D486–D489 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 1 December 2011*
doi:10.1093/nar/gkr1150

ccPDB: compilation and creation of data sets from Protein Data Bank

Harinder Singh[1], Jagat Singh Chauhan[1], M. Michael Gromiha[2], Open Source Drug Discovery Consortium[3] and Gajendra P. S. Raghava[1,*]

[1]Bioinformatics Centre, Institute of Microbial Technology, Chandigarh, [2]Department of Biotechnology, Indian Institute of Technology Madras, Chennai 600036 and [3]The Open Source Drug Discovery (OSDD) Consortium, Council of Scientific and Industrial Research, Anusandhan Bhavan, 2 Rafi Marg, Delhi 110001, India

Received August 25, 2011; Revised November 9, 2011; Accepted November 10, 2011

ABSTRACT

ccPDB (http://crdd.osdd.net/raghava/ccpdb/) is a database of data sets compiled from the literature and Protein Data Bank (PDB). First, we collected and compiled data sets from the literature used for developing bioinformatics methods to annotate the structure and function of proteins. Second, data sets were derived from the latest release of PDB using standard protocols. Third, we developed a powerful module for creating a wide range of customized data sets from the current release of PDB. This is a flexible module that allows users to create data sets using a simple six step procedure. In addition, a number of web services have been integrated in ccPDB, which include submission of jobs on PDB-based servers, annotation of protein structures and generation of patterns. This database maintains >30 types of data sets such as secondary structure, tight-turns, nucleotide interacting residues, metals interacting residues, DNA/RNA binding residues and so on.

INTRODUCTION

Annotating the structure and function of a protein is one of the major challenges in the post-genomic era. Development of bioinformatic methods for such annotation requires experimentally proven data for training, testing and validation. Hence, clean/refined data sets (e.g. non-redundant, experimentally validated) are the heart of bioinformatic methods. Protein Data Bank (PDB) is one of the major sources of experimentally obtained data and it contains >74 000 protein structures determined using X-ray crystallography, NMR spectroscopy and other techniques (1). It plays a vital role in the field of protein structure/function annotation as most of the bioinformatic tools rely on the data derived from PDB. In order to facilitate protein community, a large number of secondary databases have been derived from PDB, which includes SCOP (2), CATH (3), SuperSite (4), PDB-ligand (5), PDBsum (6), etc. In addition, a number of tools have been developed for extracting useful information from PDB and secondary databases such as DSSP (7), PROMOTIF (8), LPC (9) and HBPLUS (10).

Recently, Joosten *et al.* (11) described a series of PDB-related databases that are heavily used for developing bioinformatic techniques. They described mainly the databases developed and maintained by their own group, which include DSSP (secondary structure of proteins), HSSP (12) (multiple sequence alignment), PDBFINDER (summaries of PDB file), PDB_SELECT (13) (non-redundant proteins) and WHY_NOT (explanation as to why entries in other databases cannot exist). These databases are useful for developing new methods in the field of structural bioinformatics. However, additional scripts/software are required for extracting or parsing the data to obtain clean/refined data sets. Wang and Dunbrack (14) developed a server PISCES to cull protein sequences/structures from a set of PDB codes or FASTA sequences with an option to select the cutoff for sequence identity. This server is specific to a set of protein sequences/structures.

In this work, we developed a database, which is a collection of commonly used data sets for structural or functional annotation of proteins. We have accumulated numerous data sets from the literature, which were used for developing methods to annotate proteins at the sequence (or residue) level. In order to provide updated data sets on the latest release of PDB, we created and maintained important data sets from various releases of PDB. In addition, for providing customized data sets, we also developed a series of web-based tools for creating new data sets. These newly created data sets can be used to benchmark the existing and newly developed methods

*To whom correspondence should be addressed. Tel: +91 172 2690557; Fax: +91 172 2690632; Email: raghava@imtech.res.in

such as Protein Classification Benchmark Collection (15), prediction of protein secondary structures (16), evaluation of multiple sequence alignment (17), identification of binding sites in DNA/RNA binding proteins (18,19), ATP and ligand binding sites (20,21) and so on.

SYSTEMS AND METHODS

Data collection and organization

We extracted most of the bioinformatics related papers from Pubmed and other resources, and obtained the data sets from supplementary materials, databases, websites and/or directly from the authors. These data sets were classified with their contents and maintained at ccPDB. In order to compile data sets, we downloaded all PDB files from http://www.pdb.org/. These PDB files are maintained/mirrored at our server using rsync command, which allows users to create customized data sets from the latest release of PDB. We also maintain DSSP database in our server, which provides secondary structure and other related information. In ccPDB database, we used various software packages for deriving useful information from PDB. The following are major software used in ccPDB: (i) PROMOTIF for identifying structural motifs (8); (ii) LPC for generating ligand–protein interaction data (9); (iii) PDIdb for identifying amino acid residues, which are interacting with DNA/RNA (22,23); and (iv) In-house Perl scripts for wide range of calculations and analyzing PDB files.

Database architecture

ccPDB is built on Apache HTTP server 2.2 with MySQL server 5.1.47 as the back end and the PHP 5.2.9, HTML and JavaScript as the front end. Apache, MySQL and PHP technology were preferred as they are open-source software and platform independent.

IMPLEMENTATION

This is a comprehensive database, which maintains existing data sets collected from the literature and compiled data sets derived from PDB. In addition, database server part also allows users to create customized data sets. The database is broadly divided into three sections and the brief description of each section is given below.

Collection of data sets

This section maintains published data sets that were used for developing prediction methods. These data sets were collected from the literature after an extensive search. These data sets are divided into various categories as described below:

- Protein secondary structure: in this category, we maintain data sets used for developing secondary structure prediction of proteins.
- Nucleotide interacting residues: it contains data sets used for developing prediction methods for DNA or RNA interacting residues in proteins.
- Ligand interacting residues: it maintains data sets used for predicting ligand-protein interacting residues (e.g. ATP, GTP, FAD, MAN, etc.).

Compilation of data sets

Data sets in the section 'Collection of data sets' are useful for benchmarking any newly developed methods with existing methods. For developing a new method, one should generate data sets from the latest release of PDB, as the performance of a method mainly depends on the size of the data set. Hence, old data sets generated earlier would become obsolete as the number of protein structures in PDB is rapidly increasing. In order to reduce the task of developing data sets to protein community, we compile and maintain data sets from the latest releases (July 2011) of PDB. In addition, we will also maintain data sets generated from the previous releases of PDB. Data sets compiled from PDB are listed in Table 1 along with their compilation procedures.

Creation of data set

This is a major module of ccPDB developed for creating customized data sets. In order to facilitate users, we developed a six step procedure for creating customized data sets that provides full flexibility in each step (Table 2). Following is a brief description of each step.

- *Extract protein chains*: this option allows users to extract specific chains from PDB, for example, extraction of ATP binding protein chains from PDB. User may extract protein chains with desired structure or function. In addition, this option allows users to extract PDB chains of desired function from the list of PDB IDs provided by the users.
- *General filters*: these filters allow users to extract PDB chains from the latest release with desired conditions,

Table 1. Brief description of major data sets created at ccPDB

Type of data set	Description of data set	Software package
Secondary structure	Data sets related to secondary structure, helix, strand, coil, etc.	DSSP
Tight turns	Data sets created for variuos types of tight turns (e.g. β-turn).	PROMOTIF
Nucleotide interacting	Data sets created for small nucleotide and metal (e.g. ATP, GTP, Fe, Mg etc) binding residues.	LPC
DNA/RNA binding residues	Data sets of DNA and RNA binding proteins/residues.	PDIdb
Metal binding residues	Data set of metal binding proteins/residues (e.g. Zn, Ca interacting residues).	LPC

Table 2. Description of each process/step of data set creation module of ccPDB

Process/step	Description of process	Example
Protein/chain	Allows users to extract PDB chains having desired structure or function.	Extract ATP binding protein chains from PDB.
General filters	Extract chains using various filters like resolution, experimental technique.	Protein chains having resolution better than 3 Å, solved by X-ray crystallography.
Combination of sets	Allows to combine two sets of data.	ATP binding protein chains having resolution better than 3 Å, solved by X-ray crystallography.
Extract sequences	Extract the amino acid sequences of PDB chains.	Extract sequences of ATP binding proteins.
Non-redundant data sets	Creation of non-redundant data sets using BLASTClust.	Generate non-redundant data set at 25% of ATP binding proteins.
Annotation of residues	Assigning structure/function of each residue in PDB chains.	Mapping of ATP interacting residues in ATP binding protein chains.

for example users can select protein chains solved by 'X-ray' crystallography solved at a resolution better than 2.5 Å. Major filters included in this option are (i) experimental method, (ii) resolution and (iii) length of amino acid sequence. This option also allows users to remove redundancy in extracted proteins.

- *Combination of sets*: this option allows users to generate a new set of protein chains from two sets of data using various combinations. For example, it allows users to select chains, which are common in two sets or unique chains in two sets. This is useful for combining sets extracted from the above two steps.
- *Extracted sequences*: the above three steps allow users to extract protein chains as per their requirement. This step allows users to extract amino acid sequences of these chains from PDB.
- *Non-redundant sequences*: creation of non-redundant data set is important for training, testing and validating any prediction model. This page provides an option to remove the redundant sequences from a set of protein sequences.
- *Annotation of residues*: this interface allows users to create data sets at residue level. For example, users can assign secondary structure of each residue in a protein. This option is designed to assign ligand/DNA/RNA interacting residues in a protein.

These six steps will help users for creating different types of data sets from PDB.

Web services

ccPDB provides a number of web services for facilitating PDB users. These services allow users to perform various types of tasks including the analysis of PDB files. Following is a brief description of tools integrated in ccPDB.

Analysis of PDB_ID. In the last two decades, a number of web-based services have been developed for analyzing PDB files. These tools have been developed by various groups over the years and are available at different web sites. Hence, one has to visit various sites and submit their PDB ID to use these tools. In order to facilitate users, we developed a web interface that integrates >40 servers, where users can submit their PDB ID on these servers from our interface.

Structure information. This option provides following type of information about a PDB ID: (i) amino acid composition of chains, (ii) number and type of ligand/metal interacting residues and (iii) tight-turns in proteins.

Search in PDB files. This search module allows users to search PDB on major fields such as, ligand, organism, PDB code, etc. This option also allows users to display various types of information such as type of interacting residues (ligands/metals), secondary structure (DSSP states), tight-turns, amino acid compositions, etc.

Generate pattern. This allows users to create patterns from protein chains in the desired format suitable to various packages of machine learning techniques like SVM_light, Weka and SNNS. It allows users to generate patterns at protein level as well as at residue level.

Download information. This server allows users to download PDB files and related information that includes PDB, DSSP and PDBFINDER2 files.

UPDATE OF DATABASE

This database will be updated manually as well as automatically. In order to update the contents in 'Collection of data sets' section, we check recent data sets from the literature. We are also providing online submission facility that will allow users to submit data sets to our database. 'Compilation of data sets' section will be updated every 6 months using in-house written scripts. Creation of data set section will be updated every 3 months.

AVAILABILITY AND REQUIREMENTS

ccPDB is freely available at http://crdd.osdd.net/raghava/ccpdb

ACKNOWLEDGEMENTS

We thank Dr Roman Laskowski, European Bioinformatics Institute, UK, for providing databases and tools developed in his group. We are grateful to Prof. G. Vriend and his colleagues for providing PDB-related databases to community. We also acknowledge Tomas Norambuena for providing the customized version of PDIdb package.

FUNDING

OSDD, CSIR and DBT (India). Funding for open access charge: Council of Scientific and Industrial Research.

Conflict of interest statement. None declared.

REFERENCES

1. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al*. (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*., **39**, D392–D401.
2. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*., **36**, D419–D425.
3. Cuff,A.L., Sillitoe,I., Lewis,T., Clegg,A.B., Rentzsch,R., Furnham,N., Pellegrini-Calace,M., Jones,D., Thornton,J. and Orengo,C.A. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res*., **39**, D420–D426.
4. Bauer,R.A., Gunther,S., Jansen,D., Heeger,C., Thaben,P.F. and Preissner,R. (2009) SuperSite: dictionary of metabolite and drug binding sites in proteins. *Nucleic Acids Res*., **37**, D195–D200.
5. Shin,J.M. and Cho,D.H. (2005) PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res*., **33**, D238–D241.
6. Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res*., **37**, D355–D359.
7. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
8. Hutchinson,E.G. and Thornton,J.M. (1996) PROMOTIF–a program to identify and analyze structural motifs in proteins. *Protein Sci*., **5**, 212–220.
9. Sobolev,V., Sorokine,A., Prilusky,J., Abola,E.E. and Edelman,M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
10. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol*., **238**, 777–793.
11. Joosten,R.P., te Beek,T.A., Krieger,E., Hekkelman,M.L., Hooft,R.W., Schneider,R., Sander,C. and Vriend,G. (2010) A series of PDB related databases for everyday needs. *Nucleic Acids Res*., **39**, D411–D419.
12. Dodge,C., Schneider,R. and Sander,C. (1998) The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res*., **26**, 313–315.
13. Griep,S. and Hobohm,U. (2010) PDBselect 1992–2009 and PDBfilter-select. *Nucleic Acids Res*., **38**, D318–D319.
14. Wang,G. and Dunbrack,R.L. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
15. Sonego,P., Pacurar,M., Dhir,S., Kertesz-Farkas,A., Kocsor,A., Gaspari,Z., Leunissen,J.A. and Pongor,S. (2007) A Protein Classification Benchmark collection for machine learning. *Nucleic Acids Res*., **35**, D232–D236.
16. Chatterjee,P., Basu,S., Kundu,M., Nasipuri,M. and Plewczynski,D. (2011) PSP_MCSVM: brainstorming consensus prediction of protein secondary structures using two-stage multiclass support vector machines. *J. Mol. Model*., **17**, 2191–2201.
17. Raghava,G.P., Searle,S.M., Audley,P.C., Barber,J.D. and Barton,G.J. (2003) OXBench: a benchmark for evaluation of protein multiple sequence alignment accuracy. *BMC Bioinformatics*, **4**, 47.
18. Zhang,T., Zhang,H., Chen,K., Ruan,J., Shen,S. and Kurgan,L. (2010) Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. *Curr. Protein Pept. Sci*., **11**, 609–628.
19. Carson,M.B., Langlois,R. and Lu,H. (2010) NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res*., **38**, W431–W435.
20. Chauhan,J.S., Mishra,N.K. and Raghava,G.P. (2009) Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics*, **10**, 434.
21. Roche,D.B., Tetchner,S.J. and McGuffin,L.J. (2011) FunFOLD: an improved automated method for the prediction of ligand binding residues using 3D models of proteins. *BMC Bioinformatics*, **12**, 160.
22. Norambuena,T. and Melo,F. (2010) The protein-DNA interface database. *BMC Bioinformatics*, **11**, 262.
23. Lewis,B.A., Walia,R.R., Terribilini,M., Ferguson,J., Zheng,C., Honavar,V. and Dobbs,D. (2010) PRIDB: a protein-RNA Interface Database. *Nucleic Acids Res*., **39**, D277–D282.