SCIENCE CHINA
Life Sciences

• REVIEWS •

November 2011 Vol.54 No.11: 988–998
doi: 10.1007/s11427-011-4247-x

An overview of human protein databases and their application to functional proteomics in health and disease

ZHANG YanQiong[1,2], ZHU YunPing[2*] & HE FuChu[1,2*]

[1] *Institute of Basic Medical Sciences, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing 100730, China;*
[2] *State Key Laboratory of Proteomics, Beijing Proteome Research Center, Beijing Institute of Radiation Medicine, Beijing 102206, China*

Received March 13, 2011; accepted November 23, 2011

Functional proteomics can be defined as a strategy to couple proteomic information with biochemical and physiological analyses with the aim of understanding better the functions of proteins in normal and diseased organs. In recent years, a variety of publicly available bioinformatics databases have been developed to support protein-related information management and biological knowledge discovery. In addition to being used to annotate the proteome, these resources also offer the opportunity to develop global approaches to the study of the functional role of proteins both in health and disease. Here, we present a comprehensive review of the major human protein bioinformatics databases. We conclude this review by discussing a few examples that illustrate the importance of these databases in functional proteomics research.

human protein database, bioinformatics, functional proteomics, health, disease

Citation: Zhang Y Q, Zhu Y P, He F C. An overview of human protein databases and their application to functional proteomics in health and disease. Sci China Life Sci, 2011, 54: 988–998, doi: 10.1007/s11427-011-4247-x

The Human Genome Sequencing Project triggered a revolution in biology and medicine. Following this proteomics has developed to study the proteome, that is, the proteins expressed in a genome, a cell or a tissue. Proteomics is an alternative to existing analytical methods that are used to describe life in molecular terms. Proteome investigations have focused on two main areas: expression proteomics and functional proteomics. Functional proteomics in particular, can be defined as a tool that aims to couple proteomic information with biochemical and physiological analyses to understand the functions of proteins in normal and diseased organs [1]. The approaches of functional proteomics are focused mainly towards two major targets: the elucidation of the biological functions of unknown proteins and the definition of cellular mechanisms at the molecular level [2]. In cells, many proteins display their biological functions through a rapid and transient association within large protein complexes. Therefore, understanding protein functions as well as unraveling molecular mechanisms within the cell depends on the identification of the interacting protein partners [3]. Because most physiological and pathological processes are manifested at the protein level, biological scientists are becoming increasingly interested in applying functional proteomics strategies to achieve a better understanding of basic molecular biology and disease processes and to advance the discovery of novel diagnostic, prognostic and therapeutic targets for numerous diseases.

In recent years, a variety of publicly available bioinformatics databases have been developed to support protein-related information management and biological knowledge discovery. These databases provide and organize biological annotations for protein sequences, structures, functions and evolutionary analyses in the context of biological pathways, networks and systems. In addition to the

*Corresponding author (email: zhuyunping@gmail.com; hefc@nic.bmi.ac.cn) annotations, these databases offer a global approach to the study of the functional role of proteins both in health and disease. Here, we present a comprehensive review of the major human protein bioinformatics databases highlighting those that are recent, of high quality, publicly available, and that we judged to be of interest to researchers in functional proteomics. We also discuss the important roles that these databases can play in functional proteomics research.

1 An overview of human protein databases

Based on the topics and types of data that are stored, human protein databases can be classified primarily as sequence databases, structure databases, databases of protein-protein interactions and complexes, family databases and proteomics databases as described in Table 1.

1.1 Protein sequence databases

Protein sequence databases serve as repositories for collections of protein sequences. In addition to the protein sequence they also contain annotations that reflect the existing knowledge of the protein's function and the residues that contribute to that function. A reliable sequence can form the basis of investigations into the biological role of the protein. Therefore, protein sequence databases are the foundation for medical and functional studies.

1.1.1 Reference sequences (RefSeq)

The National Center for Biotechnology Information Reference Sequence database (NCBI RefSeq) [4] is a comprehensive resource for curated non-redundant sequences of genomic regions, transcripts and proteins. RefSeq is one of the best sources for reliable nucleotide and protein sequences. The RefSeq collection is derived from the sequence data available in the redundant archival database GenBank. RefSeq sequences are annotated and include coding regions, conserved domains, variations, references, names, and database cross-references. The annotation is performed using a combination of automated prediction and manual curation. The RefSeq data can be accessed from NCBI web sites by Entrez query, BLAST, and FTP download. RefSeq sometimes stores more than one protein sequence per gene. In such a case, it may be useful to align all available RefSeq protein sequences for the gene of interest to see where they differ and to assess whether or not substantial differences require further investigation.

RefSeq release 46 (11 March 2011) includes 12167392 proteins from 11734 organisms.

1.1.2 UniProtKB/Swiss-Prot

The UniProt Knowledgebase (UniProtKB) [5] is a comprehensive, high-quality and freely accessible resource of protein sequence and functional information. UniProtKB has two sections: one that contains manually-annotated records with information extracted from literature and curator-evaluated computational analysis (UniProtKB/Swiss-Prot), and another that holds computationally analyzed records that await full manual curation (UniProtKB/TrEMBL). Swiss-Prot was established in 1986 and has been maintained collaboratively since 1987 by Amos Bairoch and his group, first at the Department of Medical Biochemistry of the University of Geneva and now at the Swiss Institute of Bioinformatics (SIB) and the EMBL Data Library. The UniProtKB/Swiss-Prot protein sequence database contains entries composed of different line types, each with its own format. For consistency, the format of UniProtKB/ Swiss-Prot follows as closely as possible that of the EMBL Nucleotide Sequence Database. UniProtKB/Swiss-Prot entries include information that clearly describes the type of evidence available for the existence of a particular protein.

The UniProtKB/Swiss-Prot database is different from other protein sequence databases by four distinct criteria:

(i) Annotation: The annotation consists of the following features: function(s) of the protein; post-translational modification(s); domains and sites; secondary structure; quaternary structure; similarities to other proteins; disease(s) associated with deficiencies in the protein; sequence conflicts; variants; and others.

(ii) Minimal redundancy: UniProtKB/Swiss-Prot merges all relevant data so as to minimize the redundancy of the database. If conflicts exist between various sequencing reports, they are indicated in the feature table of the corresponding entry.

(iii) Integration with other databases: UniProtKB/ Swiss-Prot is currently cross-referenced with about 30 different databases. Cross-references are provided in the form of pointers to information related to a UniProtKB/ Swiss-Prot entry that is available in other data collections. The extensive network of cross-references makes UniProtKB/ Swiss-Prot a major focal point of biomolecular database interconnectivity.

(iv) Documentation: UniProtKB/Swiss-Prot is distributed with a large number of index files and specialized documentation files. Some of these files have existed for a long time; however, many have been created recently and new files are continuously being adding. The release notes always contain an up-to-date descriptive list of all distributed document files.

UniProtKB/Swiss-Prot release 2011_04 (05 April 2011) contains 526969 sequence entries, comprising 186402391 amino acids abstracted from 196878 references.

1.1.3 Database of Protein Disorder (DisProt)

The Database of Protein Disorder (DisProt) [6] is a curated database that provides sequence, structure and function information for intrinsically disordered proteins (IDPs) that lack a fixed 3D structure in their putatively native state, either in their entirety or in part. Although they lack a fixed

Table 1 Human protein bioinformatics databases

Category	Name	Content	URL	References
Protein sequence databases	Reference Sequence (RefSeq)	Containing many reliable protein sequences	http://www.ncbi.nlm.nih.gov/RefSeq/	[4]
	Entrez Protein Database	Collection of protein sequences from a variety of sources, and translations from annotated coding regions in GenBank and RefSeq	http://www.ncbi.nlm.nih.gov/sites/	[7]
	UniProtKB/Swiss-Prot	Containing the most reliable sequence and annotations	http://www.uniprot.org/	[5]
	UniProt Archive (UniParc)	Containing most of the publicly available protein sequences in the world	http://www.uniprot.org/help/uniparc	[8]
	UniProt Reference Clusters (UniRef)	Clustered sets of sequences from UniProt Knowledgebase and selected UniParc records	http://www.uniprot.org/help/uniref	[9]
	Consensus CDS protein set (CCDS)	Containing human and mouse protein sequences	http://www.ncbi.nlm.nih.gov/CCDS/	[10]
	Database of protein disorder (DisProt)	Experimentally verified disordered regions in proteins	http://www.disprot.org	[6]
	PhosphoSite	Extensive information on known phosphorylation sites	http://www.phosphosite.org	[11]
Protein structure databases	Molecular Modeling Database (MMDB)	Containing 3D structures of proteins and polynucleotides.	http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml	[12]
	Worldwide Protein Data Bank (wwPDB)	Containing the 3D structures of proteins, nucleic acids and large macromolecular complexes that have been determined using X-ray crystallography, NMR and electron microscopy techniques	http://www.wwpdb.org/	[13]
	ModBase	Containing 3D protein models calculated by comparative modeling	http://www.modbase.compbio.ucsf.edu/modbase-cgi/index.cgi	[14]
	SWISS-MODEL	Containing 3D models of proteins	http://www.swissmodel.expasy.org/repository/	[15]
	CATH	Hierarchical classification of protein domain structures in the Protein Data Bank	http://www.cathdb.info/	[16]
	Structural Classification Of Proteins (SCOP)	Description of the evolutionary and structural relationships of the proteins with known structures	http://www.scop.mrc-lmb.cam.ac.uk/scop/	[17]
	KineticDB	Containing the experimental data on protein folding kinetics	http://www.kineticdb.protres.ru/db/index.pl	[18]
	RESID	Containing the structures for protein pre-, co- and post-translational modifications	http://www.ebi.ac.uk/RESID/	[19]
	Phospho3D	Containing the 3D structures of phosphorylation sites that stores information retrieved from the phospho.ELM database	http://www.cbm.bio.uniroma2.it/phospho3d/	[20]
Databases of protein protein interactions and complexes	Database of Interacting Proteins (DIP)	Containing the binary protein–protein interactions that were manually curated by experts.	http://www.dip.doe-mbi.ucla.edu/dip/Main.cgi	[21]
	Molecular INTeraction database (MINT)	Containing the experimentally verified protein-protein interactions, which are mined from the scientific literature by expert curators	http://mint.bio.uniroma2.it/mint/	[22]
	IntAct	An open source database and software framework; Containing the interaction data that are manually extracted from public literature and annotated to a high level of detail through the extensive use of controlled vocabulary; Also containing a suite of tools that can be used to visualize and analyze the interaction data.	http://www.ebi.ac.uk/intact/main.xhtml	[23]

(*To be continued on the next page*)

(*Continued*)

Category	Name	Content	URL	References
	Human Protein Reference Database (HPRD)	Containing an extensive list of interaction data of human proteins, which are manually extracted from public literature and curated by a team of trained biologists.	http://www.hprd.org/	[24]
	STRING	Containing interaction data of human proteins from numerous sources, not only from experimental repositories, but also includes computational prediction methods, and automated text mining of public text collections such as PUBMED.	http://string-db.org/	[25]
	Unified Human Interactome (UniHI)	Containing human protein interaction data from various sources including both computational and experimental repositories.	http://www.unihi.org/	[26]
	Reactome	Containing curated knowledge of biological pathways	http://www.reactome.org/	[27]
	Kyoto Encyclopedia of Genes and Genomes (KEGG)	Pathway maps on the molecular interaction and reaction networks for metabolism	http://www.genome.jp/kegg/pathway.html	[28]
Protein family databases	Pfam	A highly comprehensive resource providing an optimised set of Hidden Markov Model profiles for protein domain families	http://pfam.jouy.inra.fr/	[29]
	Simple modular architecture research tool (SMART)	Resource for identification and annotation of protein domains and the analysis of domain architectures	http://www.smart.embl.de/	[30]
	PRINTS	Containing conserved motifs used to characterize a protein family	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php	[31]
	ProDom	Containing protein domain families automatically generated from the UniProtKB	http://www.prodom.prabi.fr/prodom/current/html/home.php	[32]
	InterPro	Integrated resource of protein families, domains and functional sites from Pfam, PRINTS, PROSITE, ProDom, SMART, PIRSF etc.	http://www.ebi.ac.uk/interpro/	[33]
	TIGRFAMs	Containing the protein families which are built in a similar fashion to Pfam but also containing whole protein chains	http://www.tigr.org/TIGRFAMs/index.shtml	[34]
	PROSITE	Containing protein domains, families and functional sites as well as associated patterns and profiles to identify them	http://expasy.org/prosite/	[35]
	Clusters of Orthologous Groups of proteins (COGs)	Phylogenetic classification of proteins encoded in complete genomes	http://www.ncbi.nlm.nih.gov/COG/	[36]
Proteomics databases	WORLD-2DPAGE Constellation	List of World-2DPAGE database servers, World-2DPAGE Portal that queries simultaneously worldwide proteomics databases, and World-2DPAGE Repository	http://world-2dpage.expasy.org/	[37]
	Global Proteome Machine Database (GPMDB)	Containing mass spectral library for data from a variety of organisms, the identified peptides are matched to the Ensembl genome database	http://www.thegpm.org/GPMDB/index.html	[38]
	PRoteomics IDEntifications database (PRIDE)	Containing protein and peptide identifications that have been described in the scientific literature together with the evidence supporting these identifications	http://www.ebi.ac.uk/pride/	[39]
	PeptideAtlas	Containing peptides identified in a large set of LC–MS/MS proteomics experiments	http://www.peptideatlas.org/	[40]
	Peptidome	Containing tandem mass spectrometry peptide and protein identification data	http://www.ncbi.nlm.nih.gov/peptidome/	[41]

structure, IDPs carry out important biological functions being typically involved in regulation, signaling and control. They often carry out these functions through high-specificity low-affinity interactions involving the multiple binding of one protein to many partners or the multiple binding of many proteins to one partner. DisProt collects and organizes knowledge regarding the experimental characterization and the functional associations of IDPs. DisProt is a collaborative effort between the Center for Computational Biology and Bioinformatics at Indiana University School of Medicine and the Center for Information Science and Technology at Temple University.

Disprot release 5.7 (28 February 2011) includes 643 proteins and 1375 disordered regions.

1.1.4 PhosphoSite

PhosphoSite [11] is an online systems biology resource providing comprehensive information and tools for the study of protein post-translational modifications (PTMs). PhosphoSite provides information about phosphorylated residues and the surrounding sequences, orthologous sites in other species, the location of the site within known domains and motifs, and relevant literature references. Cross-references are provided to a number of external resources for protein sequences, structures, PTMs and signaling pathways, as well as to sources of phospho-specific antibodies and probes. As the knowledgebase expands, users will be able to retrieve information about the kinases, phosphatases, ligands, treatments, and receptors that have been shown to regulate the phosphorylation status of the sites and the pathways in which the phosphorylation sites function. PhosphoSite provides an extensive, manually curated phosphorylation site database and also includes other commonly studied PTMs. This resource provides an easily accessible overview of the role of different phosphorylation sites, the experimental evidence for the modification, and the cell types in which the modification was found.

The latest PhosphoSite release (03 November 2011) contains 100884 non-redundant phosphorylation sites, 19185 ubiquitination sites, 7849 acetylation sites, 379 di-methylation sites, 303 mono-methylation sites, 139 methylation sites, 622 sumoylation sites, and 602 O-GlcNAc sites.

1.2 Protein structure databases

Protein structure databases describe experimentally determined protein structures and provide useful links, analyses, and schematic diagrams that relate 3D structure to biological function. A growing body of experimental data supports the notion that the structure of a protein reflects the nature of its role and therefore determines its biological function. Some the databases classify 3D structures by their folds because this can often reveal evolutionary relationships that may be hard to detect from sequence comparisons alone. A large number of databases for more specialized users, for example, those dealing with specific families, diseases, and structural features are also available.

1.2.1 Worldwide Protein Data Bank (wwPDB)

Since 2003, the Protein Data Bank Archive (PDB Archive) [13] has been managed by an international consortium called the Worldwide Protein Data Bank (wwPDB) whose partners comprise the RSCB PDB, the Macromolecular Structure Database (MSD, now known as the PDBe) at the European Bioinformatics Institute (EBI), the Protein Data Bank Japan (PDBj) at Osaka University and, more recently, the BioMagResBank (BMRB) at the University of Wisconsin-Madison. The PDB Archive is a collection of flat files in three different formats: the legacy PDB file format; the PDB exchange format that follows the mmCIF syntax (http://www.deposit.pdb.org/mmcif/); and the PDBML/XML format. For many of the structures, the original experimental data is also available. Thus, for structural models solved by X-ray crystallography, the structure factors from which the model was derived can be downloaded, while for structures solved by nuclear magnetic resonance (NMR) spectroscopy, the original distance and angle restraints can be obtained. An important task of the wwPDB was to remedy the legacy the PDB archive related mainly to ligands and literature references by fixing and making consistent all the PDB data.

wwPDB (27 April 2011) contains 69177 structures each of which is identified by a unique four-character reference code, the PDB identifier.

1.2.2 Structural Classification of Proteins (SCOP)

The Structural Classification of Proteins (SCOP) [17] database provides a comprehensive and detailed description of the evolutionary and structural relationships of proteins with known structures. The SCOP classification is constructed based on the domains in experimentally determined protein structures and includes family, superfamily, fold and class based on the secondary structure content and organization of the folds. SCOP also contains information on species and on groups with similar structures and similar functions.

SCOP release 1.75 (June 2009) includes 38221 PDB entries, 1195 folds, 1962 superfamilies and 3902 families.

1.2.3 Phospho3D

Phospho3D [20], a database of 3D structures of phosphorylation sites, stores information retrieved from Phospho.ELM (a database of S/T/Y phosphorylation sites) that is enriched with structural information and annotations at the residue level. Phospho3D also collects the results of large-scale structural comparisons of the 3D zones versus a representative dataset of structures, thus each P-site is associated to a number of structurally similar sites. Phospho3D has several additional features that include new structural descriptors, the possibility of selecting non-redundant sets of 3D structures and the availability for download of non-redundant sets of structurally annotated P-sites. Users can browse the data, search the database using kinase name, PDB identification code or keywords, and submit a protein structure and scan it against the 3D zones in the Phospho3D database.

Phospho3D (August 2010) includes 1770 mapped Phospho.ELM instances, 2083 distinct PDB files, 2158 distinct PDB chains and 5387 phosphorylation sites mapped onto a PDB structure.

1.3 Databases of protein-protein interactions and complexes

To perform their functions, proteins work together through various forms of direct or indirect interaction mechanisms. For a variety of basic functions, many proteins form a large complex representing a molecular machine or a macromolecular super-structural building block. High-throughput techniques for the detection of protein-protein interactions have matured and protein interaction data is now available on a large scale. Curated databases of protein–protein interactions have become a necessity for efficient research into how proteins function. Databases of protein-protein interactions and complexes maintain information about inter-molecular interactions, metabolic pathways, regulatory pathways, and the complexes that underlie many biological processes. With the development of retrieval tools and integration into annotation pipelines, these databases will become important resources for the discovery of new biomolecular mechanisms.

1.3.1 IntAct

IntAct [23] is an open source database and software framework for the storage, presentation and analysis of protein interaction data. Most of the interaction data is from protein-protein interactions, but IntAct also captures data for non-protein molecular interactors such as DNA, RNA, and small molecules. IntAct uses a flexible data model that can accommodate high levels of experimental details. Technical details about the experiment, binding sites, protein tags and mutations are annotated with the Molecular Interaction Ontology of the Proteomics Standard Initiative (PSI-MI). The IntAct website provides both textual and graphical views of protein interactions. The interactive viewer provides a number of unique features such as highlighting the node based on the molecule type, Gene Ontology, InterPro annotation, experimental and biological role, and species. Users can iteratively develop complex queries, exploiting the detailed annotation with hierarchical controlled vocabularies. Results can be obtained at any stage in a simplified, tabular view. A specialized view allows ‘zooming in’ on the full annotation of interactions, interactors and their properties.

IntAct version 2.0 contains 266855 binary interactions, 56486 proteins, 13103 experiments, and 1665 controlled vocabulary terms.

1.3.2 Human Protein Reference Database (HPRD)

HPRD [24] is a database of curated proteomic information pertaining to human proteins in healthy and diseased states. Although it is not a protein–protein interaction database, it contains an extensive list of interaction data of human proteins. All data in HPRD are manually extracted from public literature and curated by a team of trained biologists. The data are freely available for academic users and can be downloaded in either tab-delimited or XML formats. The whole database or only protein-protein interaction data without annotations can be downloaded in tab-delimited or PSI-MI format. HPRD is also linked to NetPath, a compendium of human signaling pathways, which currently contains annotations for several cancers and immune signaling pathways.

HPRD release 9 (13 April 2010) includes 30047 protein entries, 39194 protein-protein interactions, 93710 PTMs, 112158 protein expression, 22490 subcellular localization, 470 domains, and 453521 PubMed links.

1.3.3 Unified Human Interactome (UniHI)

Unified Human Interactome (UniHI) [26] is a comprehensive database of computational and experimental based human protein interaction networks. The database is intended to integrate diverge maps, providing a flexible and direct entry gate into the human interactome.

A variety of protein identifiers are supported by UniHI. Users can submit a set of proteins to obtain their functional information and interacting partners. The results are returned as a list of matched proteins together with names of the original source databases. UniHI also provides an interactive viewer to visualize the interaction networks. To analyze the human interactome, the UniHI website provides two powerful tools, UniHI Express and UniHI Scanner. UniHI Express can be used to refine the interaction networks based on gene expression in selected tissues to construct a tissue-specific network. UniHI Scanner can be used to compare the extracted networks with the pathways from Kyoto Encyclopedia of Genes and Genomes (KEGG) to detect new components in existing pathways. Proteins involved in multiple pathways that might be useful for disease-related studies can also be identified by UniHI Scanner.

UniHI 4 version 4.0 contains 253980 distinct interactions between 22307 unique human proteins.

1.3.4 Kyoto Encyclopedia of Genes and Genomes (KEGG)

KEGG [28] is an integrated database resource consisting of 16 main databases, broadly categorized into systems information, genomic information and chemical information. Systems information represents functional aspects of the biological systems, such as the cell and the organism, that are built from the building blocks. Genomic and chemical information represents the molecular building blocks of life in the genomic and chemical spaces, respectively. KEGG has been widely used as a reference knowledge base for the biological interpretation of large-scale datasets generated by sequencing and other high-throughput experimental technologies.

On 4 May 2011 KEGG contained 134607 pathway maps, 38205 functional hierarchies, 392 pathway modules and complexes, 375 human diseases, 9347 drugs, 835 crude drugs and other natural products, 14615 orthology groups, 1588 organisms, 6405661 genes in high-quality genomes, 372418 genes in draft genomes, 3792883 genes as EST contigs, 669846 genes in metagenomes, 17379 metabolites and other small molecules, 10978 glycans, 8451 biochemical reactions, 12547 reactant pair chemical transformations, 2324 reaction classes and 5391 enzyme nomenclatures.

1.4 Protein family databases

The databases storing information on the sequences and structures of proteins have been used to develop new resources with value-added information by classifying proteins into families according to their evolutionary relationships. These resources can provide extensive insights into evolution and, in particular, can support investigations into how proteins mutate and how function evolves over time. Such analyses have greatly assisted the transfer of functional annotations between experimentally characterized and uncharacterized genes.

1.4.1 Pfam

The Pfam database [29] is a large collection of multiple sequence alignments and hidden Markov models covering many common protein families. The database categorizes 75 percent of known proteins to form a library of protein families. The open access resource was established at the Welcome Trust Sanger Institute in 1998. Its vision is to provide a tool which allows experimental, computational and evolutionary biologists to classify protein sequences and answer questions about what they do and how they have evolved. The Pfam website also provides information on domain compositions.

Pfam version 25.0 (March 2011) contains alignments and models for 12273 protein families, based on the UniProtKB/Swiss-Prot and UniProtKB/TrEMBL protein sequence databases.

1.4.2 InterPro

InterPro [33] is a database of protein families, domains and functional sites in which identifiable features found in known proteins can be applied to new protein sequences to functionally characterize them. It classifies sequences at the superfamily, family and subfamily levels, predicting the occurrence of functional domains, repeats and important sites. The contents of InterPro are based on diagnostic signatures and the proteins that contain regions that significantly match these signatures. Each of the member databases of InterPro contributes towards a different niche, from very high-level, structure-based classifications (SUPERFAMILY and CATH-Gene3D) through to quite specific sub-family classifications (PRINTS and PANTHER). InterPro adds in-depth annotation, including Gene Ontology (GO) terms, to the protein signatures. Users can analyze an entire new genome using a downloadable version of InterProScan which can be incorporated into existing local pipelines. InterPro provides structural information from PDB, its classification in CATH and SCOP, as well as homology models from ModBase and SwissModel. Therefore, users can perform a direct comparison of the protein signatures with the available structural information.

InterPro release 32.0 (18 April 2011) contains 21516 entries, representing: 100 active sites, 66 binding sites, 628 conserved sites, 5974 domains, 14469 families, 16 PTMs and 263 repeats.

1.4.3 PROSITE

PROSITE [35] is a database of protein families and domains. It consists of entries describing the protein families, domains and functional sites as well as the amino acid patterns, signatures, and profiles in them. It is based on the observation that while there is a huge number of different proteins, most of them can be grouped, on the basis of similarities in their sequences, into a limited number of families. Proteins or protein domains belonging to a particular family generally share functional attributes and are derived from a common ancestor. The content of PROSITE is manually curated by a team at the Swiss Institute of Bioinformatics and tightly integrated into the UniProtKB/Swiss-Prot protein annotation. PROSITE was created in 1988 by Amos Bairoch who directed the group for more than 20 years. In July 2009, Ioannis Xenarios took over as the director of the PROSITE, UniProtKB/Swiss-Prot and bioinformatics competence center Vital-IT groups. PROSITE currently contains patterns and profiles specific to more than a thousand protein families or domains. Each of these signatures comes with documentation providing background information on the structure and function of these proteins. On the PROSITE website, users can perform keywords-based searches, and browse the motif entries, ProRule description, taxonomic scope, and number of positive hits. PROSITE provides the ScanProsite tool which can be used either to scan protein sequences for the occurrence of PROSITE motifs by entering UniProtKB or PDB identifier(s) or protein sequence(s), or to scan the UniProtKB or PDB databases for the occurrence of a pattern by entering the PROSITE identifier or the user's own pattern(s). ScanProsite can also be accessed programmatically through a simple HTTP web service.

PROSITE release 20.72 (7 April 2011) contains 1609 documentation entries, 1308 patterns, 922 profiles and 917 ProRule containing functional and structural information on PROSITE profiles.

1.5 Proteomics databases

Protein function analysis has evolved from the careful design of assays that address specific questions to the high-throughput 2D-gel and mass spectrometry (MS) based proteomics technologies that yield proteome-wide maps of protein expression or interaction. Because these technologies depend heavily on information storage, representation and analysis, several proteomics databases are available while new resources are still emerging.

1.5.1 World-2DPAGE Constellation

The World-2DPAGE Constellation [37] is composed of the established WORLD-2DPAGE List of 2-D PAGE database servers, the World-2DPAGE Portal that simultaneously queries world-wide proteomics databases, and the recently created World-2DPAGE Repository. The WORLD-2DPAGE List is an index to known federated 2-D PAGE database, services and related servers. Databases are grouped by species and classified in three categories according to their implementation of the rules defining a federated 2-DE database. WORLD-2DPAGE List currently lists up to 60 databases totalizing nearly 400 gel images. The World-2DPAGE Portal is a dynamic portal that can be used to simultaneously query world-wide gel-based proteomics databases. The Portal can be seen as a virtual unique database with up to 133 reference maps for 20 species totalizing nearly 18,000 identified spots. The World-2DPAGE Repository is a public standards-compliant repository for gel-based proteomics data linked to protein identification published in the literature.

The World-2DPAGE (15 March 2011) consists of 18 articles with 28 reference maps for 18 species and 5700 identified spots.

1.5.2 Proteomics Identifications Database (PRIDE)

PRIDE [39] is a prominent public data repository of MS based proteomics data that is maintained by the European Bioinformatics Institute as part of the Proteomics Services Team. PRIDE stores three different kinds of information: peptide and protein identifications derived from MS or tandem MS (MS/MS) experiments, MS and MS/MS mass spectra as peak lists, and any or all associated metadata. PRIDE was established as a production service in 2005. Several other proteomics databases have been established over the past few years, including GPMDB [38], PeptideAtlas [40], and the NCBI Peptidome [41]. NCBI Peptidome and PRIDE are structured data repositories that store the original experimental data from the researchers and do not assume any editorial control over the submitted data. PRIDE contains data from about 60 species; the biggest fraction is from human samples, followed by the fruitfly *Drosophila melanogaster* and mouse. Users can submit data obtained using different MS-based proteomics technologies in PRIDE XML or mzData XML formats. PRIDE can be queried by experiment accession number, protein accession number, literature reference, and by sample parameters including species, tissue, sub-cellular location and disease state. The query results can be retrieved as PRIDE XML, mzData XML, or HTML. PRIDE also provides access to public PRIDE data from a query-optimized data warehouse as well as programmatic web service access in a BioMart [42] interface that allows complex queries to be constructed.

The PRIDE database (23 Jun 2011) contains 16476 experiments, 5024880 identified proteins, 24498871 identified peptides, 3299204 unique peptides and 146864975 spectra.

1.5.3 Global Proteome Machine Database (GPMDB)

The GPMDB [38] was constructed to use the information obtained by Global Proteome Machine (GPM) servers to aid in the difficult process of validating peptide MS/MS spectra and protein coverage patterns. This database has been integrated into the GPM server pages [38], allowing users to quickly compare their experimental results with the best results that have been observed previously by users of the machine. GPMDB does not hold complete records of proteomics experiments; rather it holds the minimum amount of information necessary for bioinformatics-related tasks such as sequence assignment validation. Most of the data is held in a set of XML files and the database serves as an index to those files, allowing for very rapid lookups and reduced database storage requirements.

2 Discussion

Human protein bioinformatics databases offer scientists the opportunity to access a wide variety of biologically relevant data, including protein sequences, structures and functions. However, the growing interest in functional proteomics is fuelled not only by the prospect of a true functional understanding but also by substantial improvements in technology and methodology. Advances in protein identification technologies, in particular MS, have made possible the establishment of proteomics databases and extended the sensitivity, accuracy and speed of analysis, making it possible to routinely identify several thousand proteins per experiment [43]. The introduction of MS methods for accurate relative and absolute protein quantification and the large-scale analysis of PTMs, such as phosphorylation and ubiquitylation, have allowed truly functional proteomics to be carried out [44]. MS is now joined by antibody and protein-protein interaction arrays, fluorescence- and flow cytometry-based detection of proteins and PTMs, and optical spectroscopic methods of proteome analysis [45]. These latter techniques are promoted by an ever increasing repertoire of specific antibodies against proteins and PTMs, and bring single-cell proteomics into reach. Therefore, the functional proteomics approaches, when integrated with the information from human protein bioinformatics databases, offer an effective route to biomarker and drug discovery by pinpointing signaling pathways and components that are differentially regulated in particular diseases. Some examples of how protein databases have been used in functional proteomics are described in the following section.

2.1 Functional proteomics mapping of Smad signaling involved in several human pathologies

Access to the human genome functional proteomics databases have led to descriptions of large protein-protein interaction networks. An efficient strategy for the functional exploration of complex proteomes requires (i) the in-depth annotation of protein-protein interaction maps with all the available information on proteins, protein domains, and interactions; (ii) an exploration tool that allows easy navigation of complex databases; and (iii) streamlined functional assays to validate newly identified proteins. Colland *et al.* [46] have applied an integrated strategy of this kind for the identification of new factors implicated in the Smad signaling pathway involved in several human pathologies. They used two-hybrid screening to map Smad signaling protein-protein interactions and established a network of 755 interactions involving 591 proteins, 179 of which were poorly annotated or not annotated in the existing databases of protein-protein interactions and complexes. The exploration of the databases was improved by the use of PIMRider [47], a dedicated navigation tool accessible through the Web. In their study, they successfully illustrated the biological meaning of the network after indentifying the presence of 18 known Smad-associated proteins. Colland *et al.* then performed functional assays including siRNA knock-down experiments in mammalian cells and identified eight novel proteins involved in Smad signaling. The success of this study demonstrates the validity of using an integrated functional proteomics approach.

2.2 Intrinsically disordered proteins (IDPs) and functional proteomics

The recent discovery of IDPs has significantly broadened the view of the scientific community and increased the number of groups systematically studying these intriguing proteins [48]. IDPs and ID regions are typically involved in regulation, signaling and control pathways where they complement the functional repertoire of the more ordered regions that typically carry out efficient catalysis [49]. The many IPD databases (for example, ProDDO [50] and DisProt [6]) are increasingly being used in individual and high-throughput experiments (i) to improve estimations of the commonness of ID regions and their functional repertoire; (ii) to aid or improve prediction of other protein features such as protein PTM sites or other types of binding regions; and (iii) to gain insight into structural and dynamic properties of the proteins of interest.

2.3 The use of human protein bioinformatics databases to retrieve biological information on the myc proto-oncogene protein (c-myc)

The UniProtKB/Swiss-Prot entry for human c-myc (accession number P01106) contains the following information: (i) the ‘Comments’ section provides an overview of its biological function; (ii) the list of keywords gives a very useful quick impression about the protein and can be used to find other proteins that have been annotated with the same keyword; and (iii) the ‘Features’ sections contains annotations that are localized to particular residues or regions of the sequence including a helix-loop-helix motif, a potential leucine zipper, and a basic motif all towards the C-terminal end. Information about the 353 amino acids at the N-terminal end (about two-thirds of the protein) consists mostly of PTMs and areas with compositional bias. Therefore, what might be the function of this largely unannotated region, and which residues might be involved? The features section ends with a list of the positions of secondary structure elements that have been experimentally validated, in c-myc this is three alpha-helices at the C-terminal end. The c-myc RefSeq entry (accession number NP_002458.2) contains comments on a protein isoform, created using a downstream alternative start codon, which may have some role in the cell. The ‘FEATURES’ section of this entry contains details on the residues involved in particular functions, in this case DNA binding and dimerization; however, no new information on the N-terminal part of the sequence is available. To check if the lack of annotation might be due to a high degree of structural and/or functional flexibility in this region of the sequences, we used DisProt to see if the c-myc regions had been annotated with experimentally validated intrinsic disorder. A keyword search for ‘c-myc’, found the entry ‘DP00260’, which lists a series of experiments that show the propensity for intrinsic disorder in the N-terminal part of the protein including the region around the threonine at position 58 (T58); this residue was annotated to be sometimes phosphorylated and sometimes glycosylated in UniProtKB/Swiss-Prot and RefSeq. Because disordered regions often harbor PTM sites that modulate molecular interactions, and because phosphorylation is the best understood PTM, we consulted PhosphoSite. Using a keyword search for ‘T58’, we found that various functions have been associated with this site. To obtain more detailed information about the interactions or pathways that this protein might be involved in, we continued our search in protein structure, family, and interaction databases. Five PDB IDs (1A93, 1EE4, 1MV0, 1NKP and 2A93) associated with P01106 were found. The Pfam database has annotated protein-myc as belonging to three protein families: PF00010 (HLH family, helix-loop-helix DNA binding domain), PF01056 (Myc-N family, myc amino-terminal region) and PF02344 (Myc-LZ family, myc leucine zipper domain). The KEGG database revealed that P01106 is involved in 14 pathways, including the MAPK signaling pathway, the ErbB signaling pathway, cell cycle and the Wnt signaling pathway all of which play important roles in the progression of various human cancers.

3 Future perspectives

Functional proteomics strategies that uses information from the human protein databases are expected to provide an integrated picture of the expression levels and properties of the thousands of protein components of organelles, pathways, and cytoskeletal systems, both in health and disease. Recent major developments in protein-complex purification, MS and bioinformatics databases will progress the analysis of the human proteome.

What are the major challenges and future goals? First, although a variety of protein bioinformatics databases have been developed to catalog and store different information about proteins, it is still important to develop new solutions to facilitate comparative analysis, data-driven hypothesis generation, and biological knowledge discovery. Second, before a complete human proteomic analysis can be applied to the study of diseases, many challenges need to be addressed. They include the heterogeneity of biopsy material, the need to develop better image analysis systems for supporting gel comparisons, quantitation and databasing, determination of interacting partners, functional aspects, and the lack of procedures for identifying and functionally characterizing target genes that lie in disease pathways.

We are confident that all these challenges will be addressed as increasing numbers of scientists begin to apply protein bioinformatics databases to functional proteomics research as well as to clinically relevant questions in biomedical research. We predict that medicine will profit enormously from the development of integrated functional proteomics strategies for the identification and characterization of biomarkers and drug targets for disease diagnosis and therapeutics.

This work was supported by the National Basic Research Program of China (Grant Nos. 2010CB912700 and 2011CB910601).

1 Godovac-Zimmermann J, Brown L R. Perspectives for mass spectrometry and functional proteomics. Mass Spectrom Rev, 2001, 20: 1–57

2 Gavin A C, Bosche M, Krause R, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 2002, 415: 141–147

3 Monti M, Orru S, Pagnozzi D, *et al*. Functional proteomics. Clinica Chimica Acta, 2005, 357: 140–150

4 Pruitt K D, Tatusova T, Maglott D R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res, 2007, 35: D61–D65

5 The UniProt Consortium. The universal protein resource (UniProt) in 2010. Nucleic Acids Res, 2010, 38: D142–D148

6 Sickmeier M, Hamilton J A, LeGall T, *et al*. DisProt: the database of disordered proteins. Nucl Acids Res, 2007, 35: D786–D793

7 Sayers E W, Barrett T, Benson D A, *et al*. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res, 2007, 35: D5–D12

8 Leinonen R, Diez F G, Binns D, *et al*. UniProt archive. Bioinformatics, 2004, 20: 3236–3237

9 Suzek B E, Huang H, McGarvey P, *et al*. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics, 2007, 23: 1282–1288

10 Rebhan M. Protein sequence databases. Methods Mol Biol, 2010, 609: 45–57

11 Hornbeck P V, Chabra I, Kornhauser J M, *et al*. PhosphoSite: a bioinformatics resource dedicated to physiological protein phosphorylation. Proteomics, 2004, 4: 1551–1561

12 Wang Y, Addess K J, Chen J, *et al*. MMDB: annotating protein sequences with Entrez's 3D-structure database. Nucleic Acids Res, 2007, 35: D298–D300

13 Berman H, Henrick K, Nakamura H, *et al*. The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res, 2007, 35: D301–D303

14 Pieper U, Eswar N, Webb B M, *et al*. MODBASE, a database of annotated comparative protein structure models and associated resources. Nucleic Acids Res, 2009, 37: D347–D354

15 Kiefer F, Arnold K, Künzli M, *et al*. The SWISS-MODEL repository and associated resources. Nucleic Acids Res, 2009, 37: D387–D392

16 Cuff A L, Sillitoe I, Lewis T, *et al*. The CATH classification revisited-architectures reviewed and new ways to characterize structural divergence in superfamilies. Nucleic Acids Res, 2009, 37: D310–D314

17 Andreeva A, Howorth D, Chandonia J M, *et al*. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res, 2008, 36: D419–D425

18 Bogatyreva N S, Osypov A A, Ivankov D N. KineticDB: a database of protein folding kinetics. Nucleic Acids Res, 2009, 37: D342–D346

19 Garavelli J S. The RESID database of protein modifications as a resource and annotation tool. Proteomics, 2004, 4: 1527–1533

20 Zanzoni A, Ausiello G, Via A, *et al*. Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. Nucleic Acids Res, 2007, 35: D229–D231

21 Salwinski L, Miller C S, Smith A J, *et al*. The database of interacting proteins: 2004 update. Nucleic Acids Res, 2004, 32: D449–D451

22 Zanzoni A, Montecchi-Palazzi L, Quondam M, *et al*. MINT: a Molecular INTeraction database. FEBS Lett, 2002, 513: 135–140

23 Aranda B, Achuthan P, Alam-Faruque Y, *et al*. The IntAct molecular interaction database in 2010. Nucleic Acids Res, 2010, 38: D525–D531

24 Keshava Prasad T S, Goel R, Kandasamy K, *et al*. Human Protein Reference Database—2009 update. Nucleic Acids Res, 2009, 37: D767–D772

25 Snel B, Lehmann G, Bork P, *et al*. STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res, 2000, 28: 3442–3444

26 Chaurasia G, Malhotra S, Russ J, *et al*. UniHI 4: new tools for query, analysis and visualization of the human protein-protein interactome. Nucleic Acids Res, 2009, 37: D657–D660

27 Matthews L, Gopinath G, Gillespie M, *et al*. Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res, 2009, 37: D619–D622

28 Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res, 2000, 28: 27–30

29 Finn R D, Tate J, Mistry J, *et al*. The Pfam protein families database. Nucleic Acids Res, 2008, 36: D281–D288

30 Letunic I, Doerks T, Bork P. SMART 6: recent updates and new developments. Nucleic Acids Res, 2009, 37: D229–D232

31 Bru C, Courcelle E, Carrère S, *et al*. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res, 2005, 33: D212–D215

32 Attwood T K. The PRINTS database: a resource for identification of

protein families. Brief Bioinform, 2002, 3: 252–263

33 Hunter S, Apweiler R, Attwood T K, *et al*. InterPro: the integrative protein signature database. Nucleic Acids Res, 2009, 37: D211–D215

34 Haft D H, Selengut J D, White O. The TIGRFAMs database of protein families. Nucleic Acids Res, 2003, 31: 371-373

35 Hulo N, Bairoch A, Bulliard V, *et al*. The 20 years of PROSITE. Nucleic Acids Res, 2008, 36: D245–D249

36 Tatusov R L, Fedorova N D, Jackson J D, *et al*. The COG database: an updated version includes eukaryotes. BMC Bioinformatics, 2003, 4: 41–54

37 Hoogland C, Mostaguir K, Appel R D, *et al*. The World-2DPAGE Constellation to promote and publish gel-based proteomics data through the ExPASy server. J Proteomics, 2008, 71: 245–248

38 Craig R, Cortens J C, Fenyo D, *et al*. Using annotated peptide mass spectrum libraries for protein identification. J Proteome Res, 2006, 5: 1843–1849

39 Vizcaíno J A, Côté R, Reisinger F, *et al*. The proteomics identifications database: 2010 update. Nucleic Acids Res, 2009, 38: D736–D742

40 Deutsch E W, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep, 2008, 9: 429–434

41 Slotta D J, Barrett T, Edgar R. NCBI peptidome: a new public repository for mass spectrometry peptide identifications. Nat Biotechnol, 2009, 27: 600–601

42 Kasprzyk A. BioMart: driving a paradigm change in biological data management. Database (Oxford), 2011, bar049

43 Schulz K R, Danna E A, Krutzik P O, *et al*. Single-cell phospho-protein analysis by flow cytometry. Curr Protoc Immunol, 2007, Chapter 8: Unit 8.17

44 Fournier F, Guo R, Gardner E M, *et al*. Biological and biomedical applications of two-dimensional vibrational spectroscopy: proteomics, imaging, and structural analysis. Acc Chem Res, 2009, 42: 1322–1331

45 Faley S L, Copland M, Wlodkowic D, *et al*. Microfluidic single cell arrays to interrogate signalling dynamics of individual, patient derived hematopoietic stem cells. Lab Chip, 2009, 9: 2659–2664

46 Colland F, Jacq X, Trouplin V, *et al*. Functional proteomics mapping of a human signaling pathway. Genome Res, 2004, 14: 1324-1332

47 Formstecher E, Aresta S, Collura V, *et al*. Protein interaction mapping: a *Drosophila* case study. Genome Res, 2005, 15: 376-384

48 Dyson H J, Wright P E. According to current textbooks, a well-defined three-dimensional structure is a prerequisite for the function of a protein. Is this correct? IUBMB Life, 2006, 58: 107–109

49 Radivojac P, Iakoucheva L, Oldfield Christopher, *et al*. Intrinsic disorder and functional proteomics. Biophys J, 2007, 92: 1439–1456

50 Sim K L, Uchida T, Miyano S. ProDDO: a database of disordered proteins from the Protein Data Bank (PDB). Bioinformatics, 2001, 17: 379–380