26. Biological Databases

Mario Cannataro, Pietro H. Guzzi, Giuseppe Tradigo, Pierangelo Veltri

Biological databases constitute the data layer of molecular biology and bioinformatics and are becoming a central component of some emerging fields such as clinical bioinformatics, and translational and personalized medicine. The building of biological databases has been conducted either considering the different representations of molecular entities, such as sequences and structures, or more recently by taking into account high-throughput platforms used to investigate cells and organisms, such as microarray and mass spectrometry technologies. This chapter provides an overview of the main biological databases currently available and underlines open problems and future trends.

This chapter reports on examples of existing biological databases with information about their use and application for the life sciences. We cover examples in the areas of sequence, interactomics, and proteomics databases. In particular, Sect. 26.1 discusses sequence databases, Sect. 26.2 presents structure databases including protein contact maps, Sect. 26.3 introduces a novel class of databases representing the interactions among proteins, Sect. 26.4 describes proteomics databases, an area of biological databases that is being continuously enriched by proteomics experiments, and finally Sect. 26.5 concludes the chapter by underlining future developments and the evolution of biological databases.

In recent years, the availability of high-performance computational platforms and communication networks has enabled the use of algorithms to support biological studies. Bioinformatics applications have allowed the study and evaluation of biological experiments in the areas of proteomics and genomics [26.1]. Indeed, thanks to such high-performance computational platforms, bioinformatics is supporting the pipelines of experiments: from data preparation, to data preprocessing and manipulation, and results extraction. Simulations of biological systems have also used computational platforms. Moreover, the increasing use of computational platforms is related to the necessity to design, create, and publish databases of biological data and information. Biological sequences as protein amino acids, microarray experimental results, spatial and geometrical molecule conformations, experimental configurations, spectral data resulting from mass spectrometry of tissues, and ontologies for classifying data and experiments are only limited examples of the data and information that biologists may found interesting to query, navigate, and use in their experiments.

Biological databases can be seen as large and available (web-based) biological information sources for experiments and data analysis tools. They can be classified depending on the information stored and their availability, modifiability, etc. We classify databases based on information type and based on their use for bioinformatics applications as well as by biologists. We consider sequence, structure, interactomics, proteomics, and genomics databases. Sequence databases contain information in the form of alphanumeric sequences representing proteins and genes, whereas structure databases contain information about complex molecules and their structures (for instance, protein structures) as well as their functionalities or potential interactions. Interactomics databases can be considered as recent arrivals, storing the results of complex molecular interactions (for instance, protein–protein interactions).

The available databases have been enriched during recent decades as new experimental results have become available and been approved by the international community. This is the case, for instance, of the Protein Data Bank (PDB) database that includes protein structures and information that are published as soon as they are validated. Today most of these databases are available online and publish their own application programming interfaces (APIs) as well as extensible markup language (XML)-based exchange formalism to maximize interoperability among different data sources and to maximize their use. Finally, databases are also created by single research groups, then quickly made available and published to the international research community, while the development and validation of large databases are carried out by the international community. Thus, biological databases can be seen as large libraries related to biology and life sciences data and information resulting from experiments conducted in laboratories worldwide as well as from simulation analysis and published papers. They contain information from research areas related to biology, chemistry, and life sciences, and most of them specialize in particular life science areas (e.g., proteomics or genomics) as well as in research analysis and simulation (microarray gene expression databases).

26.1 Sequence Databases

These databases store information about the primary sequence of proteins. Each sequence is generally annotated by several types of information, e.g., the name of the scientist who discovered the sequence, or about the posttranslational modification. The user can query these databases by using a protein identifier or a fragment of sequence in order to retrieve the most similar proteins.

26.1.1 The EMBL Nucleotide Sequence Database

The European Molecular Biology Laboratory (EMBL) nucleotide sequence database [26.2, 3], maintained at the European Bioinformatics Institute (EBI), collects nucleotide sequences and annotations from publicly available sources. The database is involved in an international collaboration, being synchronized with the DNA Data Bank of Japan (DDBJ) and GenBank (USA) (see next sections). The core data are the protein and nucleotide sequences. The annotations describe the following items:

1. Function(s) of the protein
2. Posttranslational modification(s)
3. Domains and sites
4. Disease(s) associated with deficiencies
5. Secondary structure.

Webin is the preferred tool for individual submissions of nucleotide sequences, including third-party annotations (TPAs) and alignments. Automated procedures are provided for submissions from large-scale sequencing projects and data from the European Patent Office. New and updated data records are distributed daily, and the whole EMBL nucleotide sequence database is released four times a year. Currently, it contains more than 180 000 000 000 nucleotides in more than 100 000 000 entries. The entries are structured as textual files, composed of lines. Different types of lines, each with their own format, are used to record the various data that make up the entry. Each line begins with a two-character line code, which indicates the type of data contained in the line; for instance, the code ID represent the identifiers of the protein, while the code AC represents the accession number. Data are accessible via file transfer protocol (ftp) and several web interfaces. Moreover, the web-based sequence retrieval system (SRS) links nucleotide data to other specialized databases maintained at the EBI. Many other tools for sequence comparison and sequence similarity search, such as FASTA or the basic local alignment search tool (BLAST), are available through web interfaces.

26.1.2 GenBank

The GenBank database [26.4, 5] stores information about nucleotide sequences, maintained by the National Center of Biotechnology Information (NCBI).

GenBank entries are structured as flat files (like the EMBL database) and share the same structure as EMBL and DDBJ. All entries are grouped following both taxonomic and biochemical criteria. In this way it is possible to distinguish entries belonging to taxonomic groups (e.g., bacteria, viruses, and primates) as well as expressed sequences tags (EST) or core nucleotide sequences. Compared with the EMBL flat file structure, the main difference in the GenBank flat files is in the abbreviations used in the structure and in their interpretation.

GenBank is accessible through a web interface. Through the ENTREZ system, the entries of GenBank are integrated with many data sources, enabling searching for information about proteins and their structures, as well as literature about the functions of genes.

26.1.3 Uniprot Database

After the introduction of the nucleotide sequence databases, the scientific community worked towards the development of databases storing information about protein sequences. This work was helped by the introduction of methods and technologies able to investigate proteins, leading to the accumulation of such data. Historically, one of the first databases was Swiss-Prot [26.3]. The main characteristics of Swiss-Prot were:

1. The use of flat files organized in multiple subfields as the storage system
2. Avoidance of redundancies, achieved by employing a manual curation workflow
3. Extensive usage of annotation, providing useful information about each entry.

In 2002 the Swiss-Prot databases merged with two related projects: the Tr-EMBL database [26.3] (a supplement of Swiss-Prot that stores information about sequences that are candidates to be introduced into Swiss-Prot but are under verification) and the Protein Sequence Database that developed into the Protein Information Resource project (PSD–PIR [26.6]). The result of this process was the introduction of the UniProt [26.7] consortium, which is structured on three main knowledge bases:

1. UniProt (also referred to as the UniProt knowledge base), which is the main archive storing information about protein sequences and annotations extracted from Swiss-Prot, TrEMBL, and PSD-PIR
2. UniParc (Uniprot archive), which contains publicly available information about proteins extracted from the main archives
3. UniRef (Uniprot reference), a set of databases that organize entries of UniProt by their similarity sequence; e.g., UniRef90 groups entries of UniProt that present at least 90% sequence similarity into a single record.

The Uniprot database is publicly accessible at http://www.uniprot.org/, and it is freely downloadable.

26.2 Structure Databases

26.2.1 Protein Data Bank (PDB)

The Protein Data Bank (PDB) [26.8] is a worldwide archive of structure data of biological macromolecules. Such data are generated by crystallography and nuclear magnetic resonance (NMR) experiments. Each PDB entry is stored in a single flat file. There is an underlying ontology of about 1700 terms that define the macromolecular structure and crystallographic experiment. This ontology is called the macromolecular crystallographic information file (mmCIF) dictionary. Although distributed as a flat file, PDB is based on a relational model. There are three distinct query interfaces:

1. Status Query
2. Search Lite
3. Search Field.

Search Lite has a single text field in which it is possible to write keywords. Search Field is a customizable query form allowing queries based on author citation, sequences (via FASTA algorithm), dates, and chemical formulas. Many interfaces present information related to the results. A query result browser interface allows detailed information to be browsed and a set of files storing the structures found to be downloaded. PDB files in XML format are currently being tested. Data are acquired from the research community by submission.

26.2.2 Databases of Structural Classifications

A structural domain of a protein is an element of its ternary structure that often folds independently of the rest of the protein chain and represents a biologically relevant module of the protein itself. Despite the large number of different proteins expressed in eukaryotic systems, there are many fewer different domains, structural motifs, and folds. Many domains are not unique to a protein that is produced by one gene or gene family but instead appear in a variety of proteins as a consequence of evolution, which has conserved spatial conformation better than primary sequence.

Consequently, several methods have been developed for structural classification of proteins, and a number of different databases have been introduced, such as the Structural Classification of Proteins (SCOP) [26.9] and CATH (class, architecture, topology, and homologous superfamily) databases [26.10].

SCOP

The Structural Classification of Proteins (SCOP, http://scop.mrc-lmb.cam.ac.uk/scop) database aims to order all the proteins whose structure has been published according to their structural domains. Protein domains in SCOP are hierarchically classified into *families, superfamilies, folds, and classes*. Firstly, proteins are grouped together into families on the basis of evolutionary or functional similarities, such as sequence alignment. Then, proteins whose sequences have low similarity but whose structure or functions are close are grouped into superfamilies. The secondary structure of the proteins in these two groups is analyzed, and when proteins in two groups, e.g., two families, have similar secondary structure, they have a common fold. Finally, folds are grouped into classes as follows:

1. All α, if the structure is essentially formed by α-helices
2. All β, if the structure is essentially formed by β-sheets
3. α/β, for those with α-helices and β-strands
4. α + β, for those in which α-helices and β-strands are largely segregated
5. Multidomain, for those with domains of different classes.

The SCOP database is available on the world wide web. The user has many options to browse its content. The main possibility is to start at the top of the hierarchy and then navigate through the levels from the root to the leaves, which are individual PDB entries. Alternatively, the user can search a protein starting from an amino acid sequence to retrieve the most similar proteins categorized in SCOP. The user can then download the found protein as a single PDB file.

CATH

CATH [26.10] stores a hierarchical classification of PDB structures obtained by NMR where crystal structures are solved at resolution higher than 4.0 Å (http://www.cathdb.info/latest/index.html). Protein structures are classified using a combination of automated and manual procedures on the protein domains. To divide multidomain protein structures into their constituent domains, a combination of automatic and manual techniques are used. The hierarchy is organize in four major levels: *class*, *architecture*, *topology* (fold family), and *homologous superfamily*. Class is determined according to the secondary structure composition; currently three major classes have been recognized: mainly α, mainly β, and α–β, which includes both alternating α/β structures and α + β structures. A fourth class contains protein domains which have low secondary structure content. The architecture level considers the overall shape of the domain structure as determined by the orientations of the secondary structures, being assigned manually. The topology level groups the structures depending on the overall shape and on the connectivity of the secondary structure by applying the SSAP (sequential structure alignment program for protein structure comparison) algorithm [26.11]. Finally, the homologous superfamily level groups protein domains which are thought to share a common ancestor and can therefore be described as homologous as recognized by SSAP. Currently, it contains 30 028 PDB structures. CATH contains PDB structures organized in a relational model. CATH can be searched by submitting a protein identifier, or by browsing the hierarchical structure. Moreover, the user can access data via ftp and download them.

26.2.3 Protein Contact Map

Protein contact maps are bidimensional data structures representing a view of the three-dimensional (3-D) structure of a protein. They are used to store the presence or absence of contacts among protein residue pairs. Two residues are said to be in contact if their mutual distance is lower than a certain distance threshold (i. e., 8 Å). Contact maps have a key role in most state-of-the-art protein structure prediction pipelines, i. e., the prediction of the three-dimensional space conformation of the amino acids composing a protein.

Indeed, contact information may be used to drive the computational folding process, to select structural templates, or to assess the quality of structural predictions. Thus, it is critical to develop accurate predictors of contact maps. Correct contact maps have been shown to lead to reasonably good 3-D structures [26.12, 13], and predicted contact maps have been used for driving protein folding in the ab initio case (that is, when a protein is folded without relying on homology to another protein of known structure), for selecting and ranking folded protein models, and for predicting folding times, protein domain boundaries, secondary structures, etc. Virtually no contact map databases exist (except for [26.14]), but many prediction tools are available [26.15–17]. In fact, in the case of an unknown protein for which no exact 3-D structure has been experimentally determined, a contact map can be directly predicted from its primary structure or derived from its predicted 3-D structure, while in the case of a known protein, the 3-D structure retrieved from a structure database (i. e., Protein Data Bank) can easily give the protein contact map.

26.3 Interactomics Databases

The accumulation of protein interaction data led to the introduction of several databases. Here we concentrate on *databases of experimentally determined interactions*, which include all databases storing interactions extracted from both literature and high-throughput experiments, and databases of *predicted interactions* that store data obtained by in silico predictions. Another important class that we report is constituted by *integrated databases* or metadatabases, i. e., databases that aim to integrate data stored in other publicly available datasets. Currently, there exist many databases that differ in various biological and information science aspects: the organism covered, the kind of interactions, the kind of interface, the query language, the file format, and the visualization of results.

Data produced in low- or high-throughput experiments are stored in *databases of experimentally determined interactions* after subsequent verification by a committee of database curators. Researchers can submit their own data directly to the databases, e.g., to Intact [26.18], or they can publish data in the literature and then the database curators will extract them, e.g., the MINT (Molecular INTeraction) database [26.19]. For a more complete description of interactomics databases see [26.20, 21].

For simpler organisms, such as yeasts, worms, or flies, the process of the whole coverage of the interaction network seems to be almost completed. This process led to the introduction of a huge amount of data that may be mined for many objectives. Conversely, the complexity of the interactomes of higher eukaryotes has prevented these experiments for humans. In this scenario, the need for the introduction of algorithms and tools able to use these experimental data to predict protein interactions arose. Thus, starting from existing databases of verified interactions, a number of algorithms have been developed to predict putative interactions that are accumulated into *databases of predicted interactions*. The common approach is based on the consideration that the interaction mechanisms are conserved through evolution; i. e., if two protein A and B interact in a simple organism, then the corresponding orthologous proteins A_1 and B_1 may interact in a complex organism. Thus, starting from the interacting proteins in a simple organism, predictions are made for other organisms.

Despite the existence of many databases, the resulting data present three main problems [26.22]: the low overlap among databases, the resulting lack of completeness with respect to the real interactome, and the absence of integration. Consequently, in order to perform an exhaustive data collection (e.g., for an experiment), researchers have to manually query different data sources. This problem is being addressed through the introduction of databases based on the integration of existing ones. Nevertheless, in the interactomics field, the integration of existing databases is not an easy task.

The integration of data from different laboratories and sources can be done through the adoption of an accepted system of interaction identifiers. It should be noted that, while in other biological database systems, such as sequence databases, there exists a common system of identifiers, and cross-referencing can be used to retrieve the same biological entity from different databases, PPI (protein-to-protein interaction) interactions are currently not identified by a unique identifier, but through the names of the corresponding partners.

Despite these problems, different approaches for data integration and building larger interaction maps have been pro posed. The rationale for these approaches is based on a three-step process:

1. Collection of data from different data sources
2. Transformation of data into a common model
3. Annotation and scoring of the resulting dataset.

All the existing databases go beyond storage of the interactions, also integrating them with functional annotations, sequence information, and references to corresponding genes. Finally, they generally provide some visualization that presents a subset of interactions in a comprehensive graph.

Nevertheless, currently there are some problems and characteristics that are common to almost all databases:

1. Errors in the databases
2. Lack of naming standards
3. Little overlap among interactions.

Any published dataset may contain errors, so any database may contain false interactions, often called false positives, i. e., proteins erroneously reported as interacting. This may be due, for instance, to technical (i. e., false positives that are due to the detection method) and biological problems (i. e., proteins that are reported to be interacting in vitro but that are never co-located).

In other biological database communities, such as those storing protein sequences or structures, there exist many projects providing common accepted identifiers for biological objects, or at least a system for cross-referencing the same object between almost all databases. In interactomics there is no such common identifier, and in general interactions are not identified by a single code but rather by using the identifiers of the interacting proteins.

It has been noted [26.22] that existing databases present little overlap with respect to the dimension of the interactomes. Despite this, integration of databases remains an open problem due to the difficulties resulting from the absence of a naming standard.

Conversely, common aspects of existing datasets are:

1. Simple web-based interface for querying
2. Simple visualization of results in both tabular and graphical form
3. Data available for download in different formats.

It should be noted that almost all these databases offer the user the possibility of retrieving data and some annotations through a simple web-based interface. Despite this, querying of protein networks aims to go beyond the simple retrieval of a set of interactions stored in databases.

Databases can actually be queried through simple key-based searches, e.g., by inserting one or more protein identifiers. The output of such a query is in general a list of interacting protein pairs. These pairs share a protein, as specified in the query. Such an approach, despite its conceptual simplicity and easy practical use, presents some limitations. Let us consider, for instance, a researcher who wishes to compare patterns of interactions among species, or a researcher who wants to search for interactions related to a given biological compartment or biological process. The existing query interfaces, in general, do not enable such queries.

Thus, a more powerful querying system should provide a semantically more expressive language, e.g., enabling retrieval of all interaction patterns that share the same structure. Then, the query system should map the query, expressed in a high-level language (e.g., using a graph formalism), into suitable graph structures and search for them by applying appropriate algorithms. Unfortunately this problem is not easy from a computational point of view, and it requires:

1. Modeling of the PPI network in a suitable data structure
2. Appropriate algorithms for mapping, i. e., identification of the correspondence between the nodes in a subnetwork and those stored in the database [26.23].

26.4 Proteomics Databases

26.4.1 Global Proteome Machine Database

The Global Proteome Machine Database (http://www.thegpm.org/GPMDB/index.html) [26.24] was constructed to utilize information obtained from the different servers included in the Global Proteome Machine project (GPM), to validate peptide tandem mass spectrometry (MS/MS) spectra and protein coverage. GPM is a system for analyzing, storing, and validating proteomics information derived from tandem mass spectrometry. The system is based on a relational database on different servers for data analysis, and on a user-friendly interface to retrieve and analyze data. This database has been integrated into GPM server pages. The gpmDB data model is based on a modification of the Hupo-PSI minimum information about a proteomics experiment (MIAPE) [26.25] scheme. With respect to the proposed standard, the database is conceived to hold only the information needed in certain bioinformatics-related tasks, such as sequence assignment validation. Data are mainly held in a set of XML files: the database serves as an index to those files. This combination of a relational database with XML is called by the authors XML information about a proteomics experiment (XIAPE). The system is available both through a web interface and as a standalone application, allowing users to compare their experimental results with others previously observed by other scientists.

26.4.2 PeptideAtlas

PeptideAtlas [26.26] is a database that aims to annotate the human genome with protein-level information (http://www.peptideatlas.org/overview.php). It contains data coming from identified peptides analyzed by liquid chromatography tandem mass spectrometry (LC-MS/MS) and thus mapped onto the genome. PeptideAtlas is not a simple repository for mass spectrometry experiments, but uses spectra as a primary information source to annotate the genome, combining different information. Consequently, population of this database involves two main phases:

1. A proteomic phase in which samples are analyzed through LC-MS/MS, and resulting spectra are mined to identify the contained peptides
2. An in silico phase in which peptides are processed by applying a bioinformatic pipeline and each peptide is used to annotate a genome. Resulting derived data, both genomics and proteomics, are stored in the PeptideAtlas database.

Data submitted by researchers are organized in a relational model. PeptideAtlas is based on the Systems Biology Experiment Analysis Management System (SBEAMS) project, which is a framework for collecting, storing, and accessing data produced by a variety of different experiments. It combines a relational database management system back-end providing integrated access to remote data sources. User can query data through a web interface or can download a whole dataset, organized by the original publications.

26.4.3 NCI Repository

The National Cancer Institute Clinical Proteomics Databank (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp), is not a database, but stores different datasets obtained by mass spectrometry. Currently it holds datasets obtained in different experimental conditions, coming from different mass spectrometry platforms, for both human and animals. It contains six different surface-enhanced laser desorption ionization time-of-flight (SELDI-TOF) mass spectrometry datasets. This technique is very similar to MALDI-TOF and generates spectra which have similar characteristics. Datasets are stored as flat files, each containing a whole SELDI-TOF spectrum. The datasets are freely available to download as flat files.

26.4.4 PRIDE

The proteomics identifications database (PRIDE) (http://www.ebi.ac.uk/pride) [26.27] is a database of protein and peptide identifications that may be annotated with supporting mass spectra. PRIDE stores information about a complete proteomic experiment, starting from the title and a brief description of the experiment itself. Each experiment is annotated by a description of the sample under analysis and of the instrumentation used to perform the analysis. The core element of each entry is the protein identifications, sorted by unique accession numbers and supported by a corresponding list of one or more peptide identifications. For each peptide identified, the database also stores the sequence and coordinates of the peptide within the protein for which it provides evidence. Optionally, an entry can contain a reference to any submitted mass spectra that form the evidence for the peptide identification, encoded in the versatile proteomics standard initiative (PSI) mzData format. Users can directly submit protein and peptide identification data to be published in peer-reviewed publications by using the PRIDE 2.1 XML schema. The PRIDE database currently contains 3178 experiments, 339 696 identified proteins, 2 145 505 identified peptides, 309 889 unique peptides, and 2 582 614 spectra. PRIDE is based on a relational database based on structured query language (SQL) and is currently available for ORACLE and MySQL. PRIDE provides the user with a web interface to retrieve data. Data can be exported in PRIDE XML schema, a format which embeds mzData as a subelement, or using the mzData XML schema.

26.4.5 2-D Gel Databases

Databases of data produced by using gel electrophoresis generally store both images and identification, the core data, and metadata relating to the experiment. Metadata are relative to the parameters of the experiment, while core data store the proteins contained in the associated image. In the following we present the SWISS-2DPAGE database.

SWISS-2DPAGE

The SWISS-2DPAGE (http://www.expasy.ch/ch2d/ch2d-top.html) [26.28] database was established in 1993 and is maintained collaboratively by the Swiss Institute of Bioinformatics (SIB) and the Biomedical Proteomics Research Group (BPRG) of the Geneva University Hospital. Current content includes identified spots and corresponding protein entries in 36 reference maps from human, mouse, *Arabidopsis thaliana*, *Dictyostelium discoideum*, *Escherichia coli*, *Saccharomyces cerevisiae*, and *Staphylococcus aureus*.

The protein entries in SWISS-2DPAGE are structured text files. Each entry is composed of defined lines, used to record various kinds of data. For standardization purposes, the format of SWISS-2DPAGE entries is similar to that used in the Swiss-Prot database, in addition to specific lines dedicated to the two-dimensional (2-D) polyacrylamide gel electrophoresis (PAGE) data:

1. The master line (MT) lists the reference maps where the entry has been identified
2. The images line (IM) lists the 2-D PAGE images available for the entry
3. The 2-D lines group different topics including the mapping procedure, spot coordinates, protein amino acid composition, protein expression levels, and modifications.

SWISS-2DPAGE is available through the ExPASy molecular biology server. The SWISS-2DPAGE top page provides text searches, and displays results with active links to other databases. It is also possible to get a local copy of SWISS-2DPAGE via ftp from the ExPASy ftp server. On the ExPASy webserver the data image associated with a protein entry displays the experimental location of the protein on the chosen map, in addition to a theoretical region computed from the protein sequence.

26.5 Conclusions

Biological databases have been developed as autonomous, specialized, but not integrated repositories of data; For instance, the first databases stored data regarding the different representations of DNA and proteins, such as sequences and structures. Since molecular medicine research needs information from different biological databases, the bioinformatics community started to develop integrated databases where integration is often obtained by cross-referencing common identifiers. Following the expansion of omics sciences, such as genomics, proteomics, and interactomics, and the diffusion of high-throughput experimental platforms, novel biological databases have been produced. Among them, mass spectrometry and gel electrophoresis data have improved proteomics databases, while data about protein interactions form the main interactomics protein–protein interaction databases. A future trend will be the further integration of this plethora of biological databases and especially the annotation of existing data with information contained in different knowledge bases and ontologies such as Gene Ontology (http://www.geneontology.org/).

References

26.1 R. Matthiesen: Methods, algorithms and tools in computational proteomics: A practical point of view, Proteomics **7**(16), 2815–2832 (2007)

26.2 EMBL Nucleotide Sequence (European Molecular Biology Laboratory, EMBL Heidelberg, Heidelberg) available online at http://www.ebi.ac.uk/embl

26.3 B. Boeckmann, A. Bairoch, R. Apweiler, M.-C.C. Blatter, A. Estreicher, E. Gasteiger, M.J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout, M. Schneider: The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res. **31**(1), 365–370 (2003)

26.4 GenBank database (National Center for Biotechnology Information, National Library of Medicine, Bethesda) USA available online at www.ncbi.nlm.nih.gov/genbank/

26.5 D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, D.L. Wheeler: GenBank, Nucleic Acids Res. **36**, D25–D30 (2008)

26.6 W.C. Barker, J.S. Garavelli, P.B. Mcgarvey, C.R. Marzec, B.C. Orcutt, G.Y. Srinivasarao, L.S. Yeh, R.S. Ledley, H.W. Mewes, F. Pfeiffer, A. Tsugita, C. Wu: The PIR-international protein sequence database, Nucleic Acids Res. **27**(1), 39–43 (1999)

26.7 The UniProt Consortium: The Universal Protein Resource (UniProt) in 2010, Nucleic Acids Res. **38**(suppl 1), D142–D148 (2010)

26.8 H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne: The protein data bank, Nucleic Acids Res. **28**(1), 235–242 (2000)

26.9 T.J.P. Hubbard, A.G. Murzin, S.E. Brenner, C. Chothia: SCOP: A structural classification of proteins database, Nucleic Acids Res. **25**(1), 236–239 (1997)

26.10 C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton: CATH – a hierarchic classification of protein domain structures, Structure **5**(8), 1093–1108 (1997)

26.11 C. Orengo, W. Taylor: SSAP: Sequential structure alignment program for protein structure comparison. In: *Computer Methods for Macromolecular Sequence Analysis*, Methods in Enzymology, Vol. 266, ed. by S.P. Colowick, R.F. Doolittle, N.O. Kaplan (Academic, New York 1996) pp. 617–635

26.12 M. Vendruscolo, E. Kussell, E. Domany: Recovery of protein structure from contact maps, Fold. Des. **2**(5), 295–306 (1997)

26.13 I. Walsh, D. Baú, A.J.M. Martin, C. Mooney, A. Vullo, G. Pollastri: Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks, BMC Struct. Biol. **9**(1), 5 (2009)

26.14 P. Chen, C. Liu, L. Burge, M. Mohammad, B. Southerland, C. Gloster, B. Wang: IRCDB: A database of inter-residues contacts in protein chains, 1st Int. Conf. Adv. Databases (2009) pp. 1–6

26.15 D. Baú, A. Martin, C. Mooney, A. Vullo, I. Walsh, G. Pollastri: Distill: A suite of web servers for the prediction of one-, two- and three-dimensional structural features of proteins, BMC Bioinformatics **7**, 1–8 (2006)

26.16 R.M. MacCallum: Striped sheets and protein contact prediction, Bioinformatics **20**(suppl 1), i224–i231 (2004)

26.17 B. Rost, M. Punta: PROFcon: Novel prediction of long-range contacts, Bioinformatics **21**(9), 2960–2968 (2005)

26.18 H. Hermjakob, L. Montecchi-Palazzi, C. Lewington, S. Mudali, S. Kerrien, S. Orchard, M. Vingron, B. Roechert, P. Roepstorff, A. Valencia, H. Margalit, J. Armstrong, A. Bairoch, G. Cesareni, D. Sherman, R. Apweiler: IntAct: An open source molecular interaction database, Nucleic Acids Res. **1**(32), 452–455 (2004)

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

26.20 M. Cannataro, P.H. Guzzi, P. Veltri: Protein-to-protein interactions: Technologies, databases, and algorithms, ACM Comput. Surv. **43**, 1 (2010)

26.21 A. Batemen: NAR Database ISSUE, Nucleic Acids Res. **35**(Suppl. 1) (2007)

26.22 G. Chaurasia, Y. Iqbal, C. Hanig, H. Herzel, E.E. Wanker, M.E. Futschik: UniHI: An entry gate to the human protein interactome, Nucleic Acids Res. **35**(suppl1), D590–594 (2007)

26.23 S. Zhang, X.-S. Zhang, L. Chen: Biomolecular network querying: A promising approach in systems biology, BMC Syst. Biol. **2**(1), 5 (2008)

26.24 C. Robertson, J.P. Cortens, R.C. Beavis: Open source system for analyzing, validating, and storing protein identification data, J. Proteome Res. **3**(6), 1234–1242 (2004)

26.25 C.F. Taylor, H. Hermjakob, R.K. Julian, J.S. Garavelli, R. Aebersold, R. Apweiler: The work of the human proteome organisation's proteomics standards initiative (HUPO PSI), OMICS **10**(2), 145–151 (2006)

26.26 F. Desiere, E.W. Deutsch, N.L. King, A.I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S.N. Loevenich, R. Aebersold: The PeptideAtlas project, Nucleic Acids Res. **34**(Suppl. 1), D655–D658 (2006)

26.27 P. Jones, R.G. Côté, L. Martens, A.F. Quinn, C.F. Taylor, W. Derache, H. Hermjakob, R. Apweiler: PRIDE: A public repository of protein and peptide identifications for the proteomics community, Nucleic Acids Res. **34**(Suppl. 1), D659–D663 (2006)

26.28 J.-C. Sanchez, D. Chiappe, V. Converset, C. Hoogland, P.-A. Binz, S. Paesano, R.D. Appel, S. Wang, M. Sennitt, A. Nolan, M.A. Cawthorne, D.F. Hochstrasser: The mouse SWISS-2D PAGE database: A tool for proteomics study of diabetes and obesity, Proteomics **1**(1), 136–163 (2001)