IJPT
0000-0000/03/31-1-11
IRANIAN JOURNAL OF PHARMACOLOGY & THERAPEUTICS

IJPT 3:1-11, 2004

RIDR

Protein Databases

ZARRIN MINUCHEHR and BAHRAM GOLIAEI

National Institute for Genetic Engineering and Biotechnology, Tehran, Iran (Z.M.); Bioinformatics Center, Laboratory of Biophysics and Molecular Biology, Institute of Biochemistry and Biophysics, University of Tehran, Tehran, Iran (B.M.)

Received December 6, 2003; Revised March 14, 2004; Accepted April 6, 2004

This paper is available online at *http://ijpt.iums.ac.ir*

ABSTRACT

Proteins are sources of many peptides with diverse biological activity. Some of them are considered as valuable components of foods and drug targets with desired and designed biological activity. We are now entering an era rich in biological data in which the field of bioinformatics is poised to exploit this information in increasingly powerful ways. There are currently many databases all over the world housing information that floods from the genome project. Nevertheless, as databases become more comprehensive, as the volume of sequence data expands and search outputs become more complex, the different databases play an increasingly major role in the post genomic era. However, the rate of sequence generation and overwhelmingly proliferation of databases have made it difficult to keep up with developments. Researchers now need rapid, easy to use, reliable tools for their use in functional characterization of newly determined sequences. This paper aims to provide a comprehensive overview on the status of protein databases available in the World Wide Web. As this is a fast moving area, a list of all the mentioned databases along with their current URLs are presented in a table at the end of the article.

Keywords: Protein, Database, Bioinformatics, Proteomics, Databank

The development of protein-sequencing methods [1] led to the sequencing of several new proteins. Margaret Dayhoff [2] and her colleagues worked to assemble databases of protein sequences, their work led to a database known as PIR or Protein Identification Resources. PIR can be mentioned as the first organized protein database. Biological databases and resources may contain many different kinds of information, each record of the database, is typically called an entry and each entry has two main parts:

1. Descriptive information or Annotation.
   a. Description
   b. Literature References
2. The raw data sequences or observations, which are sometimes, also called the core data.

Annotation is the information added to primary information or raw data in a biological database, such as references, organism, comments, protein structure, binding site etc. Many database groups all over the world have defined groups of experts to help annotate proteins, protein structures, protein families, etc. Protein databases play a key role in modern biology and are especially important in Molecular Biology, Genomics, Functional Genomics, System Biology, Protein Design, Protein Engineering, Drug Design, Pharmacological Development and Medicine.

A question that may come to mind is that why we have so many databases instead of holding all the data in a single database. The reason is that each database is designed to answer a specific question and different databases hold different types of data. The field of Bioinformatics and constructing biological databases is young and fast growing. Therefore, because of the fast growing databases all over the world and the lack of a brief introduction to existing protein databases, we here introduce some important protein databases available worldwide. In addition a relatively large number of available protein databases, along with a small description of the database including its URL and their citations are summarized in a comprehensive table at the end of the article.

RETRIEVAL TOOLS FOR PROTEIN DATABASES

Although each database may have its unique search tool in retrieving data, there are two major tools for retrieving information from molecular protein databases, Entrez and SRS (Sequence Retrieval System).

Entrez

Entrez is the retrieval system which accesses different protein sequences and structures via MMDB (Molecular Modeling Database), Entrez also integrates with

biomedical literatures via PubMed and Online Mendelian Inheritances in Man (OMIM). The protein sequences in Entrez are obtained from a variety of sources such as SWISS-PROT, PIR and GenBank protein translations. This service provides searching bibliographic records for sequences using Boolean queries as well as presenting links to related information. Some links are for example from a protein sequence to its related DNA sequence or to some specialized protein databases or to the abstract of the related paper which is used for obtaining the protein sequence [3].

SRS

SRS was initially developed at the EMBL (European Molecular Biology Laboratory) in the early 1990's as a "Sequence Retrieval System" to address the anticipated growth of the four public sequence repositories. But SRS evolved to address a far more complicated problems the rapidly growing number of biological databases are now containing diverse, but highly relevant biological data. Today, SRS is a flexible and scalable integration platform for biological data, which integrates data independently of the data source and format. The result is that SRS is a remarkably powerful tool in accommodating new data types and new formats.

The Sequence Retrieval System (SRS) is the world's premier data integration, analysis and display tool for bioinformatics, genomics and related databases. The power of SRS lies in its ability to effectively integrate heterogeneous data sources and analysis tools behind a single interface and integration framework. It ensures an effective integration of heterogeneous data without losing data in file and format conversions. At the same time, it provides a level of integration that adds value to the whole application.

SRS achieves this flexibility through a meta-level approach to integration. SRS captures and uses meta data containing all of the relevant information about the structure, format, and syntax of the underlying databases. This meta data also includes information about database cross references, enabling some of the most powerful SRS functionalities: cross database queries and views [4].

PRIMARY PROTEIN DATABASES

These databases are called primary since they have the information of the protein sequence. The major two primary protein databases are PIR and SWISS-PROT.

PIR (Protein Information Resource)

This protein database can be mentioned as the first classical database introduced to the protein science community. The PIR website allows sequence similarity and free-text searching of the protein sequence database. PIR produces the Protein Sequence Database (PSD) of functionally annotated protein sequences, which grew out of the Atlas of Protein Sequence and Structure [2] edited by Margaret Dayhoff and has been incorporated into an integrated knowledge based system of value-added databases and analytical tools, this annotated protein database now contains over 283000 sequences covering the entire taxonomic range. For over three decades, PIR has provided many protein databases and analysis tools freely accessible to the scientific community. iProClass, a central point for exploration of protein information, provides summary descriptions of protein family, function and structure for PIR-PSD, Swiss-Prot, and TrEMBL sequences, with links to over 50 biological databases Release 2.35, 24-Nov-2003, contains 1,169,177 entries. PIR-NREF, a comprehensive database for sequence searching and protein identification, contains non-redundant protein sequences from PIR-PSD, Swiss-Prot, TrEMBL, RefSeq, GenPept, and PDB. The PSD database Release 1.35, 24-Nov-2003, contains 1,397,398 entries, as a whole we could say that the PSD Protein Sequence Database is an integrating database from PIR, MIPS and JIPD, this database is publicly available and meant to be non redundant. PIR also provides context cross references between its own database entries. The web site corresponding to this database is http://pir.georgetown.edu/.

SWISS-PROT and TrEMBL

SWISS-PROT is a protein sequence database, which provides a high level of annotation. This database has a minimal level of redundancy and a high level of integration with other databases. TrEMBL is a computer annotated supplement to SWISS-PROT. SWISS-PROT database was created at the department of Medical Biochemistry of the University of Geneva and since then it has been a collaborative effort with the European Molecular Biology Laboratory (EMBL). Since 1987, the SWISS-PROT database is distinguished from other protein sequence databases by three distinct criteria i) annotation ii) minimal redundancy and iii) integration with other databases. The current URL for the database is http://www.expasy.ch/sprot. The 42.4 version of SWISS-PROT released 14-Nov-2003 contains 138,347 entries and the TrEMBL Release 25.4 on 14-Nov-2003 contains 1,013,930 entries.

SBASE

SBASE is a collection of protein domain sequences collected from the literature, from protein sequence databases and from genomic databases [5-6]. The protein domains are defined by their sequence boundaries given by the publishing authors or in one of the primary sequence databases (Swiss-Prot, PIR, TREMBL etc.). The SBASE database uses a set of theoretical approaches for representing similarities. Sequences that have an above threshold BLAST similarity score to at least one member of the group are called the neighbor of the group. In order to access this database the URL http://www.icgeb.trieste.it/sbase should be used.

PROTEIN PATTERN DATABASES

These databases are mainly created by analytical methods from the primary database. At the heart of the analysis methods that underpin pattern databases is the multiple sequence alignment method and different tech-

niques have evolved to exploit the fact of classifying proteins in different pattern databases. The different analytical methods to create pattern databases are given below:

Single motif methods. The idea is that a particular protein family can be characterized by the single most conserved region, for example an enzyme active site. The motif is then reduced to a consensus expression, for example the expression P-x-{DR}-[KQ] means that we have a conserve Proline (P) followed by an arbitrary residue (x) followed by any residue except Asp (D) or Arg (R) finally we have Lys (K) or Gln (Q).

Multiple motif methods. This method finds several motifs that characterize the aligned family within a sequence alignment in a database search, therefore a greater chance of identifying a distant relative exists. For example, we have a sequence, which matches only four of seven motifs, but may be diagnosed as a true match this is when the motifs are matched in the correct order in the sequence and the distances between them are consistent with those expected of true neighboring motifs. The ability to tolerate mismatches makes multiple motif matching a powerful diagnostic approach.

Profile methods. This method uses the variable regions between conserved motifs. In this method, the complete conserved position of the alignment (including gaps), becomes a discriminator or a profile, this profile defines which residues are allowed at the given positions, which residues are allowed at conserved and which degenerate. Profiles are sometimes related weight matrices and they provide a sensitive mean of detecting. Distant sequence relationships, where only a few residues are well conserved can be found using this method.

The different methods mentioned above have given rise to different pattern databases, despite the data source; these pattern databases are emerged from the conserved motifs shared in homologous sequences, which is thought to be crucial to the structure and function of proteins. Therefore searching the pattern database theoretically offers an insight to the sequence biological function. Since these pattern databases are derived from the multiple sequence alignment methods, one can identify a distant relationship between the sequences [7, 8].

Some important pattern databases are briefly introduced below and some of the rest are summarized in Table 2.

PROSITE

PROSITE is derived from the SWISS-PROT database. Entries are stored in PROSITE in two distinct files:

I. Pattern file which houses the pattern and lists the matches with other SWISS-PROT files.
II. Document file which is a plain annotation file, which describes the biological role of the motif and gives supporting bibliographic information, release 18.16, of 23-Nov-2003 contains 1231 documentation entries that describe 1671 different patterns, rules and profiles/matrices. The web site, which supports this database, is http://www.expasy.org/prosite/.

PRINTS

Motifs in PRINTS database is stored in the form of finger prints, PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterize a protein family; its diagnostic power is refined by iterative scanning of a SWISS-PROT/TrEMBL composite. Fingerprints can encode protein folds; they function more flexibly and powerfully than single motifs. Each entry in PRINTS is manually annotated, which makes this database a precious database to rely. The more motifs in a fingerprint, the better it will be able to identify distant relatives [9-12]. Release 36.0 of PRINTS 11-Apr-2003 contains 1800 entries, encoding 10,931 individual motifs. In order to access the database the URL http://bioinf.man.ac.uk/dbbrowser/PRINTS/ should be used.

Blocks

Blocks is based on the PROSITE identified families. Blocks are multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. Block Searcher, Get Blocks and Block Maker are aids to detection and verification of protein sequence homology. Matches to sets of blocks from the same family are unlikely to have arisen by chance but high scoring matches to single blocks seldom have biological significance. Since the Blocks database is derived automatically, their entries are not annotated, but at the same time, the entry includes links to PROSITE and PRINTS documentation files. The last Version of Blocks database is 13.0 and consists of 8656 blocks representing 2101 groups documented in InterPro 3.1 keyed to SWISS-PROT 39.17 and TrEMBL obtained from the InterPro server. In order to get in contact to the database you can connect to http://www.blocks.fhcrc.org/.

There is an ever growing of the pattern databases, each has a different diagnostic strength and weakness and is highly difficult to asses the quality of a particular resource, therefore a researcher should be aware in choosing the suitable database in his or her research.

PROTEIN TERTIARY STRUCTURE DATABASE

The Protein Databank

The Protein databank (PDB) is the single worldwide archive of structural data available on biological macromolecules particularly proteins. The Protein Data Bank (PDB) was established at Brookhaven National Laboratories (BNL) in 1971. At the beginning, the archive held seven crystal structures known by that time. In 1980 the number of deposited structures increased dramatically. In October 1998, the management of the PDB became the responsibility of the Research Collaboratory for Structural Bioinformatics (RCSB). The PDB distributes coordinate data, structure factor files and NMR constraint files. The Protein Data Bank now con-

tains a total of 21,117 Proteins, Peptides, and Viruses, 981 Protein/Nucleic Acid Complexes, 1,266 Nucleic Acids and 18 Carbohydrates structures which makes a total of 23,382 entries till 25-Nov-2003, of which 19,864 entries are resolved by X-ray diffraction and 3,518 are NMR structures. Current URL is http://www.rcsb.org/pdb/.

PROTEIN STRUCTURE CLASSIFICATION DATABASES

These databases provide structural comparisons for the proteins currently in the Brookhaven PDB and access to the sequences of these proteins. Each database offers different family coverage and different levels of annotation. This is vital for a user, who not only wants to discover whether a sequence has matched a predefined motif, but also needs to understand its biological significance.

SCOP

The SCOP database (Structural Classification Of Proteins) is based on expert definition of structural similarities. Release 1.63 of SCOP is obtained from 18,946 PDB Entries of 1-March-2003. Since this database is especially important in structural classification of proteins, some information about the number of folds, superfamilies and families in the SCOP database in its latest release are summarized in Table 1. In order to browse this database the site http://scop.berkeley.edu/ should be contacted.

CATH

The CATH protein structure database (Classification by Class, Architecture, Topology and Homology) protein structure database resides at http://www.Biochem .ucl.ac.uk/bsm/cath. Proteins are classified into hierarchical levels by class, this database is similar to SCOP except that α/β and α+β proteins are considered to be in one class.

FSSP

FSSP (Fold classification based on Structure-Structure alignment of Proteins) is based on a structural alignment of the proteins in the Brookhaven structural database by the structural program DALI. The FSSP database is based on "an exhaustive all-against-all 3D structure comparison of protein structures currently in the Protein Data Bank (PDB)". The classification and alignments are automatically maintained and continuously updated using the Dali (distance matrix alignment) search engine. Chains are divided into a representative set and sequence homologs of structures in another set. An all-against-all structure comparison is performed on the set. This database contained 2,860 entries till 06-Aug-2002. The database is produced by the Sander group at EMBL, Heidelberg, and made available on the Web by the European Bioinformatics Institute (EBI), for more information contact http://www.bioinfo .biocenter.helsinki.fi:8080/dali/index.html.

HSSP

The HSSP (Homology-derived structures of proteins) database of protein structure sequence alignments and family profiles is a derived database merging information from three dimensional structures and one dimensional sequences of proteins. The database is useful for analyzing residue conservation in structures context for defining structurally meaningful sequence patterns, integral studying protein evolution, folding and design. This database contained 20,574 entries until 30-Nov-2003. The address to this database is http://www.cmbi.kun.nl/gv/hssp/

DSSP

The DSSP (Definition of Secondary Structure of Proteins) database of secondary structure and solvent accessibilities is a useful and widely used resource for this purpose. For more information read the Dictionary of Secondary Structure of Protein [13]. This database distinguishes eight secondary structural classes that can be grouped into alpha helices, beta strands and coils, the database contained 21,882 entries till 04-Dec-2003. The address to this site is http://www.sander.ebi.ac.uk/dssp/.

MMDB

Proteins of known structure in the Brookhaven PDB have been categorized into structurally related groups in MMDB (Molecular Modeling Database) by the VAST structural alignment program. The Molecular Modeling Database (MMDB) contains 3-D macromolecular structures, including proteins and polynucleotides. MMDB contains over 20,000 structures and is linked to the rest of the NCBI databases, including sequences, bibliographic citations, taxonomic classifications, and sequence and structure neighbors. In order to contact this database this URL http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml should be used.

SARF Database

The SARF database (Spatial Arrangement of backbone Fragments) is a structural database located at http://123d.ncifcrf.gov/sarf2.html. SARF can find structural similarity rapidly based on a search for secondary structural elements.

Since there are plenty of databases, each appropriate for a particular task, an alphabetical list of molecular biology databases relevant to protiens is summarized in Table 2.

Table 1. SCOP 1.63

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	171	286	457
All beta proteins	119	234	418
Alpha and beta proteins (a/b)	117	192	501
Alpha and beta proteins (a+b)	224	330	532
Multi-domain proteins	39	39	50
Membrane and cell surface proteins	34	64	71
Small proteins	61	87	135
Total	765	1232	2164

Table 2. Protein databases and their URL.

DATABASE	Description	URL	Reference
3DALI	A database of aligned protein structures and related sequences .	http://www.embl-heidelberg.de/argos/ali/ali_info.html	Pascarella and Argos (1992) [14]
3DPSSM	A Fast, Web-based Method for Protein Fold Recognition using 1D and 3D Sequence Profiles coupled with Secondary Structure and Solvation Potential Information.	http://www.sbg.bio.ic.ac.uk/~3dpssm/	Kelley et al .(2000) [15]
AACID	Amino acid indices and similarity matrices.	http://www.genome.ad.jp/dbget/aaindex.html	Kawashima et al. (1999) [16]
AARSDB	Aminoacyl-tRNA synthetase Database.	http://rose.man.poznan.pl/aars/index.html	Szymanski et al. (2000, 2001) [17, 18]
AraC	The AraC-XylS database contains information about a family of positive transcriptional regulators broadly distributed in bacteria.	http://www.arac-xyls.org/	Tobes et al .(2002) [19]
ASPD	ASPD is a new curated database that incorporates data on full-length proteins.	http://wwwmgs.bionet.nsc.ru/mgs/gnw/aspd/	Afonnikov et al. (2000) and Valuev et al. (2000) [20,21]
BIND	The Biomolecular Interaction Network Database.	http://www.bind.ca/	Bader et al. (2001, 2003) [22, 23]
BLOCKS	Multiple alignments of conserved regions of protein families.	http://blocks.fhcrc.org/	Henikoff et al. (1991,1998,2000) [24-26] Pietrokovski et al. (1996) [27]
BMRB	The BMRB database contains NMR chemical shifts derived from proteins and peptides.	http://www.bmrb.wisc.edu/	Seavey et al. (1991) [28]
CATH	Protein domain structures.	http://www.biochem.ucl.ac.uk/bsm/cath	Orengo et al. (1997) [29]
CCDC	The Cambridge Crystallographic Data Centre.	http://www.ccdc.cam.ac.uk/index.html	Allen et al. (2002) [30]
CDD	Alignment models for conserved protein domains.	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	Marchler-Bauer (2003) [31]
CluSTr	Automatic classification of SWISS-PROT+TrEMBL proteins.	http://www.ebi.ac.uk/clustr/	Kriventseva (2001) [32]
CSDBase	This is an interactive database which provides detailed information on proteins containing the so-called cold shock domain (CSD).	http://www.chemie.uni-marburg.de/~csdbase/	Michael et al. (2002) [33]
COGS	The CluSTr (Clusters of Swiss-Prot+TrEMBL proteins) database offers an automatic classification of Swiss-Prot + TrEMBL proteins into groups of related proteins.	http://www.ncbi.nlm.nih.gov/COG	Tatusov et al. (1997) [34]
DART	Drug Adverse Reaction Target A database for facilitating the search for drug adverse reaction target	http://xin.cz3.nus.edu.sg/group/drt/dart.asp	Web site
Database of Macromolecular Movements	Descriptions of protein and macromolecular motions, including movies.	http://bioinfo.mbb.yale.edu/MolMovDB	Echols et al. (2003) [35]
Decoys 'R' Us	Computer-generated protein conformations based on sequence data.	http://dd.stanford.edu/	Xia et al. (2000) [36]
DEF	Determination of expected protein folds - Prediction Server.	http://www.stepc.gr/~synaptic/def2.html	Reczko et al. (1997) [37]
DExH/D Family Database	This database is an attempt to gather information about the DExH/D protein family with particular emphasis on the biochemical and functional characteristics of these proteins.	http://www.helicase.net/dexhd/dbhome.htm	Jankowsky et al. (2000) [38]
DIP	Database of Interacting Proteins Experimentally-determined protein-protein interactions.	http://dip.doe-mbi.ucla.edu	Xenarios et al.(2000) [39]
DSDBASE	database on disulphide bonds in proteins that provides information on native disulphides and those which are stereochemically possible between pairs of residues in a protein.	http://www.ncbs.res.in/~faculty/mini/dsdbase/dsdbase.html	Sowdhamini et al. (1989) [40]
DSMM	ADatabase of Simulated Molecular Motions.	http://projects.eml.org/mcm/database/dsmm	Finocchiaro et al. (2003) [41]
DSMP	DATABASE OF STRUCTURAL MOTIFS IN PROTEINS representative protein data set derived using the PDB_ SELECT program.	http://www.cdfd.org.in/dsmp.html	Guruprasad et al.(2000) [42]
EMOTIF	Protein sequence motif determination and searches.	http://dna.Stanford.EDU/emotif/	Huang et al. (2001) [43]
ENZYME	Enzyme nomenclature database.	http://www.expasy.ch/enzyme/	Bairoch (2000) [44]
ESTHER	Esterases and alpha/beta hydrolase enzymes and relatives.	http://www.ensam.inra.fr/cholinesterase/	Cousin et al. (1996) [45]

Table 2. Protein databases and their URL (Continue).

DATABASE	Description	URL	Reference
EXProt	EXProt is a non-redundant protein database containing a selection of entries from genome annotation projects and public databases, aiming at including only proteins with an experimentally verified function.	http://www.cmbi.nl/exprot	Ursing et al. (2002) [46]
FSSP	FSSP stands for "Fold classification based on Structure-Structure alignment of Proteins".	http://www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html	Holm et al. (1996) [47]
FUNPEP	Low-complexity or compositionally-biased protein sequences.	http://www.cmbi.kun.nl/swift/FUNPEP/library/	Web site
GCRDB	G-Protein Coupled Receptor Database.	http://receptor.mgh.harvard.edu/GCRDBHOME.html	Kolakowski et al. (1994) [48]
Het-PDB Navi	Protein-small molecule interactions.	http://daisy.bio.nagoya-u.ac.jp/golab/hetpdbnavi.html	Web site
Histone Database	The Histone Database is a searchable, periodically updated collection of histone-fold containing sequences derived from sequence-similarity searches of public databases.	http://genome.nhgri.nih.gov/histones/	Sullivan et al. (2002) [49]
Homeobox Page	Information relevant to homeobox proteins, classification, and evolution.	http://www.biosci.ki.se/groups/tbu/homeo.html	Bürglin (1997) [50]
HSSP	Structural families and alignments; structurarlly-conserved regions and domain architecture.	http://swift.embl-heidelberg.de/hssp http://www.sander.ebi.ac.uk/hssp/	Dodge et al. (1998) [51]
HUGE	Large (>50 kDa) human proteins and cDNA sequences.	http://www.kazusa.or.jp/huge/	Kikuno et al. (2000, 2002) [52,53]
IDENTIFY	Stores motifs in the form o fregular expressions.	http://dna.stanford.edu/identify	Huang et al. (2001) [43]
IMGT	Integrated database specializing in protein sequences of immunoglobulins, T-cell receptors (LIGM-DB) and HLA molecules (HLA-DB).	http://imgt.cines.fr/	Robinson et al. (2003) [54]
INBASE	InBase is a curated database devoted to inteins (intervening protein sequence).	http://www.neb.com/neb/inteins.html	Perler (2002) [55]
INTERPRO	Protein families and domains.	http://www.ebi.ac.uk/interpro	Apweiler et al. (2001) [56]
iProClass	Annotated protein database with family, function, and structure information.	http://pir.georgetown.edu/iproclass/	Wu et al. (2001) [57, 58]
ISSD	Integrated sequence and structural information.	http://www.protein.bio.msu.su/issd/	Adzhubei et al. (1999) [59]
JenPep	Functional and quantitative thermodynamic data on peptide binding to immunological biomacromolecules.	http://www.jenner.ac.uk/Jenpep2	McSparron et al. (2003) [60]
KABAT	The Kabat database is a repository for protein and nucleotide data that are of immunological interest.	http://www.hgmp.mrc.ac.uk/Bioinformatics/Databases/kabatn-help.html	Johnson et al. (2001) [61]
KEGG	enzyme database & metabolic pathways.	http://www.genome.ad.jp/kegg/kegg2.html	Kanehisa et al. (2002) [62]
KinMutBase	Disease-causing protein kinase mutations.	http://www.uta.fi/imt/bioinfo/KinMutBase/	Stenberg et al. (2000) [63]
Klotho	Biochemical Compounds Declarative Database.	http://www.ibc.wustl.edu/klotho/ http://www.biocheminfo.org/klotho/	Web site
LIGAND	Database of Chemical Compounds and Reactions in Biological Pathways.	http://www.genome.ad.jp/dbget/ligand.html	Goto et al. (2002) [64]
Ligand Gated Ion channels	Ligand Gated Ion Channel Database contains entries of ligand-activated ion channel subunits.	http://www.pasteur.fr/units/neubiomol/LGIC.html	Le Novere et al. (2001) [65]
LOOP	Loop classification database.	http://www.bmm.icnet.uk/loop/	Oliva et al. (1997) [66]
LPFC	Library of protein family core structures	http://www-smi.stanford.edu/projects/helix/LPFC/	Schmidt et al. (1997) [67]
MDB	A tools for the design of metal binding sites in proteins.د	http://metallo.scripps.edu/current/raw.html	Roberts et al. (1995) [68]
MEROPS	Proteolytic enzymes (proteases/peptidases).	http://www.merops.ac.uk/	Rawlings et al. (2002) [69]
MetaFam	Integrated protein family information.	http://metafam.ahc.umn.edu/	Silverstein (2001) [70]
Metalloprotein Database and Browser	Metal-binding sites in metalloproteins.	http://metallo.scripps.edu/	Castagnetto et al. (2002) [71]
MIPS	A database for genomes and protein sequences.	http://www.mips.biochem.mpg.de/	Mewes et al. (2002) [72]
MITOP	Database for mitochondria-related genes, proteins and diseases.	http://mips.gsf.de/proj/medgen/mitop/	Scharfe et al. (2000).[73]
MMDB	All experimentally-determined three-dimensional structures, linked to NCBI Entrez.	http://www.ncbi.nlm.nih.gov/Structure/	Wang et al.(2002) [74] Chen et al. (2003) [75]
NetOGly	The NetOglyc server produces neural network predictions of mucin type GalNAc O-glycosylation sites in mammalian proteins.	http://www.cbs.dtu.dk/services/NetOGlyc/	Hansen et al.(1998) [76]
NPD	Nuclear Protein Database Proteins localized in the nucleus.	http://npd.hgu.mrc.ac.uk/	Dellaire et al. (2003) [77]

Table 2. Protein databases and their URL (Continue).

DATABASE	Description	URL	Reference
NRL_3D	NRL_3D is a sequence--structure database derived from the 3 dimensional structure of proteins deposited with the Brookhaven National Laboratory's Protein Data Bank.	http://www.psc.edu/general/software/packages/nrl_3d/nrl_3d.html	Pattabiraman et al. (1990) [78]
O-GLYCBASE	O- and C-linked glycosylation sites in proteins.	http://www.cbs.dtu.dk/databases/OGLYCBASE/	Gupta et al. (1999) [79]
OWL	OWL is a non-redundant composite of 4 publicly-available primary sources: SWISS-PROT, PIR (1-3), GenBank (translation) and NRL-3D.	http://www.bioinf.man.ac.uk/dbbrowser/OWL/	Bleasby et al. (1994) [80]
PALI	Phylogeny and alignment of homologous protein structures.	http://pauling.mbu.iisc.ernet.in/%7Epali	Gowri et al. (2003) [81]
PASS2	Structural motifs of protein superfamilies.	http://ncbs.res.in/%7Efaculty/mini/campass/pass.html	Mallika et al. (2002) [82]
PClass	Protein structure Classification database.	http://gene.stanford.edu/PClass/	Web site
PDB	Structure data determined by X-ray crystallography and NMR.	http://www.rcsb.org	Bernstein et al. (1977) [83] Berman et al. (2000) [84]
PDB-REPRDB	Representative protein chains, based on PDB entries	http://www.cbrc.jp/pdbreprdb/	Noguchi et al. (2000, 2001) [85,86]
PDBsum	Summary of key information on structures in PDB.	http://www.biochem.ucl.ac.uk/bsm/pdbsum	Laskowski (2001) [87]
PEDB	Sequences from prostate tissue and cell type-specific cDNA libraries.	http://www.pedb.org/	Hawkins et al. (1999) [88]
PEP	Predictions for Entire Proteomes Summarized analyses of protein sequences.	http://cubic.bioc.columbia.edu/pep/	Carter et al.(2003) [89]
Peptaibol	Peptaibol (antibiotic peptide) sequences.	http://www.cryst.bbk.ac.uk/peptaibol/welcome.html	Chugh et al. (2001) [90]
Pfam	ultiple sequence alignments and hidden Markov models of common protein domains.	http://www.sanger.ac.uk/Pfam	Sonnhammer et al. (1998) [91]
PhosphoBase	Protein phosphorylation sites.	http://www.cbs.dtu.dk/databases/PhosphoBase/	Kreegipuu et al. (1999) [92]
PIMA	Thermodynamics data for wild-type and mutant proteins.	http://dot.imgen.bcm.tmc.edu:9331/seq-searcj/[rpteom-search.html	Web site
PIR-PSD	Comprehensive, annotated, non-redundant protein sequence databases.	http://www-nbrf.georgetown.edu/pirwww/pirhome.shtml	Barker et al.(1991, 2000) [93,94]
PIR-ALN	A curated database of protein sequence alignments.	http://pir.georgetown.edu/pirwww/dbinfo/piraln.html	Srinivasarao et al. (1999) [95]
PIR-NREF	The PIR-NREF is a Non-redundant REFerence protein database designed to provide a timely and comprehensive collection of all protein sequence data.	http://pir.georgetown.edu/pirwww/search/pirnref.shtml	Wu et al (2003) [96]
PMD	Compilation of protein mutant data.	http://pmd.ddbj.nig.ac.jp/	Kawabata et al. (1999) [97]
PRESAGE	Protein structures with experimental and predictive annotations.	http://presage.berkeley.edu/	Brenner et al. (1999) [98]
PRF	Protein Research Foundation Sequence Database.	http://www.genome.ad.jp/htbin/www_bfind?prf	Web site
PRINTS	A compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family.	http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/	Attwood et al. (1999) [10,11]
PROCLASS	Protein families defined by PIR superfamilies and PROSITE patterns.	http://www-nbrf.georgetown.edu/gfserver/proclass.html	Wu et al. (1996) [99]
ProDDO	PROTEIN DATABASE OF DISORDER.	http://bonsai.ims.u-tokyo.ac.jp/~klsim/database.html	Sim et al. (2001) [100]
PRODOM	Protein domain families.	http://protein.toulouse.inra.fr/prodom.html	Corpet et al (1998,1999,2000) [101-103]
PROMISE	Prosthetic centers and metal ions in protein active sites.	http://metallo.scripps.edu/PROMISE/MAIN.html	Degtyarenko et al. (1999) [104]
ProNIT	Thermodynamic database for Protein-Nucleic Acid interaction.	http://www.rtc.riken.go.jp/jouhou/pronit/pronit_search.html	Prabakaran et al. (2001) [105]
Prosite	Biologically-significant protein patterns and profiles.	http://www.expasy.ch/prosite	Hofmann et al. (1999 [106])
Proteome Analysis Database	Online application of InterPro and cluSTr for the functional classification of proteins in whole genomes.	http://www.ebi.ac.uk/proteome/	Pruess et al. (2003) [107]
Proteome BioKnowledge Library	Model organism, pathogen, and mammalian proteomes.	http://www.proteome.com/	Web site
PROTFAM	A protein sequence homology database.	http://www.mips.biochem.mpg.de/desc/protfam/	Web site
ProTherm	Thermodynamic data for wild-type and mutant proteins.	http://www.rtc.riken.go.jp/jouhou/protherm/protherm.html	Gromiha et al. (2002) [108]
ProtoMap	Automated hierarchical classification of SWISS-PROT proteins.	http://protomap.cornell.edu	Yona et al. (2000) [109]
PSORT	Is a computer program for the prediction of protein localization sites in cells.	http://psort.nibb.ac.jp	Yona et al. (2000) [109]
QUEST	2D Protein database.	http://siva.cshl.org/index.html	Web site

Table 2. Protein databases and their URL (Continue).

DATABASE	Description	URL	Reference
REBASE	Restriction enzymes and associated methylases.	http://rebase.neb.com/rebase/rebase.html	Roberts et al. (2003) [110]
RIBONUCLEASE P	The RNase P Database is a compilation of RNase P sequences, sequence alignments, secondary structures, three-dimensional models, and accessory information.	http://www.mbio.ncsu.edu/RNaseP/home.html	Brown (1999) [111]
RESID	Protein structure modifications.	http://www-nbrf.georgetown.edu/pirwww/dbinfo/resid.html	Garavelli et al. (2001, 2003) [112,113]
SARF	Spatial Arrangement of backbone Fragments.	http://123d.ncifcrf.gov/sarf2.html	Alexandrov et al. (1996) [114]
SBASE	Protein domain sequences and tools.	http://www3.icgeb.trieste.it/~sbasesrv/	Vlahovicek et al. (2002, 2003); [5,6] Pongor et al. (1993) [115]
SCOP	Familial and structural protein relationships.	http://scop.mrc-lmb.cam.ac.uk/scop	Murzin et al. (1995) [116]
SDAP	Sequences, structures, and IgE epitopes of allergenic proteins.	http://fermi.utmb.edu/SDAP	Ivanciuc et al. (2003) [117]
SignalPWeb server	The SignalP World Wide Web server predicts the presence and location of signal peptide cleavage sites in amino acid sequences from different organisms.	http://www.cbs.dtu.dk/services/SignalP/	Nielsen et al. (1997) [118]
SLOOP	Classification of protein loops.	http://www-cryst.bioc.cam.ac.uk/~sloop/	Web site
SMART	Simple Modular Architecture Research Tool.	http://smart.embl-heidelberg.de	Schultz et al.(1998, 2000) [119,120]
SRPDB	SRPDB (Signal Recognition Particle Database) Provides Aligned, Annotated and Phylogenetically Ordered Sequences Related to Structure and Function of SRP.	http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html	Nagai et al. (2003) [121]
SUPERFAMILY	Assignments of proteins to structural superfamilies.	http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY/	Gough et al. (2001) [122]
SUPFAM	Grouping of sequence families into superfamilies.	http://pauling.mbu.iisc.ernet.in/%7Esupfam	Pandit et al. (2002) [123]
SWISS-2DPAGE	Annotated two-dimensional polyacrylamide gel electrophoresis database.	http://www.expasy.org/ch2d/	Hoogland et al. (2000) [124]
SWISS-3DIMAGE	Images of crystallized proteins.	http://us.expasy.org/sw3d/	Peitsch et al. (1995) [125]
SWISS-PROT/TrEMBL	Curated protein sequences.	http://www.expasy.ch/sprot	Boeckmann et al. (2003) [126]
SYSTERS	SYSTERS (SYSTEmatic Re-Searching) Protein Family Database.	http://systers.molgen.mpg.de/	Krause et al. (2000) [127]
ooTFD	object-oriented Transcription Factors Database.	http://www.ifti.org/cgi-bin/ifti/ootfd.pl	Ghosh et al. (2000) [128]
The Pharmacogenomics and Pharmacogenetics Knowledge Base	Variation in drug response based on human variation.	http://www.pharmgkb.org/	Hewett et al. (2002) [129]
TIGRFAMs	Functional identification of proteins.	http://www.tigr.org/TIGRFAMs/	Haft et al. (2001) [130]
TMBASE	A Database of Membrane Spanning Protein Segments.	http://www.ch.embnet.org/software/tmbase/TMBASE_doc.html	Hofmann et al. (1993) [131]
Wnt Database	Wnt proteins and phenotypes.	http://www.stanford.edu/~rnusse/wntwindow.html	Web site
YPD	Yeast Protein Database.	http://www.proteome.com/YPDhome.html	Hodges et al. (1999) [132]

ACKNOWLEDGEMENT

We hereby like to thank Dr. Sanati and Dr. Yakhchali for their kind support to the work. We are also highly indebted to Dr. A. M. Sobhani for carefully reading the manuscript and for his critical comments to the article. This research was partly supported by the grant no: 521/4/551 from the research council of the University of Tehran.

REFERENCES

1. Sanger F, Tuppy H. The amino-acid sequence in the phenylalanyl chain of insulin 2. The investigation of peptides from enzymic hydrolysates. *Biochem J* 1951;**49:**481-490
2. Dayhoff MO. Atlas of Protein Sequence and Structure. Vol 5 1972.
3. Schuler GD, Epstein JA, Ohkawa H, Kans JA Entrez: molecular biology database and retrieval system. *Methods Enzymol* 1996;**266:**141-62.
4. Etzold T, Ulyanov A, Argos P. SRS: information retrieval system for molecular biology data banks. *Methods Enzymol* 1996;**266:**114-28.
5. Vlahovicek K, Kajan L, Murvai J, Hegedus Z, Pongor S. The SBASE domain sequence library, release 10: domain architecture prediction. *Nucleic Acids Res* 2003;**31(1):**403-5.
6. Vlahovicek K, Murvai J, Barta E, Pongor S. The SBASE protein domain library, release 9.0: an online resource for protein domain identification. *Nucleic Acids Res* 2002;**30(1):**273-5.
7. Attwood TK. The role of pattern databases in sequence analysis. *Brief Bioinform* 2000;**1(1):**45-59.
8. Attwood TK. Introduction to Bioinformatics. Addison Wesley Longman Limited; 1999 Chapter 3, 35-67.
9. Attwood TK, Beck ME, Flower DR, Scordis P, Selley J, The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Res* 1998;**26(1):**304-308.
10. Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, Scordis P, et al. "PRINTS prepares for the new millennium". *Nucleic Acids Res* 1999;**27(1):**220-225.
11. Attwood TK, Croning MD, Flower DR, Lewis AP, Mabey JE, Scordis P, Selley JN, Wright W. PRINTS-S: the database formerly known as PRINTS. *Nucleic Acids Res* 2000;**28(1):**225-7.
12. Attwood TK, Bradley P, Flower DR, Gaulton A, Maudling N, Mitchell A, et al. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res* 2003;**31(1):**400-402.
13. Kabsch W, Sander C. Dictionary of the secondary structure of proteins. *Biopolymers* 1983;**22:**2577-2637.
14. Pascarella S, Argos P. A data bank merging related protein structures and sequences. *Protein Eng* 1992 Mar;5(2):121-37.
15. Kelley LA, MacCallum RM, Sternberg MJ. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J Mol Biol* 2000;**299(2):**499-520.
16. Kawashima S, Ogata H, Kanehisa M; AAindex: amino acid index database. *Nucleic Acids Res* 1999;**27:**368-369.
17. Szymanski M, Deniziak MA, Barciszewski J. Aminoacyl-tRNA synthetases database. *Nucleic Acids Res* 2000;**29:**288-290.
18. Szymanski M, Deniziak MA, Barciszewski J. Aminoacyl-tRNA synthetases database. *Nucleic Acids Res* 2001;**29:**288-290.
19. Tobes R, Ramos JL. AraC-XylS database: a family of positive transcriptional regulators in bacteria. *Nucleic Acids Res* 2002;**30(1):**318-21.
20. Afonnikov DA, Valuev VP, Kashinskaya JO, Orlov YL. The ASPD Database on synthetic peptides. *Computational Technologies* 2000;**5:**75-78.
21. Valuev VP, Afonnikov DA, Petrenko O, Beylina AG, Lokhova IV, Grigorovich DA, et al. The database ASPD on experiments with application of phage display technique.: Proceedings of the IId International Conference on Bioinformatics of Genome Regulation and Structure, Novosibirsk, 2000 eds. Kolchanov NA et al., vol.II, p. 160-163.
22. Bader GD, Donaldson I, Wolting C, Ouellette BF, Pawson T, Hogue CW. BIND: The Biomolecular Interaction Database. *Nucleic Acids Res* 2001;**29(1):**242-245.
23. Bader GD, Betel D, Hogue CW. BIND: The Biomolecular INteration Network Databae. *Nucleic Acids Res* 2003;**31(1):**248-50.
24. Henikoff S, Henikoff JG. Automated assembly of protein blocks for database searching. *Nucleic Acids Res* 1991;**19(23):**6565-72.
25. Henikoff JG, Henikoff S. Blocks database and its applications. *Methods Enzymol*1996;**266:**88-105.
26. Henikoff JG, Greene EA, Pietrokovski S, Henikoff S, Increased coverage of protein families with the blocks database servers, *Nucleic Acids Res* 2000;**28:**228-230.
27. Pietrokovski S, Henikoff JG, Henikoff S. The Blocks database--a system for protein classification. *Nucleic Acids Res* 1996;**24(1):**197-200.
28. Seavey BR, Farr EA, Westler WM, Markley JL. A Relational Database for Sequence-Specific Protein NMR Data. *J Biomol NMR* 1991;**1:**217-236.
29. Orengo CA, Michi AD, Jones S, Jones DT, Swindells MB, Thornton JM, CATH: A Hierarchic Classification of Protein Domain Structures. *Structure* 1997;**5:**1093-1108.
30. Allen FH. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallogr B* 2002;**58:**380-8.
31. Marchler-BA, Anderson JB, DeWeese-SC, Fedorova ND, Geer LY, He S, et al.. CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res* 2003;**31:**383-387.
32. Kriventseva EV, FleischmannW, Apweiler R, CluSTr: a database of Clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res* 2001;**29:**33-36.
33. Michael H, Weber W, Fricke I, Doll N, Marahiel MA CSDBase: An Interactive Database for Cold Shock Domain Containing Proteins and the Bacterial Cold Shock Response. *Nucleic Acids Res* 2002;**30:**375-8.
34. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997;**278:**631-7.
35. Echols N, Milburn D, Gerstein M. MolMovDB: analysis and visualization of conformational change and structural flexibility *Nucleic Acids Res* 2003;**31:**478-82.
36. Xia Y, Huang ES, Levitt M, Samudrala R. Ab initio construction of protein tertiary structures using a hierarchical approach. *J Mol Biol* 2000;**300:**171-185.
37. Reczko M, Karras D, Bohr H, An update of the DEF database of protein fold class predictions. *Nucleic Acids Res* 1997;**25(1):**235
38. Jankowsky E, Jankowsky A. The DExH/D protein database. *Nucleic Acids Res* 2000;**28:**333-334.
39. Xenarios I, Rice DW, Salwinski L, Baron MK, Marcotte EM, Eisenberg D. DIP: the database of interacting proteins. *Nucleic Acids Res* 2000;**28(1):**289-91.
40. Sowdhamini R, SrinivasanN, Shoichet B, Santi DV, Ramakrishnan C, Balaram P. Stereochemical modelling of disulfide bridges: Criteria for introduction into proteins by site-directed mutagenesis. *Prot Eng* 1989;**3:**95-103.
41. Finocchiaro G, Ting Wang2, Rene Hoffmann2, Aitor Gonzalez1,2 andRebecca C. Wade DSMM: a Database of Simulated Molecular Motions. *Nucleic Acids Res* 2003;**31:**456–457
42. Guruprasad K, Prasad MS, Kumar GR. Database of structural motifs in proteins. *Bioinformatics* 2000;**16(4):**372-5.
43. Huang JY, Brutlag DL. The EMOTIF database. *Nucleic Acids Res* 2001;**29(1):**202-4.

44. Bairoch A. The ENZYME database in 2000. *Nucleic Acids Res* 2000;**28:**304-305.
45. Cousin X, Hotelier T, Lievin P, Toutant JP, Chatonnet A. A cholinesterase genes server (ESTHER): a database of cholinesterase-related sequences for multiple alignments, phylogenetic relationships, mutations and structural data retrieval. *Nucleic Acids Res* 1996;**24(1):**132-6.
46. Ursing B M, Enckevort V, Frank H J, Leunissen JAM, Siezen RJ. EXProt: a database for proteins with an experimentally verified function. *Nucleic Acids Res* 2002;**30:**50-51.
47. Holm L, Sander C. Mapping the protein universe. *Science* 1996;**273:**595-602.
48. Kolakowski LF. GCRDb: a G-protein-coupled receptor database. *Receptors Channels* 1994;**2(1):**1-7.
49. Sullivan S, Sink DW, Trout KL, Makalowska I, Taylor PM, Baxevanis AD, Landsman D. The Histone Database. *Nucleic Acids Res* 2002;**30(1):**341-2.
50. Bürglin TR. Analysis of TALE superclass homeobox genes (MEIS, PBC, KNOX, Iroquois, TGIF) reveals a novel domain conserved between plants and animals. *Nucleic Acids Res* 1997;**25:**4173-4180.
51. Dodge C, Schneider R, Sander C. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res* 1998;**26(1):**313-5.
52. Kikuno R, Nagase T, Suyama M, Waki M, Hirosawa M, Ohara O. HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res* 2000;**28(1):**331-2.
53. Kikuno R, Nagase T, Waki M, Ohara O. HUGE: a database for human large proteins identified in the Kazusa cDNA sequencing project. *Nucleic Acids Res* 2002;**30(1):**166-8.
54. Robinson J, Waller MJ, Parham P, de Groot N, Bontrop R, Kennedy LJ, Stoehr P, Marsh SG. IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Res* 2003;**31(1):**311-4
55. Perler FB, InBase, the Intein Database. *Nucleic Acids Res* 2002;**30:**383-384.
56. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, et al. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res* 2001;**29(1):**37-40.
57. Wu CH, Shivakumar S, Huang H. ProClass Protein Family Database. *Nucleic Acids Res* 1999;**27(1):**272-4.
58. Wu C, Xiao C, Hou Z, Huang H, Barker WC. iProclass: an integrated, comprehensive and annotated protein classification database. *Nucleic Acids Res* 2001;**29:**52-54.
59. Adzhubei IA, Adzhubei AA. Taxonomic range extended. *Nucleic Acids Res* 1999;**27:**268-271.
60. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, Flower DR. JenPep: a novel computational information resource for immunobiology and vaccinology. *J Chem Inf Comput Sci* 2003;**43(4):**1276-87
61. Johnson G, Wu TT. Kabat Database and its applications: future directions. *Nucleic Acids Res* 2001;**29:**205-206.
62. Kanehisa M, Goto S, Kawashima S, Nakaya A. The KEGG databases at GenomeNet. *Nucleic Acids Res* 2002;**30(1)**:42-6.
63. Stenberg KA, Riikonen PT, Vihinen M. KinMutBase, a database of human disease-causing protein kinase mutations. *Nucleic Acids Res* 2000;**28(1):**369-71.
64. Goto S, Okuno Y, Hattori M, Nishioka T, Kanehisa M. LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res* 2002;**30:**402-404.
65. Le Novere N, Changeux JP. LGICdb: the ligand-gated ion channel database. *Nucleic Acids Res* 2001;**29:**294-295.
66. Oliva B, Bates PA, Querol E, Aviles FX, Sternberg MJ. An automated classification of the structure of protein loops. *J Mol Biol* 1997;**266(4):**814-30.
67. Schmidt R, Gerstein M, Altman RB. LPFC: an Internet library of protein family core structures. *Protein Sci* 1997;**6(1):**246-8.
68. Roberts VA, Getzoff ED. Metalloantibody design. *FASEB J* 1995;**9:**94-100.
69. Rawlings ND, O'Brien EA, Barrett AJ, MEROPS: The protease database. *Nucleic Acids Res* 2002;**30:**343-346.
70. Silverstein KA, Shoop E, Johnson JE, Ernest F, Retzel EF. MetaFam: a unified classification of protein families. I. Overview and statistics. *Bioinformatics* 2001;**17:**249-261.
71. Castagnetto JM, Hennessy SW, Roberts VA, Getzoff ED, Tainer, JA, Pique ME. MDB: the Metalloprotein Database and Browser at The Scripps Research Institute. *Nucleic Acids Res* 2002;**30(1):**379-382.
72. Mewes, H. W., D. Frishman, U. Guldener, G. Mannhaupt, K. Mayer, M. Mokrejs, B. Morgenstern, M. Munsterkotter, S. Rudd, B. Weil. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res* 2002;**30:**31-34.
73. Scharfe C, Zaccaria P, Hoertnagel K, Jaksch M, Klopstock T, Dembowski M, Lill R, Prokisch H, Gerbitz KD, Neupert W, Mewes HW, Meitinger T. MITOP, the mitochondrial proteome database. *Nucleic Acids Res* 2000;**28(1):**155-158.
74. Wang Y, Anderson JB, Chen J, Geer LY, He S, Hurwitz DI, et al. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res* 2002;**30(1):**249-252.
75. Chen J, Anderson JB, DeWeese-Scott C, Fedorova ND, Geer LY, He S, et al. MMDB: Entrez's 3D-structure database. *Nucleic Acids Res* 2003;**31(1):**474-7.
76. Hansen JE, Lund O, Tolstrup N, Gooley AA, Williams KL, Brunak S. NetOglyc: prediction of mucin type O-glycosylation sites based on sequence context and surfac4e accessibility. *Glycoconj J* 1998;**15(2):**115-30.
77. Dellaire G, Farrall R, Bickmore WA. The Nuclear Protein Database (NPD): sub-nuclear localisation and functional annotation of the nuclear proteome. *Nucleic Acids Res* 2003;**31(1):**328-330.
78. Pattabiraman N, Namboodiri K, Lowrey A, Gaber BP. NRL-3D: a sequence-structure database derived from the protein data bank (PDB) and searchable within the PIR environment. *Protein Seq Data Anal* 1990;**3(5):**387-405.
79. Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE, O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. *Nucleic Acids Res* 1999;**27:**370-372.
80. Bleasby AJ, Akrigg D, Attwood TK. OWL - a non-redundant composite protein sequence database. *Nucleic Acids Res* 1994;**22(17):**3574-7.
81. Gowri VS, Pandit SB, Karthik PS, Srinivasan N, Balaji S. Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database. *Nucleic Acids Res* 2003;**31:**486-488.
82. Mallika V, Bhaduri A, Sowdhamini R. PASS2: a semi-automated database of protein alignments organised as structural superfamilies. *Nucleic Acids Res* 2002;**30(1):**284-8.
83. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur J Biochem* 1977;**80:**319-324.
84. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. *Nucleic Acids Res* 2000;**28:**235-242.
85. Noguchi T, Onizuka K, Ando M, Matsuda H, Akiyama Y. Quick Selection of Representative Protein Chain Sets Based on Customizable Requirements. *Bioinformatics* 2000;**16(6):**520-526.
86. Noguchi T, Matsuda H, Akiyama Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB). *Nucleic Acids Res* 2001;**29(1):**219-220.
87. Laskowski RA. PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res* 2001;**29:**221-222.
88. Hawkins V, Doll D, Bumgarner R, Smith T, Abajian C, Hood L, Nelson PS. PEDB: the Prostate Expression Database. *Nucleic Acids Res* 1999;**27(1):**204-8.

89. Carter P, Liu J, Rost B. PEP: Predictions for Entire Proteomes. *Nucl Acid Res* 2003;**31:**410-413.
90. Chugh JK, Wallace BA. Peptaibols: models for ion channels. *Biochem Soc Trans* 2001;**29(Pt 4):**565-70.
91. Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 1998;**26(1):**320-2.
92. Kreegipuu A, Blom N, Brunak S. PhosphoBase: a database of phosphorylation sites: release 2.0. *Nucleic Acids Res* 1999;**27(1):**237-239
93. Barker WC, George DG, Hunt LT, Garavelli JS. The PIR protein sequence database. *Nucleic Acids Res* 1991;**19 Suppl:**2231-36.
94. Barker WC, Garavelli JS, Huang H, McGarvey PB, Orcutt BC, Srinivasarao GY, et al. The Protein Information Resource (PIR). *Nucleic Acids Res* 2000;**28:**41-44.
95. Srinivasarao GY, Yeh LS, Marzec CR, Orcutt BC, Barker WC. PIR-ALN: a database of protein sequence alignments. *Bioinformatics* 1999;**15(5):**382-90.
96. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, Chen Y, et al. The Protein Information Resource. *Nucleic Acids Res* 2003;**31(1):**345-7.
97. Kawabata T, Ota M, Nishikawa K. The protein mutant database. *Nucleic Acids Res* 1999;**27:**355-357.
98. Brenner SE, Barken D, Levitt M. The PRESAGE database for structural genomics. *Nucleic Acids Res* 1999;**27:**251-253.
99. Wu CH, Zhao S, Chen HL. A protein class database organized with ProSite protein groups and PIR superfamilies. *J Comput Biol* 1996;**3(4):**547-61.
100. Sim KL, Tomoyuki Uchida, and Satoru Miyano ProDDO: A database of disordered proteins from the Protein Data Bank (PDB). *Bioinformatics* 2001;**17:**379-380.
101. Corpet F, Gouzy J, Kahn D. The ProDom database of protein domain families. *Nucleic Acids Res* 1998;**26:**323-326.
102. Corpet F, Gouzy J, Kahn D. Recent improvements of the ProDom database of protein domain families. *Nucleic Acids Res* 1999;**27:**263-267.
103. Corpet F, Servant F, Gouzy J, Kahn D. ProDom and ProDom-CG: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Res* 2000;**28:**267-269.
104. Degtyarenko KN, North AC, Findlay JB. PROMISE: a database of bioinorganic motifs. *Nucleic Acids Res* 1999;**27:**233-236.
105. Prabakaran P, Jianghong A, Gromiha MM, Selvaraj S, Uedaira H, Kono H, et al. Thermodynamic database for protein–nucleic acid interactions (ProNIT). *Bioinformatics* 2001;**17:**1027-1034
106. Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. *Nucleic Acids Res*. 1999;**27:**215-219.
107. Pruess M, Fleischmann W, Kanapin A, Karavidopoulou Y, Kersey P, Kriventseva E, et al. The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. *Nucleic Acids Res* 2003;**31(1):**414-417.
108. Gromiha M, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A. ProTherm: Thermodynamic Database for Proteins and Mutants: Developments in Version 3.0. *Nucleic Acids Res* 2002;**31:**301-302.
109. Yona G, Linial N, Linial M, ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res* 2000;**28(1):**49-55.
110. Roberts RJ, Vincze T, Posfai J, Macelis D. REBASE - Restriction enzymes and methylases. *Nucleic Acids Res* 2003;**31:**418-420.
111. Brown JW. The Ribonuclease P Database. *Nucleic Acids Res* 1999;**27:**314.
112. Garavelli JS, Zhenglin Hou, Nagarajan Pattabiraman, and Robert M. Stephens The RESID Database of protein structure modifications and the NRL-3D Sequence-Structure Database. *Nucleic Acids Res* 2001;**29(1):**199-201.
113. Garavelli JS. The RESID Database of Protein Modifications: 2003 developments. *Nucleic Acids Res* 2003;**31(1):**499-501.
114. Alexandrov NN. SARFing the PDB. *Protein Eng* 1996;**9:**727-732.
115. Pongor S, Skerl V, Cserzo M, Hatsagi Z, Simon G, Bevilacqua V. The SBASE domain library: a collection of annotated protein segments. *Protein Eng* 1993;**6(4):**391-5.
116. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;**247:**536-540.
117. Ivanciuc O, Schein CH, Braun W, SDAP Database and Computational Tools for Allergenic Proteins. *Nucleic Acids Res* 2003;**31:**359-362.
118. Nielsen H, Engelbrecht J, Brunak S, von Heijne G. A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Int J Neural Syst* 1997;**8(5-6):**581-99.
119. Schultz J, Milpetz F, Bork P, Ponting CP. SMART, a simple modular architecture research tool: identification of signaling domains. *Proc Natl Acad Sci* 1998;**95(11):**5857-64.
120. Schultz J, Copley RR, Doerks T, Ponting CP, Bork P. SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Res* 2000;**28(1):**231-4.
121. Nagai K, Oubridge C, Kuglstatter A, Menichelli E, Isel C, Jovine L. Structure, function and evolution of the signal recognition particle. *EMBO J* 2003;**22:**3479-3485.
122. Gough J, Karplus K, Hughey R, Chothia C, Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure. *J Mol Biol* 2001;**313(4):**903-919.
123. Pandit SB, Gosar D, Abhiman S, Sujatha S, Dixit SS, Mhatre NS. SUPFAM - Database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: Implications for structural genomics and function annotation in genomes. *Nucleic Acids Res* 2002;**30:**289-293.
124. Hoogland C, Sanchez JC, Tonella L, Binz PA, Bairoch A, Hochstrasser DF, Appel RD. The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res* 2000;**28:**286-288.
125. Peitsch MC, Stampf DR, Wells TNC, Sussman JL. The Swiss-3DImage collection and PDB-Browser on the World-Wide Web. *Trends Biochem Sci* 1995;**20:**82-84.
126. Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL. *Nucleic Acids Res* 2003;**31:**365-370.
127. Krause A, Stoye J, Vingron M. The SYSTERS protein sequence cluster set. *Nucleic Acids Res* 2000;**28(1):**270-2.
128. Ghosh D, Object-oriented Transcription Factors Database (ooTFD). *Nucleic Acids Res* 2000;**28:**308-310.
129. Hewett M, Oliver DE, Rubin DL, Easton KL, Stuart JM, Altman RB, Klein TE. PharmGKB: the Pharmacogenetics Knowledge Base. *Nucleic Acids Res* 2002;**30(1):**163-5.
130. Haft DH, Loftus BJ, Richardson DL, Yang F, Eisen JA, Paulsen IT, White O. TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Res* 2001;**29(1):**41-3.
131. Hofmann K, Stoffel W. TMBASE - A database of membrane spanning protein segments. *Biol Chem Hoppe-Seyler* 1993;**374:**166.
132. Hodges PE, McKee AH, Davis BP, Payne WE, Garrels JI. The Yeast Proteome Database (YPD): a model for the organization and presentation of genome-wide functional data. *Nucleic Acids Res*, 1999;**27(1):**69-73.

Address correspondence to: Zarrin Minuchehr, Ph.D. National Institute for Genetic Engineering & Biotechnology, P.O.Box: 14155-6343, Tehran, Iran
E-mail: minuchhr@nrcgeb.ac.ir