for Genetic Medicine, Johns Hopkins University (Baltimore, MD) and the National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD). The database offers a wealth of textual information provided in each entry, some of which can be useful in the context of protein studies
http://www.ncbi.nlm.nih.gov/omim/

PDB. The Protein Data Bank is a collection of 3D structures of proteins, nucleic acids and other biological macromolecules. PDB is a resource of critical importance in the discovery of new pharmacological agents, new catalysts, new biomaterials and possibly nanodevices
http://www.rcsb.org/pdb/

SCOP. Provides a detailed and comprehensive description of the structural and evolutionary relationships between proteins whose structure is known
http://scop.mrc-lmb.cam.ac.uk/scop/

Swiss-2Dpage. Contains 2D-PAGE and SDS PAGE reference maps and information on identified proteins from a variety of human biological samples. It is maintained collaboratively by the Central Clinical Chemistry Laboratory of the Geneva University Hospital and the Swiss Institute of Bioinformatics
http://www.expasy.org/ch2d/

Swiss-Prot. A non-redundant curated protein knowledge resource that provides a high level of annotation. Besides the stark protein sequence, a Swiss-Prot entry offers the description of the function of a protein, its domain structure, posttranslational modifications, variants and links to other databases
http://www.expasy.org/sprot

TrEMBL. Consists of computer-annotated entries in Swiss-Prot format derived from the translation of coding sequences in the European Molecular Biology Laboratory nucleotide sequence database. Hence TrEMBL entries are preliminary Swiss-Prot entries that have not yet been manually annotated
http://www.ebi.ac.uk/swissprot

World-2Dpage. A complete index of 2D-PAGE databases and services
http://www.expasy.org/ch2d/2d-index.html