A Review on Several Protein Databses Available on the Internet

Changmin Liao

Library, China West Normal University
No. 1 of Shida Road, Nanchong City 637000, Sichuan Province, P.R. China
liaochangminlxh@aliyun.com

***Abstract*- Today, there have been many protein databases available on the Internet, and different databases possess different missions and functions. To submit or search easily desired protein-related data, two most commonly employed protein databases were introduced in detail, including wwPDB (the Worldwide Protein Data Bank) and UniProt (the Universal Protein Resource). Additionally, several other protein databases were also mentioned simply. This paper is beneficial for these researchers working on protein-related studies.**

Keywords- Protein; Database; wwPDB; UniProt

I. INTRODUCTION

For the rapid and ongoing accumulation of predicted protein sequences by high-throughput genome sequencing for numerous and increasingly diverse organisms, the expansion of large-scale proteomics (e.g. gene expression profiling and protein-protein interactions) and the advent of structural genomics have combined to provide a wealth of data to analyze and use. There is a widely recognized need for a centralized repository of protein sequences with comprehensive coverage and a systematic approach to protein annotation, incorporating, integrating and standardizing data from these various sources [1]. Therefore, many protein databases used for dealing with a large number of protein data were born on the internet. In this paper, wwPDB (Worldwide Protein Data Bank) and UniProt (Universal Protein Resource), as two most commonly employed protein databases, were introduced in detail, and some other protein databases were also mentioned simply.

II. WWPDB

wwPDB (the Worldwide Protein Data Bank) consists of organizations that act as deposition, data processing and distribution centers for protein data [2]. The founding members are RCSB PDB (Research Collaboratory for Structural Bioinformatics Protein Database in USA), PDBe (Protein Data Bank in Europe) and PDBj (Protein Data Bank in Japan). The BMRB (Biological Magnetic Resonance Data Bank in USA) group also joined the wwPDB in 2006 [3]. Each member's website can accept structural data and process the data. The processed data is sent to the "archive keeper", which is agreed to be RCSB PDB. This ensures that there is only one version of the data that is identical for all users. The modified database is then made available to the other wwPDB members, each of them makes the resulting structure files available through their websites to the public [4-7].

The mission of the wwPDB is to maintain a single Protein Data Bank Archive of macromolecular structural data that is freely and publicly available to the global community [2, 3]. Currently, a large number of protein data were deposited in wwPDB by above four members [8]. The statistics for PDB structures that are deposited and processed by year and site was seen in Table I. The official website of wwPDB was available at http://www.wwpdb.org/ (shown in Figure 1), this site offers information about services provided by the individual member organizations and about projects undertaken by the wwPDB, and the homepages of the four members were shown in Figure 2.

TABEL I STATISTICS FOR PDB STRUCTURES THAT ARE DEPOSITED AND PROCESSED BY YEAR AND SITE*

Year	Total Depositions	Deposited To			Processed By		
		RCSB PDB	PDBj	PDBe	RCSB PDB	PDBj	PDBe
2000	2983	2445	10	528	2297	158	528
2001	3287	2673	118	496	2408	383	496
2002	3565	2769	289	507	2401	657	507
2003	4830	3488	673	669	3135	1026	669
2004	5508	3796	900	812	3082	1614	812
2005	6678	4507	1166	1005	3563	2110	1005
2006	7282	5145	1052	1085	4252	1945	1085
2007	8130	5399	1603	1128	4703	2299	1128
2008	7073	5452	648	973	4106	1994	973
2009	8300	6715	527	1058	5069	2173	1058
2010	8878	6912	593	1373	5464	2041	1373
2011	9250	7172	582	1496	5938	1816	1496
2012	9972	7693	603	1676	6412	1884	1676
2013	3021	2317	216	488	1984	549	488
TOTAL	88757	66483	8980	13294	54814	20649	13294

* Last Updated: 24 Apr 2013. Including theoretical models and entries later withdrawn or obsoleted.

DOI: 10.5963/BER0202009

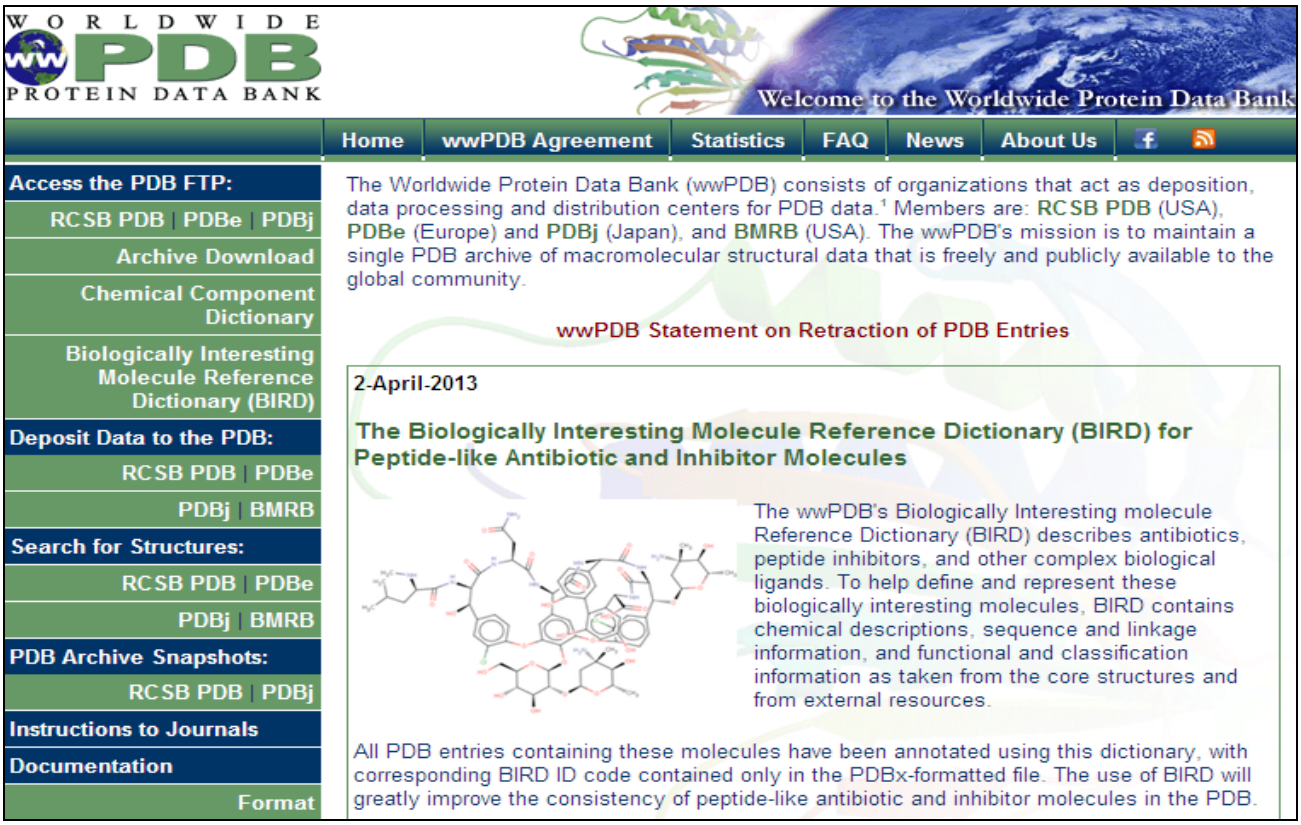

Figure 1 The homepage of wwPDB

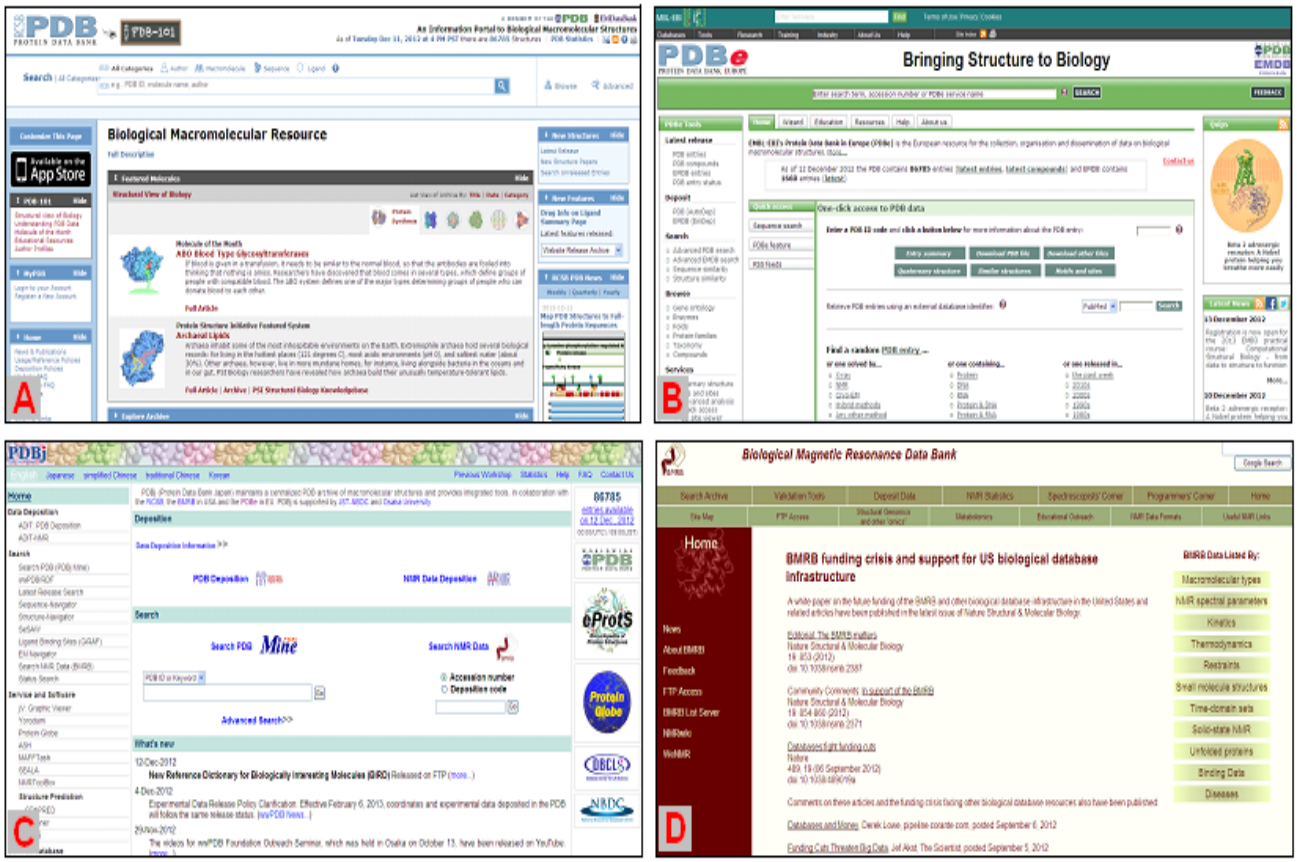

Figure 2 The homepages of the four members within wwPDB. A: RSCB-PDB (http://www.rcsb.org/pdb/home/home.do); B: PDBe (http://www.ebi.ac.uk/pdbe/); C: PDBj (http://www.pdbj.org/); D: BMRB (http://www.bmrb.wisc.edu/)

DOI: 10.5963/BER0202009

III. UNIPROT

UniProt (the Universal Protein Resource) is the central resource for storing and interconnecting information from large and disparate sources [9], and the most comprehensive catalog of protein sequence and functional annotation. It is produced by the UniProt Consortium which consists of groups from the European Bioinformatics Institute (EBI) (http://www.ebi.ac.uk/), the Swiss Institute of Bioinformatics (SIB) (http://www.isb-sib.ch/) and the Protein Information Resource (PIR) (http://pir.georgetown.edu/) [10, 11].

UniProt is comprised of four major components, each optimized for different uses. The UniProt Knowledgebase (UniProtKB) is an expertly accurate database, a central access point for integrated protein information with cross-references to multiple sources. The UniProt Archive (UniParc) is a comprehensive sequence repository, reflecting the history of all protein sequences. UniProt Reference Clusters (UniRef) merges closely related sequences based on sequence identity to speed up searches. While the UniProt Metagenomic and Environmental Sequences database (UniMES) were created to respond to the expanding area of metagenomic data [12, 13].

UniProt is freely and easily accessible for researchers who conduct interactive and customtailored analyses for proteins of interest to facilitate hypothesis generation and knowledge discovery. Simultaneously, it is a comprehensive, high-quality and freely accessible database involved in protein sequence and functional information, many of which are derived from the results of genome sequencing projects. Additionally, this database contains a large amount of information about the biological function of proteins derived from the research literature. The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information [14]. The homepage of UniProt database http://www.uniprot.org/ (Figure 3).

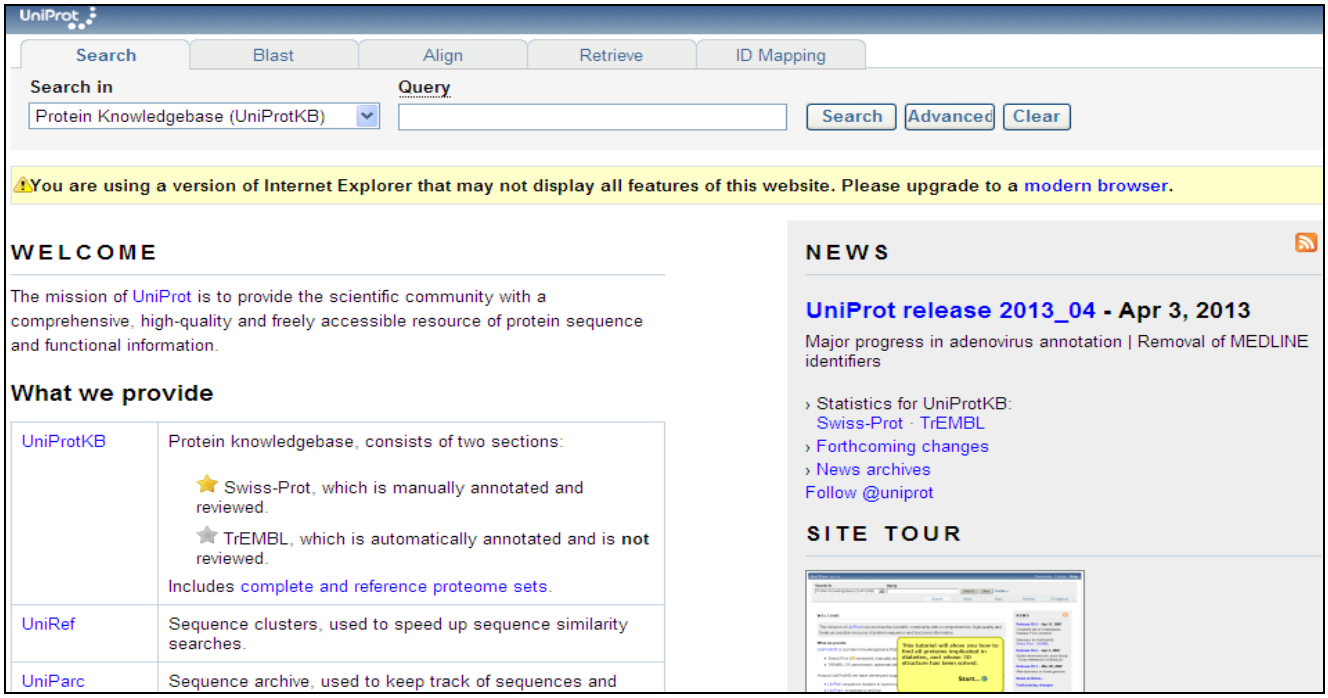

Figure 3 The homepage of UniProt database

IV. OTHER PROTEIN DATABASES

Besides wwPDB and UniProt, there are still many other protein databases available on the internet. Certainly, different databases have different websites and functions.

A. Protein Database in NCBI (The National Center for Biotechnology Information)

The homepage of this protein database is available at http://www.ncbi.nlm.nih.gov/protein, and its main function is a collection of sequences from several sources, including translations from annotated coding regions in GenBank, RefSeq and TPA, as well as records from SwissProt, PIR, PRF, and PDB.

B. ENZYME

The homepage of this protein database is available at http://enzyme.expasy.org/. It is primarily based on the recommendations of the Nomenclature Committee of IUBMB (the International Union of Biochemistry and Molecular Biology),

DOI: 10.5963/BER0202009

and it contains the following data for each type of characterized enzyme for which an EC (Enzyme Commission) number has been provided.

C. PGD

The homepage of PGD (Protein Geometry Database) is available at http://pgd.science.oregonstate.edu/. It allows us to explore either protein conformation or protein covalent geometry or the correlations between protein conformation and bond angles and lengths.

D. SCOP

The homepage of SCOP (Structural Classification of Proteins) is available at http://scop.mrc-lmb.cam.ac.uk/scop/. The aim of this protein database is to provide a detailed and comprehensive description of the structural and evolutionary relationships between all proteins whose structure is known. As such, it provides a broad survey of all known protein folds, detailed information about the close relatives of any particular protein, and a framework for future research and classification.

E. PRF

The homepage of PRF (The Protein Research Foundation) is available at http://www.proteinresearch.net/. Its main objectives are to replace imported protein for animal use with locally produced protein, but also to promote better utilization of protein. These objectives are promoted through funding research and technology transfers.

V. SUMMARY

Currently, there have been many protein databases available on the Internet, different databases possess different missions and functions. In this presented paper, two most commonly employed protein databases were detailed introduced, including wwPDB and UniProt. In addition, several other protein databases were also mentioned simply. This paper is beneficial for these researchers working on protein-related studies.

REFERENCES

[1] The UniProt Consortium, "The Universal Protein Resource (UniProt," Nucleic Acids Research, vol. 36, no. Database Issue, D190-D195, 2008.

[2] H. Berman, K. Henrick, and H. Nakamura, "Announcing the worldwide Protein Data Bank," Nat. Struct. Biol., vol. 10, no. 12, pp. 980, 2003.

[3] H. Berman, K. Henrick, H. Nakamura, and J. L. Markley, "The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data," Nucleic Acids Res., vol. 35, no. Database issue, pp. D301-D303, 2007.

[4] S. Velankar, Y. Alhroub, A. Alili, C. Best, H. C. Boutselakis, S. Caboche, et al., "PDBe: protein data bank in Europe," Nucleic Acids Res., vol. 39, no. Database Issue, pp. D402-D410, 2011.

[5] S. Dutta, K. Burkhardt, J. Young, G. J. Swaminathan, T. Matsuura, et al., "Data deposition and annotation at the worldwide protein data bank," Molecular Biotechnology, vol. 42, no. 1, pp. 1-13, 2009.

[6] J. L. Markley, E. L. Ulrich, H. M. Berman, K. Henrick, H. Nakamura, et al., "BioMagResBank (BMRB) as a partner in the worldwide protein data bank (wwPDB): new policies affecting biomolecular NMR depositions," Journal of Biomolecular NMR, vol. 40, no. 3, pp. 153-155, 2008.

[7] E. L. Ulrich, H. Akutsu, J. F. Doreleijers, Y. Harano, Y. E. Ioannidis, et al., "BioMagResBank," Nucleic Acids Res., vol. 36, no. Database issue, pp. D402-D408, 2008.

[8] http://www.wwpdb.org/stats.html.

[9] R. Apweiler, A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, et al., " UniProt: the universal protein knowledgebase," Nucleic Acids Research, vol. 32, no. Database Issue, pp. D115-D119, 2004.

[10] A. Bairoch, R. Apweiler, C. H. Wu, W. C. Barker, B. Boeckmann, et al., "The universal protein resource (UniProt)," Nucleic Acids Research, vol. 33, no. Database Issue, pp. D154-D159, 2005.

[11] E. Jain, A . Bairoch, S. Duvaud, I. Phan, N. Redaschi, et al., "Infrastructure for the life sciences: design and implementation of the UniProt website," BMC Bioinformatics, vol. 10, Article Number, 136, 2009.

[12] Y. L. Yip, M. Famiglietti, A. Gos, P. D. Duek, F. P. A. David, et al., "Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase," Human Mutation, vol. 29, no. 3, pp. 361-366, 2008.

[13] B. E. Suzek, H. Z. Huang, P. McGarvey, R. Mazumder, and C. H. Wu, "UniRef: comprehensive and non-redundant UniProt reference clusters," Bioinformatics, vol. 23, no. 10, pp. 1282-1288, 2007.

[14] http://www.uniprot.org/.

DOI: 10.5963/BER0202009