A Survey of Online Protein Database For Research In Bioinformatics and Computational Biology

Babasaheb .S. Satpute[1], Dr. Raghav Yadav[2]
[1, 2] Dept of Computer Science & I.T
[1, 2] SIET, SHUATS, Allahabad, India

Abstract- *one of the vital parts of today's modern biology, computational biology and bioinformatics is Protein databases. Lot of experiments and projects are being carried out and are producing humongous amount of data for protein structures, functions, and predominantly sequences. Many a time searching protein databases the first step in the study of a novel protein. We can obtain lot of useful information by comparing two proteins or two different protein families, compared to study of a single protein. Secondary databases which are obtained from experimental databases are also available in abundance and are being used by researchers in large number across the world. Those databases can be used to do predictions and classification purpose. It helps to use more than one database in order to understand or learn the functioning and structure of a protein. There are many databases which are known abundantly by researchers but are not being used to their full capacity. Here we present the review of some of the online protein databases available on the Internet so that it can help researchers in computational biology and bioinformatics to start their research on proteins.*

Keywords- Bioinformatics, Biological Databases, Protein Analysis, Protein Modeling

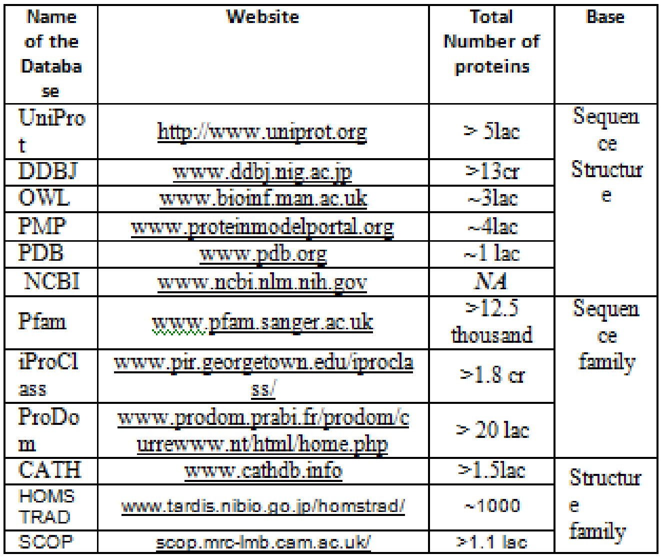
I. INTRODUCTION

Databases pertaining to proteins are of immense importance to the field or bioinformatics and computational biology predominantly to the researchers working on protein related problems like classification of proteins or predicting the family of the protein. Lot of experiments and projects using computers, internet and also by using wet labs are being carried out and they are producing humongous amount of data related to proteins and their sequences and functions. It becomes very difficult to manage, handle and store such data by using traditional methods. So it becomes imperative to use computers and computer database may be like RDBMS to store this humongous data. Many organizations have created their database and they are making those available online on the Internet so that those databases can be utilized by researchers across the globe for doing protein related research and the society by large can be benefitted from it. Many a time searching protein databases the first step in the study of a novel protein. If you don't want to miss some important information about an unknown protein or if you don't want do unnecessarily do the experiment repeatedly, you should do the extensive search on the protein databases. In order to obtain extensive useful information we need to compare proteins rather than studying a single protein, which can provide much useful information about their relation may be within the genome or may be across various species. Secondary databases which are obtained from experimental databases are also available in abundance and are being used by researchers in large number across the world. There are many databases which are known abundantly by researchers but are not being used to their full capacity.

Internet is proving to be a boon for the protein databases and it is empowering protein databases to a very large extent. It becomes very easy and economical to manage, maintain, and modify the online databases compared to traditional databases maintained using CD or hard drive. With the advent of online databases powered by Internet the Researchers with inadequate resources can do their research as they need not develop and maintain their own databases. Especially a large number of tiny databases on particular kind of proteins like EF-Hand Calcium-Binding Proteins Data Library (http://structbio.vanderbilt.edu/cabp_database/), are easily within reach. They also provide a very easy to use GUI for the researchers or users to use. We can directly search about a particular protein by using their own search engine or search facility which is easy to use. We can also submit or deposit our protein data or our findings to their repository.

There are many databases which are known abundantly by researchers but are not being used to their full capacity. Here we present the review of few of the online databases pertaining to proteins available on the Internet so that it can help researchers in computational biology and bioinformatics to start their research on proteins. Here we discuss various databases which emphasizes on structure, sequence, and family. We also assess the pros and cons of those databases. We have provided the Internet address of some of the famous online protein databases in table 1.

Table 1. Some of renowned protein databases



Name of the Database	Website	Total Number of proteins	Base
UniProt	http://www.uniprot.org	> 5lac	Sequence Structure
DDBJ	www.ddbj.nig.ac.jp	>13cr	
OWL	www.bioinf.man.ac.uk	~3lac	
PMP	www.proteinmodelportal.org	~4lac	
PDB	www.pdb.org	~1 lac	
NCBI	www.ncbi.nlm.nih.gov	*NA*	
Pfam	www.pfam.sanger.ac.uk	>12.5 thousand	Sequence family
iProClass	www.pir.georgetown.edu/iproclass/	>1.8 cr	
ProDom	www.prodom.prabi.fr/prodom/currewww.nt/html/home.php	> 20 lac	
CATH	www.cathdb.info	>1.5lac	Structure family
HOMSTRAD	www.tardis.nibio.go.jp/homstrad/	~1000	
SCOP	scop.mrc-lmb.cam.ac.uk/	>1.1 lac	

II. DATABASES BASED ON PROTEIN SEQUENCES

Because of the Human Genome Project and various other projects novel protein sequences are being generated at exponential rate. Those proteins are being viewed as and are becoming the extraordinary resource of information and are indeed very useful. What scientists or various research and government need to do is to preserve the data and sequences generated from those projects. And that's what exactly is done by various organizations and their databases comprises protein sequences with varied objectives. Out of many protein databases based on sequences UniProt is preferred by largest number of people and researchers. It has more detailed information compared to its counterparts. It also has very less redundancy. UniProtKB consists of 3 major parts: (1) Protein knowledgebase, consisting of Swiss-Prot database which is annotated by hand or mannually and TrEMBL which gets annotated automatically; (2) A cluster used for quickly searching the similarity among the sequences called UniRef and (3) A documentation for tracking various protein sequences and their identifiers called UniParc.

Notwithstanding Swiss-Prot and TrEMBL, UniProtKB incorporates data from Protein Succession Database (PSD) in the Protein Recognizable proof Asset, that constructs an entire and non-repetitive database from various protein and nucleic acids arrangement databases along with bibliographic and explained data. The National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) additionally gives wealthy data & various helpful apparatuses for protein groupings. For instance, the BLAST seek uses nr protein database (Altschul et al., 1997). It incorporates sections from the non-repetitive GenBank (Benson et al., 1999) interpretations, the Protein Data Bank (PDB), PIR, UniProt, , Protein Research Foundation (PRF) in Japan. Just sections with completely indistinguishable successions are consolidated.

The greater part of the protein sequence databases have an arrangement seek instrument and cross-references to passages of other protein and quality databases. Many succession databases, for example, UniProt, additionally give content seeking utilizing, for example, protein names or catchphrases. To contemplate another protein, the creator suggests first playing out a grouping seek utilizing BLAST in nr if the protein succession is accessible. The pursuit frequently gives passage names in the protein databases incorporated into nr. A homologous protein will be bit if the protein is found to be absent in nr, that could regularly prompt some valuable data, for example, the capacity of the inquiry protein. On the off chance that the arrangement of the question protein is inaccessible, completing a content hunt in UniProt more often than not recognizes the sequence. UniProt is likely the place to get the most data about a sequence on the off chance that it could be seen in UniProt. In any case, few extra data might be seen by checking different databases based on protein sequences. For instance, the Kyoto Encyclopedia of Genes and Genomes (KEGG; Ogata et al., 1999) comments on few quality passages with data about metabolic and administrative pathways. Someone could likewise ponder proteins in light of quality models (anticipated protein arrangements) from numerous species-particular genome assets.Despite the fact that anticipated groupings created by computational quality discovering devices in these assets may contain blunders, an expansive number of proteins are secured and are frequently sufficiently dependable to give valuable data. At the point when the protein of intrigue is from an animal categories that isn't secured by any of these databases, it is likely that some data can be recovered from its homolog of a model life form in one of the databases.

As a curated databases pertaining to protein sequences, UniProt offers an entrance to an extensive variety of explanations, covering zones, for example, work, family, area parsing, post-translational changes, & variations. The Uniform Resource Locator for the UniProt is www.uniprot.org. To find the information about a protein Human vitronectin in UniProt section, you could look either the passage name (VTNC_HUMAN) or the unique number for accessing (P04004) got from searching on BLAST. Then again, you could utilize the full-content inquiry at the website of UniProt database to do searching by using protein name in this case human vitronectin or catchphrases. We can use a blend of a various entries for searching purpose.

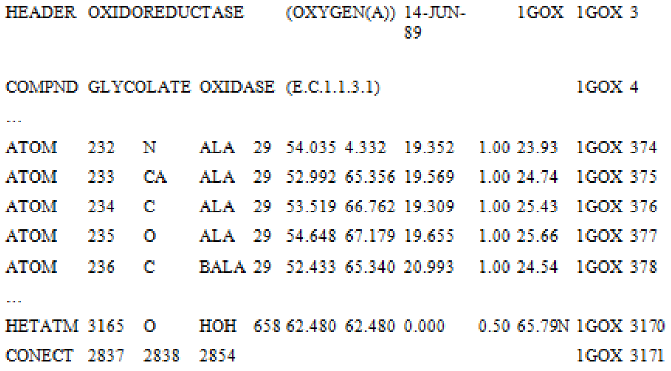
The format for the entries used for searching in UniProt is X_Y, here X is the memory code which can be of up to four letters and Y is of up to maximum five letters and is a mental helper animal varieties recognizable proof code of up to five characters for the natural wellspring of the protein. A few codes utilized for Y are full names written in English language, for example SHEEP, MOUSE, MAIZE, PIG, RAT, HUMAN, YEAST,HORSE, and WHEAT. Some are shortenings, containing BOVIN (cow-like), CHICK for chicken, and ECOLI for Escherichia coli.

If we merge the entries they can have several accession numbers. It is important to preserve accession number in order to have unambiguous references.As shown in following table format, NiceProt View design consist of accompanying things: (1) name and starting point, (2) protein properties, (3) general explanation (remarks), (4) ontologies (quality capacities), (5) twofold protein-protein connections, (6) succession comment (highlights), (7) arrangement, (8) references (writing reference), (9) web assets, (10) cross-references (connections to different databases), (11) section data, and (12) pertinent archives. The content in the general comment passage gives a capacity comment to the protein. We can record the explaination given by different protein databases with cross reference entries. Gene Cards, a database of human qualities, demonstrates chromosomal area and the contribution of the protein in specific sicknesses (if material). InterPro contains prescient protein "marks, for example, utilitarian areas, rehashes and critical destinations. Tapping the connection to InterPro from UniProt prompts a decent realistic view for space parsing.

III. DATABASES BASED ON PROTEIN STRUCTURE

It is becoming increasingly prevalent in today's biological science to search databases based on protein structures. It is important to study the 3 dimensional structure as it has got lot of importance as it defines the biological functions of the proteins and also it is important from drug design point of view. Nuclear Magnetic Resonance NMR like advanced technologies are helping in rapidly determining the structures of proteins. Protein Data Bank (PDB) is the largest repository of protein structures worldwide to process and distribute the structures of protein. The protein Structural data present in PDB was derived by using such technologies. They used NMR, X-ray crystallography, electron microscopy and few other technology to derive and decide the structures in PDB. The PDB database can be accessed online by using website www.rcsb.org/pdb and www.pdb.org.

We can search through the PDB database by using various ways. We can do the searching by using unique ID given by PDB, or we can also search by using keywords for features based on structures etc. There are two main types of file formats used by PDB for storing protein structure information. The first one is Protein Data Bank file format which was found by Bernstein in 1977. The second one is called as macromolecular crystallographic information file or mmCIF format which was introduced by Boume in 1997. The protein research fraternity still widely uses the old PDB file format. There are mainly three in the PDB file format viz. annotations, coordinates and connectivities. The chemical bonding between the atomic molecules is depicted in connectivity section which is not compulsory. Which is also placed at the end portion of a PDB file. The three dimensional coordinates of atoms are listed in coordinate section which begin with ATOM. Below is an example of a sample PDB file.



```
HEADER  OXIDOREDUCTASE      (OXYGEN(A)) 14-JUN-       1GOX  1GOX 3
                                        89

COMPND  GLYCOLATE OXIDASE  (E.C.1.1.3.1)                    1GOX 4
...
ATOM    232    N     ALA   29  54.035 4.332  19.352   1.00 23.93   1GOX 374
ATOM    233    CA    ALA   29  52.992 65.356 19.569   1.00 24.74   1GOX 375
ATOM    234    C     ALA   29  53.519 66.762 19.309   1.00 25.43   1GOX 376
ATOM    235    O     ALA   29  54.648 67.179 19.655   1.00 25.66   1GOX 377
ATOM    236    C     BALA  29  52.433 65.340 20.993   1.00 24.54   1GOX 378
...
HETATM  3165   O     HOH   658 62.480 62.480 0.000    0.50 65.79N  1GOX 3170
CONECT  2837   2838  2854                                            1GOX 3171
```

Each line demonstrates the molecule serial number, iota compose, deposit write, chain identifier (in the event of multi-chain structure), buildup serial number, orthogonal directions (three esteems), inhabitance, temperature factor, and fragment identifier.

There are three things viz. serial number of atom, type of atom, type of residue, serial number of a residue, three values of a orthogonal coordinates, occupancy, temperature factor and identifier for a segment. There are many records listed in the explanation section of the PDB file such as HEADER which contains the name of the protein molecule and date of first availability of it. The second is COMPND section which lists the molecular contents about the entry. SOURCE is the originating source. AUTHOR section lists the authors who were responsible for inventing. And few other things.

The PDB enables a client to see an atom structure intelligently through Jmol (Hanson, 2010), PDB SimpleViewer, PDB ProteinWorkshop, and RCSB-Kiosk, when the program is designed to help these free rendering apparatuses. The PDB gives related data about the protein, for example, optional structure task and geometry. Each PDB passage likewise connections to an extensive variety of explanations from auxiliary databases, including (1) outline and show databases, for example, Structural Biology Knowledgebase (SBKB, http://sbkb.org), PISA (Protein Interfaces, Surfaces and Assemblies; Krissinel and Henrick, 2007), Molecular Modeling Database (MMDB; Marchler-Bauer et al., 1999) in Entrez, PDBsum (Laskowski et al., 1997), Jena Library of Biological Macromolecules (JenaLib, http://www.fli-leibniz.de/IMAGE.html), PDBWiki (a group explained information base of natural sub-atomic structures, http://pdbwiki.org), and Proteopedia (a cooperative 3D-reference book of proteins and different particles; Prilusky et al., 2011); (2) space comment from SCOP (Murzin et al., 1995), CATH (Orengo et al., 1997), and Pfam (Finn et al., 2010); (3) structure correlation with different proteins utilizing different techniques; (4) the MEDLINE catalog; (5) protein developments recorded in Database of Macromolecular Movement (MolMovDB; Gerstein and Krebs, 1998); and (6) geometry examinations of the protein, for example, CSU Contacts of Structural Units (Sobolev et al., 1999) and castP Identification of Protein Pockets and Cavities (Liang et al., 1998).

Notwithstanding PDB and its connecting databases, other structure-related databases can likewise give helpful data. For instance, pdbLight (http://mufold.org/pdblight.php) coordinates protein grouping and structure information from numerous hotspots for protein structure expectation and investigation, together with anticipated SCOP arrangement for the week after week refreshed PDB structures. BioMagResBank (BMRB; University of Wisconsin, 1999) is a vault for NMR spectroscopy information on proteins, peptides, and nucleic acids. Especially, it gives incomplete NMR information (e.g., synthetic movements) previously the full structure is tackled. Protein Model Portal (PMP; Arnold et al., 2009) gives anticipated basic models and their quality appraisals for an expansive number of proteins.

IV. DATABASES BASED ON PROTEIN FAMILY

INTRODUCTION

Proteins can be characterized by their succession, developmental, auxiliary, or useful connections. A protein with regards to its family is significantly more instructive than the single protein itself. For instance, deposits rationed over the family regularly show unique useful parts. Two proteins arranged in the same practical family may propose that they share comparative structures, notwithstanding when their groupings don't have critical comparability.

There is no one of a kind method to arrange proteins into families. Limits between various families might be subjective. The decision of characterization framework depends to some degree on the issue; as a rule, the creator recommends investigating arrangement frameworks from various databases and contrasting them. Three sorts of order strategies are generally embraced in light of the closeness of succession, structure, or capacity. Grouping based techniques are appropriate to any proteins whose arrangements are known, while structure-based strategies are restricted to the proteins of known structures, and capacity construct strategies depend with respect to the elements of proteins being clarified. Grouping and structure-based orders can be mechanized and are adaptable to high-throughput information, though work based characterization is normally done physically. Structure- and capacity based techniques are more solid, while arrangement based strategies may bring about a false positive outcome when grouping similitude is frail (i.e., two proteins are ordered into one family by chance as opposed to by any organic importance). Likewise, since protein structure and capacity are preferable moderated over arrangement, two proteins having comparable structures or comparative capacities may not be recognized through succession based strategies.

1) Databases for Sequence-Based Protein Families

Arrangement based protein families are ordered by a profile got from a different succession arrangement. The profile can be appeared over a long space (many buildups or more) or can be uncovered in short grouping themes. Grouping techniques in view of profiles crosswise over long spaces have a tendency to be more dependable however less delicate than those in view of short arrangement themes.

A few arrangement construct strategies concentrate more in light of profiles crosswise over long areas, including Pfam (Finn et al., 2010), ProDom (Corpet et al., 1999), and Clusters of Orthologous Group (COG; Tatusov et al., 1997). These strategies vary in the procedures used to develop families. Pfam assembles different arrangement arrangements of numerous regular protein areas utilizing concealed Markov models. The ProDom protein area database comprises of homologous spaces in view of recursive PSI-BLAST looks. Machine gear-piece points toward discovering antiquated rationed areas by outlining groups of orthologs over a wide phylogenetic range. Savvy (Simple Modular Architecture Research Tool; Letunic et al., 2009) gathers space families, which are clarified regarding phyletic dispersions, utilitarian class, three-dimensional structures and practically vital deposits. It can be utilized for recognizable proof and explanation of hereditarily portable areas and investigation of space models. The iProClass database (Wu et al., 2004) joins different wellsprings of data for protein grouping. One can utilize every one of these databases for a far reaching investigation or pick one of them in view of the motivation behind the examination. Different succession based protein families have distinctive core interests. For instance, Pfam concentrates on work, ProDom on arrangement space, and COG on advancement.

Moreover, Pfam gives the arrangement among the relatives. One can distinguish a few highlights of the family through this example (i.e., from especially preserved deposits at particular arrangement positions). A few strategies depend on "fingerprints" of little saved themes in groupings, as with PROSITE (Hofmann et al., 1999), PRINTS (Attwood et al., 1999), and BLOCKS (Heniko et al., 1999). In protein succession families, a few locales have been exceptional moderated than others amid development. These districts are by and large critical for the capacity of a protein or for the upkeep of its three-dimensional structure or capacity. The fingerprints might be utilized to dole out a recently sequenced protein to a particular family. Fingerprints are gotten from gapped arrangements in PROSITE and PRINTS, yet are gotten from ungapped arrangements (relating to the very saved locales in proteins) in BLOCKS. A unique finger impression in PRINTS may contain a few themes from PROSITE, and in this way might be more adaptable and capable than a solitary PROSITE theme. Accordingly, PRINTS can give a helpful aide to PROSITE. It ought to be noticed that some practically random proteins might be arranged together because of chance matches in short themes.

2) Databases for Structure-Based Protein Families

The various leveled relationship among proteins can be unmistakably uncovered in structures through structure-structure correlation. Structure families frequently give more data on the connection between proteins than what arrangement families can offer, especially when two proteins share a comparable structure however no huge grouping character. Figure 1 demonstrates a case of a structure-structure arrangement between two proteins. Now and then, arrangement closeness between two proteins exists however isn't sufficiently solid to deliver an unambiguous arrangement. For this situation, the arrangement between two structures can create better arrangement as far as natural criticalness, and along these lines may pinpoint the developmental relationship and dynamic locales all the more precisely

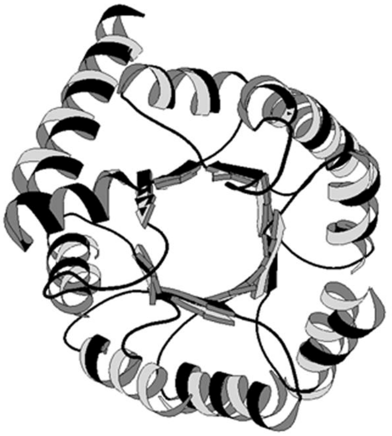


Figure 1. Structure superposition between glycolate oxidase(1gox, in black) and inosine monophosphate dehydrogenase

Diverse structure-structure correlation strategies yield distinctive structure families. CATH (Class, Architecture, Topology and Homologous superfamily; Orengo et al., 1997) is a various leveled grouping of protein space structures. CE (Combinatorial Extension of the ideal way; Shindyalov and Bourne, 1998) gives auxiliary neighbors of the PDB passages with structure-structure arrangements and three-dimensional superposition. FSSP (Fold order in view of Structure-Structure arrangement of Proteins; Holm and Sander, 1996) highlights a protein family tree and an area word reference, notwithstanding entire chain-based grouping, succession neighbors, and numerous structure arrangements. SCOP (Structural Classification of Proteins; Murzin et al., 1995) utilizes enlarged manual grouping, class, overlay, superfamily, and family arrangement. Huge (Vector Alignment Search Tool; Gibrat et al., 1996) contains delegate structure arrangements and three-dimensional superposition. Among these five databases, SCOP gives more capacity related data. Nonetheless, because of the manual work included, SCOP isn't refreshed as every now and again as the others (as of September 2011, it was last refreshed for the PDB discharge on June, 2009), though FSSP and CATH take after the PDB refreshes nearly.

SCOP is utilized here for instance to demonstrate the highlights of structure-based families. SCOP can be gotten to through its home server in the UK (http://scop.mrc-lmb.cam.ac.uk/scop/). SCOP depicts the various leveled relationship among proteins through the significant levels of (homologous) family, superfamily, and overlap. Proteins are bunched together into a (homologous) family in the event that they have noteworthy arrangement similitude. Distinctive families that have low succession comparability yet whose auxiliary and useful highlights propose a typical transformative root are set together in a superfamily.

Distinctive superfamilies are ordered into an overlay on the off chance that they have a similar real auxiliary structures in a similar course of action and with the same topological associations (the fringe components of optional structure and turn areas may vary in size and compliance). Two superfamilies in a similar crease might not have a typical developmental birthplace. Their basic likenesses may emerge from the material science and science of proteins supporting certain pressing courses of action and chain topologies.

3) Databases for Function-Based Protein Families

There are different protein practical families arranged from alternate points of view. The ENZYME information bank (Bairoch, 1993) contains the accompanying information for every catalyst: EC number, suggested name, elective names, reactant action, cofactors, pointers to the UniProt passage, and pointers to any malady related with an inadequacy of the compound. BRENDA (Scheer et al., 2011) gathers broad protein utilitarian information. Synergist Site Atlas (Porter et al., 2004) is a database of three-dimensional catalyst dynamic locales got from PDB structures. Different quality ontologies, for example, Gene Ontology (GO; The Gene Ontology Consortium, 2000) and KEGG, likewise sort out proteins into practical classifications. Comment and investigation by these ontologies for a given rundown of qualities can be completed utilizing devices and databases, for example, DAVID (Database for Annotation, Visualization and Integrated Discovery; Huang et al., 2009). What's more, there are a developing number of databases committed to unique kinds of proteins, for example, G-protein-coupled receptors, transporters, and protein kinases.

V. CONCLUSION

As the amount of protein data is incresing at an astronomical rate traditional methods are insufficiet to store, manage, manipulate and retrive it. Hence advanced computational databases which will be avaiable online is the need. Here we presented a report of few popular databases available online on the Internet. This paper can be a starting point for the researchers who want to start their research in protein analysis,computational biology and bioinformatics.

REFERENCES

[1] Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MD, Durbin R, Falquet L, Fleischmann W, Gouzy J, Hermjakob H, Hulo N, Jonassen I, Kahn D, Kanapin A, Karavidopoulou Y, Lopez R, Marx B, Mulder NJ, Oinn TM, Pagni M, Servant F, Sigrist CJ, Zdobnov EM. The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucl Acids Res. 2001;29:37–40. [PMC free article] [PubMed]

[2] Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, Scordis P, Selley J, Wright W. PRINTS prepares for the new millennium. Nucl Acids Res. 1999;27:220–225. [PMC free article] [PubMed]

[3] Bairoch A. The ENZYME data bank. Nucl Acids Res. 1993;21:3155–3156. [PMC free article] [PubMed]

[4] Bairoch A, Apweiler R. The UniProt protein sequence data bank and its supplement TrEMBL in 1999. Nucl Acids Res. 1999;27:49–54. [PMC free article] [PubMed]

[5] Barker WC, Garavelli JS, McGarvey PB, Marzec CR, Orcutt BC, Srinivasarao GY, Yeh LL, Ledley RS, Mewes H, Pfeiffer F, Tsugita A, Wu C. The PIR-international protein sequence database. Nucl Acids Res. 1999;27:39–42. [PMC free article] [PubMed]

[6] Benson DA, Boguski MS, Lipman DJ, Ostell J, Ouellette BF, Rapp BA, Wheeler DL. Genbank. Nucl Acids Res. 1999;27:12–17. [PMC free article] [PubMed]

[7] Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank: A computer based archival file for macromolecular structures. J Mol Biol. 1977;112:535–542. [PubMed]

[8] Contreras-Moreira B. 3D-footprint: a database for the structural analysis of protein-DNA complexes. Nucl Acids Res. 2010;38(Database issue):D91–97. [PMC free article] [PubMed]

[9] Corpet F, Gouzy J, Kahn D. Recent improvements of the ProDom database of protein domain families. Nucl Acids Res. 1999;27:263–267. [PMC free article] [PubMed]

[10] Etzold T, Ulyanov A, Argos P. SRS: Information retrieval system for molecular biology data banks. Methods Enzymol. 1996;266:114–128. [PubMed]

[11] Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. The Pfam protein families database. Nucl Acids Res. 2010;38(Database issue):D211–22. [PMC free article] [PubMed]

[12] Gao J, Agrawal GK, Thelen JJ, Xu D. P3DB: a plant protein phosphorylation database. Nucl Acids Res. 2009;37(Database issue):D960–D962. [PMC free article] [PubMed]

[13] Gibrat JF, Madej T, Bryant SH. Surprising similarities in structure comparison. Curr Opinion Struct Biol. 1996;6:377–385. [PubMed]

[14] Gromiha MM, An J, Kono H, Oobatake M, Uedaira H, Sarai A. Protherm: Thermodynamic database for proteins and mutants. Nucl Acids Res. 1999;27:286–288. [PMC free article] [PubMed]

[15] Gupta R, Birch H, Rapacki K, Brunak S, Hansen JE. O-GLYCBASE version 4.0: a revised database of O-glycosylated proteins. Nucl Acids Res. 1999;27:370–372. [PMC free article] [PubMed]