A Review on Several Protein Databases 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

wwwPDB (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*

<table>
<thead>
<tr>
<th>Year</th>
<th>Total Depositions</th>
<th>Deposited To</th>
<th>Processed By</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>RCSB PDB</td>
<td>PDBj</td>
<td>PDBe</td>
</tr>
<tr>
<td>2000</td>
<td>2983</td>
<td>2445</td>
<td>10</td>
</tr>
<tr>
<td>2001</td>
<td>3287</td>
<td>2673</td>
<td>118</td>
</tr>
<tr>
<td>2002</td>
<td>3565</td>
<td>2769</td>
<td>289</td>
</tr>
<tr>
<td>2003</td>
<td>4830</td>
<td>3488</td>
<td>673</td>
</tr>
<tr>
<td>2004</td>
<td>5508</td>
<td>3796</td>
<td>900</td>
</tr>
<tr>
<td>2005</td>
<td>6678</td>
<td>4507</td>
<td>1166</td>
</tr>
<tr>
<td>2006</td>
<td>7282</td>
<td>5145</td>
<td>1052</td>
</tr>
<tr>
<td>2007</td>
<td>8130</td>
<td>5399</td>
<td>1603</td>
</tr>
<tr>
<td>2008</td>
<td>7073</td>
<td>5452</td>
<td>648</td>
</tr>
<tr>
<td>2009</td>
<td>8300</td>
<td>6715</td>
<td>527</td>
</tr>
<tr>
<td>2010</td>
<td>8878</td>
<td>6912</td>
<td>593</td>
</tr>
<tr>
<td>2011</td>
<td>9250</td>
<td>7172</td>
<td>582</td>
</tr>
<tr>
<td>2012</td>
<td>9972</td>
<td>7693</td>
<td>603</td>
</tr>
<tr>
<td>2013</td>
<td>3027</td>
<td>2317</td>
<td>216</td>
</tr>
<tr>
<td>TOTAL</td>
<td>88757</td>
<td>66483</td>
<td>8980</td>
</tr>
</tbody>
</table>

* Last Updated: 24 Apr 2013. Including theoretical models and entries later withdrawn or obsoleted.
Figure 1 The homepage of wwPDB

Figure 2 The homepages of the four members within wwPDB. A: RCSB-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/)
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 custom tailored 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](image_url)

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),
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