Database and ontologies

PDBLight: a processed protein structure database for protein structure prediction and analysis

Zhiquan He, Chao Zhang,Yang Xu, Jingfen Zhang and Dong Xu*

Department of Computer Science and C.S. Bond Life Sciences Center, University of Missouri, Columbia, MO, USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Summary: PDBLight is a web-based database, which integrates protein sequence and structure data from multiple sources for protein structure prediction and analysis. Besides information collected from PDB in chain unit, computational results such as sequence profile, secondary structure, solvent accessibility, and predicted SCOP classification have also been integrated. PDBLight can easily retrieve a subset of proteins according to user-specified sequence or structure criteria, e.g., sequence identity threshold or SCOP classification. In addition, cleaned sequence and structure data for each chain could be downloaded in bulk, which is convenient for large-scale computational analyses. With friendly interface, users can also view detailed information of each protein, e.g. sequence logo and three-dimensional structure graphics. Moreover, PDBLight is weekly updated following the release of PDB.

Availability: The database is publicly available and can be accessed at http://mufold.org/pdblight.php.

Contact: xudong@missouri.edu.

1 INTRODUCTION

Protein structure data in Protein Data Bank (PDB)[1] are widely used in studies of protein function and evolution, and they serve as a basis for protein structure prediction. The number of entries in PDB has been increasing rapidly. However, there are two barriers in large-scale usage of PDB data, especially in an automatic fashion. The first barrier is that a large number of protein chains in PDB are highly similar in terms of sequence or structure. For example, many PDB files contain identical chains. Hence, a light version of DPB may be useful. In addition, it is often needed to obtain a set of PDB chains satisfying some criteria such as structure resolution and sequence length, and to classify similar chains into groups to select only a representative from each group. The second barrier in large-scale usage of PDB data is that many PDB files have issues due to inconsistency of data and standards so that automated retrieval and analysis are often difficult. For example, the sequence in a PDB header is sometimes inconsistent with that in the coordinate part. Another example is that some residues in PDB are modified and the residue types cannot be easily mapped to the original amino acids. To address these barriers, extensive manual work often has to be done for handling PDB files in large scale and various labs often do such redundant work.

Currently, several websites are available to address the first barrier. The PDB website itself can remove similar sequences with specific level of mutual sequence identity. Other websites such as PDB-Select[2], ASTRAL [3] , PDB-REPRDB[4] and PISCES[5] have similar functions, all of which allow users to download a predefined chain list or generate a customized list with some sequence or structure criteria. However, these web servers only provide a chain list and do not deal with the second barrier. More importantly, the derived chain lists from these websites are typically not updated weekly following the release of hundreds of PDB files each week. Release of non-redundant structure datasets is even slower. For example, the widely used protein structure classification database SCOP [6], which involves extensive manual annotations, was updated one and half years ago (1.75 release (June 2009)). It would be useful to automatically incorporate SCOP classification for newly released PDB files, even if the classification quality is suboptimal.

In this paper, we introduce PDBLight, which comprehensively integrates PDB data, predicted SCOP classification and additional computational data, e.g. DSSP [7] secondary structure and PSIBLAST [8] sequence profile. PDBLight provides a friendly web interface for user to browse, search and download these data. The main features of PDBLight are as follows:

(1) User can search a PDB sequence against several derived sequence databases by using BLAST with specified parameters and browse all the hit sequences.

(2) User can generate a customized list from the entire PDB sequences by setting the filtering parameters, which include experimental method (e.g., X-Ray or NMR), sequence length, structure resolution (only applied to X-Ray structures), deposit date, full or partial SCOP address and mutual sequence identity level from 90, 80 to 30 percent. This can be used for a non-redundant template database in developing protein energy function and template-based protein structure prediction.

(3) User can input a list of chain names to browse the corresponding information and quickly get the representatives of the involved clusters and their members with seven levels of mutual sequence identity, from 90 to 30 percent. As an

*To whom correspondence should be addressed.

example, this can be used for selecting multiple templates in homology modeling.

(4) User can download data for a list of chains in different data formats, including FASTA sequence files, and original or processed PDB files.

(5) User can browse and download the pre-computed sequence and SCOP representative datasets. The files can also be retrieved through a command line without going through a web browser.

(6) User can view each chain in details, including the basic information from PDB file, evolutional information in sequence logo, secondary structure and three-dimensional structure graphics with JMol (http://www.jmol.org).

(7) The database is automatically updated every week following the weakly release of PDB.

2 DATA AND METHOD

As an automatic routine, PDBLight weekly synchronizes its PDB files to ftp://ftp.wwpdb.org/pub/pdb/data/structures/all/pdb/ and organizes the processed data in chain units, as shown in Fig. 1. Original PDB files are processed to have a simplified, clean PDB format. For all cleaned PDB chains, secondary structures are computed using DSSP; non-redundant sequences (defined as kernel), are generated by mapping multiple sequences to one representative (M:1 or many-to-one mapping). A representative (kernel) sequence or a user-specified chain ID can be mapped to PDB chains through a one-to-many (1:M) mapping by sequence similarity. Sequence profile, hidden Markov Model[9] and predicted SCOP classification, are computed for kernel sequences.

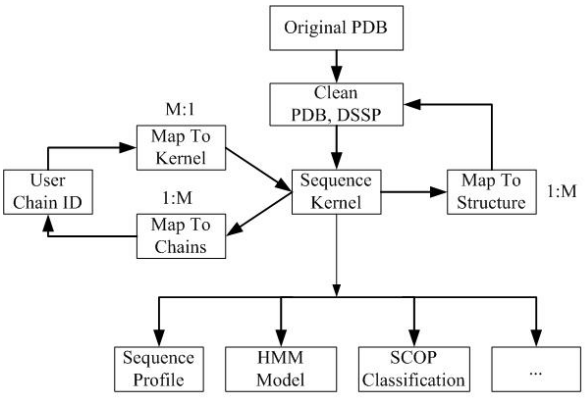

Fig. 1. Data Organization. 1: M (one-to-many mapping); M:1 (many-to-one mapping)

This data organization reduces the data redundancy and therefore saves storage space. Besides the basic information such as experimental method, resolution, deposit date, source and references are retrieved from original PDB files, the sequence and structure data are processed and cleaned as follows:

1. The sequences from the header (reference sequence) and coordinate part of PDB file for each chain are aligned through a constrained sequence alignment, in which gap is not allowed in the reference sequence.
2. We have restored the residue codes for the majority of the modified residues through "MODRES" records from PDB files.
3. Atoms of a residue beyond its standard composition are simply removed as it is difficult for various structure prediction and analysis tools, e.g. Modeller [10] to process them.
4. PDB chains are removed if the structures have only CA or sequences with length less than 30, or contain too many unknown residues ('X's).
5. If alternative conformations exist for residues or atoms, the first conformation is selected. For example, residue 1 of THR in PDB 1CBN has two conformations.

PDBLight provides a fast way to generate a subset of chains from the whole dataset with seven levels of sequence identity, from 90 to 30 percent. This is implemented by a systematic indexing scheme and pre-computed clustering results. Sequence clustering with threshold from 90 to 40 percent is done using CD-Hit[11]. Clustering into 30 percent of mutual identity is done by all-to-all sequence comparison using PSIBLAST, as the lowest identity cutoff of CD-Hit is 40%. The similarity between two sequences is computed by the PSIBLAST local alignment identity divided by the average sequence length. The selection of the representative from each cluster is based on combination of sequence length, structure resolution and deposit date. Longer sequence has higher priority to be selected; but if two sequences have a length difference less than 10 residues, the one with higher resolution will be selected. Here X-ray structures are always assumed superior to NMR structures. Sequences with later deposit date have higher priority as newly resolved structures are more likely to have better quality.

We developed an automatic protocol to classify the sequences into SCOP hierarchy. The method of assigning SCOP address to each sequence is as follows:

1. We compare each new sequence in PDB dataset against all sequences of SCOP dataset using PSIBLAST. Select those hits whose E-value is less than 0.01 and Z-Score of the corresponding CE [12]structure alignment is greater than 4.5.
2. If no hit is found in step 1, compare the query structure to the family representatives of the SCOP dataset. Select those hits whose CE Z-Score is greater than 4.5.
3. When multiple hits are found from step 1 or 2, assign the address of the new sequence to the hit with the highest CE Z-Score. When the Z-Score is identical, we choose the longest sequence as the representative.
4. Check unassigned regions: if the length is greater than 30, repeat steps 1 to 3 using the sub-sequence; otherwise merge the short unassigned regions to the neighboring domains.

Test results show that our automatic protocol approximates SCOP classification well. We selected 585 sequences from SCOP 1.75 that are not present in SCOP 1.73. 94 of them are multi-domain sequences with 186 domains in total. The remaining 491 are single domain sequences. The test is done against SCOP 1.73. Table 1 shows the assignment accuracy. For multi-domain sequences, if the predicted domain region covers more than 80% of the expected SCOP domain region, the prediction is regarded as correct.

Table 1. SCOP Classification Accuracy

	Family	Super-family	Fold
Multi-domain (186)	95.16%	96.82%	98.39%
Single domain (491)	92.87%	96.95%	97.35%

The unassigned rate is 4.14% among the 186+491=677 domains, which may represent novel folds. Our performance is comparable to the published accuracies for fastSCOP [13], where structures were assigned into SCOP super-families only. Meanwhile, fastSCOP and other websites, e.g. ASTRAL, are unavailable for weekly updates like ours. With new SCOP sequences released, our classification could be adjusted and updated to close the gap between our protocol and SCOP classification.

PDBLight integrates computational results of the sequence and structure to help user better understand the protein. We have calculated sequence profiles for all the sequences by running PSIBLAST (2007 release version) three rounds against the non-redundant (NR) database (released in 2008) with the E-value cutoff of 0.001. Sequence profile is represented as a logo image, which is generated using Weblogo[14] for the first up to 100 alignments extracted from the last round of PSIBLAST. The secondary structure and solvent accessibility are computed by DSSP. Secondary structure is represented in three states: H (helix), E (beta strand) and C (coil). And relative solvent accessibility (RSA) is computed and classified into three states: E (exposed, RSA is greater than 0.37), B (buried, RSA is less than 0.069) and I (intermediate)[15]. In addition, a structure image is generated for each chain using Raster3D [16] and MOLSCRIPT [17],and a user can view the three-dimensional structure interactively with JMol.

To the date of Nov. 14, 2010, the data of PDBLight can be summarized in Tables 2-4.

Table2.Number of processed PDB files, and the deposit time of the PDB files.

PDB Files	Total Chains	Unique Chains	Deposited From	Deposited To
65,943	158,601	44,425	Aug. 11, 1972	Oct. 29, 2010

Table 3.Number of representative sequences at each mutual sequence identity threshold level.

30	40	50	60	70	80	90
13,978	15,997	18,146	19,942	21,307	22,588	24,378

Table 4. Numbers of SCOP representatives at family, super-family and fold levels.

Family	Super-family	Fold
3872	1951	1189

PDBLight will be continuously maintained and updated. Part of the future work will be better handling of residue modification and missing coordinates. Currently, we only restore the residue codes and ignore the conformation changes in the atomic coordinates from the modified state to the apo protein. Besides, numerous residues miss some atom coordinates and some chains contain segments without coordinates. Our study for handling these issues is ongoing.

ACKNOWLEDGEMENTS

This work has been supported by National Institutes of Health Grant R33GM078601. The computations were mainly performed on the high-performance computing resources at the University of Missouri Bioinformatics Consortium.

REFERENCES

1. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, Ritter O, Abola EE: **Protein Data Bank (PDB): database of three-dimensional structural information of biological macromolecules**. *Acta Crystallogr D Biol Crystallogr* 1998, **54**(Pt 6 Pt 1):1078-1084.
2. Griep S, Hobohm U: **PDBselect 1992-2009 and PDBfilter-select**. *Nucleic Acids Res* 2010, **38**(Database issue):D318-319.
3. Brenner SE, Koehl P, Levitt M: **The ASTRAL compendium for protein structure and sequence analysis**. *Nucleic Acids Res* 2000, **28**(1):254-256.
4. 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.
5. Wang G, Dunbrack RL, Jr.: **PISCES: a protein sequence culling server**. *Bioinformatics* 2003, **19**(12):1589-1591.
6. 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**(4):536-540.
7. Kabsch W, Sander C: **Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features**. *Biopolymers* 1983, **22**(12):2577-2637.
8. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
9. Soding J: **Protein homology detection by HMM-HMM comparison**. *Bioinformatics* 2005, **21**(7):951-960.
10. Sali A, Blundell TL: **Comparative protein modelling by satisfaction of spatial restraints**. *J Mol Biol* 1993, **234**(3):779-815.
11. Li W, Godzik A: **Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences**. *Bioinformatics* 2006, **22**(13):1658-1659.

12. Shindyalov IN, Bourne PE: **Protein structure alignment by incremental combinatorial extension (CE) of the optimal path**. *Protein Eng* 1998, **11**(9):739-747.
13. Tung CH, Yang JM: **fastSCOP: a fast web server for recognizing protein structural domains and SCOP superfamilies**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W438-443.
14. Crooks GE, Hon G, Chandonia JM, Brenner SE: **WebLogo: a sequence logo generator**. *Genome Res* 2004, **14**(6):1188-1190.
15. Xu Y, Xu D, Uberbacher EC: **An efficient computational method for globally optimal threading**. *J Comput Biol* 1998, **5**(3):597-614.
16. Merritt EA, Murphy ME: **Raster3D Version 2.0. A program for photorealistic molecular graphics**. *Acta Crystallogr D Biol Crystallogr* 1994, **50**(Pt 6):869-873.
17. Kraulis PJ: **MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures**. *J Appl Cryst* 1991, **24**:946-950.