Journal of
Applied Crystallography

ISSN 0021-8898

Received 11 October 2011
Accepted 4 December 2011

PDB-2-PB: a curated online protein block sequence database

V. Suresh, K. Ganesan and S. Parthasarathy*

Department of Bioinformatics, School of Life Sciences, Bharathidasan University, Tiruchirappalli 620 024, Tamil Nadu, India. Correspondence e-mail: bdupartha@gmail.com

© 2012 International Union of Crystallography
Printed in Singapore – all rights reserved

This article describes the development of a curated online protein block sequence database, PDB-2-PB. The protein block sequences for protein structures with complete backbone coordinates have been encoded using the encoding procedure of de Brevern, Etchebest & Hazout [*Proteins* (2000), **41**, 271–287]. In the current release of the PDB-2-PB database (version 1.0), the protein entries from a recent release of the World Wide Protein Data Bank (wwPDB), which has 74 297 solved PDB entries as of 7 July 2011, have been used as a primary source. The PDB-2-PB database stores the protein block sequences for all the chains present in a protein structure. PDB-2-PB version 1.0 has the curated protein block sequences for 103 252 PDB chain entries (93 547 X-ray, 7033 NMR and 2672 other experimental chain entries). From the PDB-2-PB database, users can extract the curated protein block sequence and its corresponding amino acid sequence, which is extracted from the PDB ATOM records. Users can download these sequences either by using the PDB code or by using various parameters listed in the database. The PDB-2-PB database is freely available at http://bioinfo.bdu.ac.in/~pb/.

1. Introduction

The three-dimensional structures of proteins are very useful for understanding their biological functions. Owing to the increase in the number of protein structures deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000), the number of available structure comparison methods is rapidly increasing. Initially, protein secondary structures were used as the primary elements in structure comparison methods (Gibrat *et al.*, 1996; Singh & Brutlag, 1997; Lu, 2000). However, in recent years, researchers have found that the information obtained from local protein structures (LPSs) is also helpful for examining each and every part of the protein structure (Jones & Thirup, 1986; Unger *et al.*, 1989; Levitt, 1992), and the LPS has found many potential applications. Offmann *et al.* (2007) reviewed different kinds of LPS and their interesting applications in protein structure analysis.

The protein block (de Brevern *et al.*, 2000; de Brevern, 2005; Joseph *et al.*, 2010) is one of the effective structural prototypes of LPS and it has been used in various structure comparison methods and analyses. A protein block (PB) is made up of 16 pentapeptide structural motifs denoted by the letters from *a* to *p*. It is used to represent the three-dimensional structure of a protein in a one-dimensional PB sequence. The PB sequences are similar to one-dimensional sequence representations of secondary structure elements and have been used to find the relationships among protein structures and fragments. Recently, Joseph *et al.* (2010) reviewed interesting applications of PBs in protein structure analysis. A PB-based substitution matrix (Tyagi, Gowri *et al.*, 2006) was developed and has been effectively used for structure-based PB sequence alignment and large-scale structure comparison methods (Tyagi, Sharma *et al.*, 2006; Tyagi *et al.*, 2008). A PB-based web server, Protein Block Expert (PBE; http://bioinformatics.univ-reunion.fr/PBE/), was developed by Tyagi, Sharma *et al.* (2006), and it uses optimal alignment methods to align the given protein structure with the known three-dimensional structures in a database.

Our recent work with the PB is a development of the web-based fold recognition server called PredictFold-PB (Suresh *et al.*, 2012). In this method, we align the predicted PB sequence of a query with a library of assigned PB sequences of 953 known folds using a local pairwise alignment program. Our method uses a PB fold library with 953 assigned PB sequences that belong to 953 folds out of the 1195 folds in the SCOP 1.75 release (Murzin *et al.*, 1995). The remaining folds were not considered because their PDB entries are affected by one of the following: (i) missing amino acids in the ATOM records, (ii) ATOM records interrupted by HETATM records, (iii) non-standard amino acids in ATOM records, (iv) PDB entries with only Cα coordinate information and (v) PDB entries without any backbone coordinate information.

Even though the PBE server is a primary source for PB sequences, it has the following drawbacks: (i) it stores the PB sequences for all the PDB entries including the PDB entries with missing coordinates, (ii) there is no separate link provided for extraction of the dihedral angles φ and ψ, which are used to assign the PB letters, and (iii) only one PB sequence can be extracted at a time.

The limitations of the PBE server and the missing coordinates in some of the PDB entries led us to develop a curated PB sequence database called PDB-2-PB. The current release of the PDB-2-PB database, version 1.0, comprises the PB and amino acid (AA) sequences for 104 631 PDB chain entries. From our database, a user can download the PB sequence of a chain along with its corresponding chain information and AA sequence using two different retrieval options: (i) search using a given PDB code and (ii) search using a given parameter list. The PDB-2-PB database is freely available at http://bioinfo.bdu.ac.in/~pb/.

2. Method

First, the primary PDB entries were downloaded from the latest release of the World Wide Protein Data Bank (wwPDB; Berman *et al.*, 2007), which has 74 297 solved PDB entries as of 7 July 2011.

doi:10.1107/S0021889811052356

Table 1
Summary of PDB entries used from the wwPDB and the PDB chain entries available in the PDB-2-PB database.

No.	Experimental methods	PDB entries in wwPDB	Curated PDB chains in PDB-2-PB
1	X-ray diffraction	64775	93547
2	NMR	8941	7033
3	Other		
	Electron microscopy	371	2449
	Solid-state NMR	43	79
	Electron crystallography	31	50
	Neutron diffraction	39	34
	Fiber diffraction	39	28
	Powder diffraction	18	28
	Solution scattering	35	2
	Infrared spectroscopy	4	2
	Fluorescence	1	0
	Total	74297	103252

Then, the PDB entries with any one of the following omissions in the PDB records were removed: (i) missing amino acids in the ATOM records, (ii) ATOM records interrupted by HETATM records, (iii) non-standard amino acids in ATOM records, (iv) PDB entries with only Cα coordinate information and (v) PDB entries without any backbone coordinate information. The PDB entries solved by solution NMR methods (8941 entries) always have more than one NMR model. In this case, we have manually checked each NMR entry with the OLDERADO (Kelley & Sutcliffe, 1997) server to select the best NMR model. For some NMR entries, the best NMR models were not available in OLDERADO and such entries were not included in our database. This leaves 57% of the PDB entries (42 607 out of the 74 297), which were used to create our PDB-2-PB database.

For each PDB entry, only the chains with complete backbone coordinates were extracted, and they were encoded into 16 PB letters by using the PB encoding procedure reported by de Brevern *et al.* (2000). The encoding procedure that we have used in this work consists of the following steps: (i) extract the chain entry with sufficient backbone coordinates, (ii) splice it into overlapping fragments of five residues in length, (iii) calculate eight dihedral angles (φ, ψ) for each fragment, (iv) calculate root-mean-square deviations on angular values (RMSDa) between observed (φ, ψ) values in the fragment and the ideal (φ, ψ) values (de Brevern *et al.*, 2000) of each one of the 16 PB letters, and (v) determine the PB letter assigned to the centre residue of the fragment, which has the lowest RMSDa value among the 16 RMSDa values. PB letter assignment for the first and last two residues of the chain entry is not possible. Therefore, if the AA sequence of a chain entry has length N, the PB sequence always has length ($N - 4$).

In the present work, we have classified the primary-source PDB entries into three classes, as X-ray, NMR and other, based on their experimental methods. The 'other' class encompasses the PDB entries solved by experimental methods such as neutron, electron and power diffraction, electron crystallography, electron microscopy, fluorescence transfer, and infrared spectroscopy. To create the PDB-2-PB database, all the curated PB sequences of 103 252 PDB chain entries that belong to 42 607 PDB entries were stored within the three classes (93 547 X-ray, 7033 NMR and 2672 other). Table 1 shows a summary of the PDB entries used and the curated PDB chain entries available in our database.

Each chain entry in the PDB-2-PB database stores the following information: (i) PDB code, (ii) chain name, (iii) AA sequence, (iv) PB sequence, (v) chain length, (vi) sequence start and end positions, (vii) resolution (only for X-ray-solved entries), (viii) the experimental method used, and (ix) dihedral angles (φ and ψ). We have developed our database with a user-friendly sequence retrieval system using the following two options. A user can extract the PB sequence and other related information by using either a given PDB code or a list of the parameters available in our database.

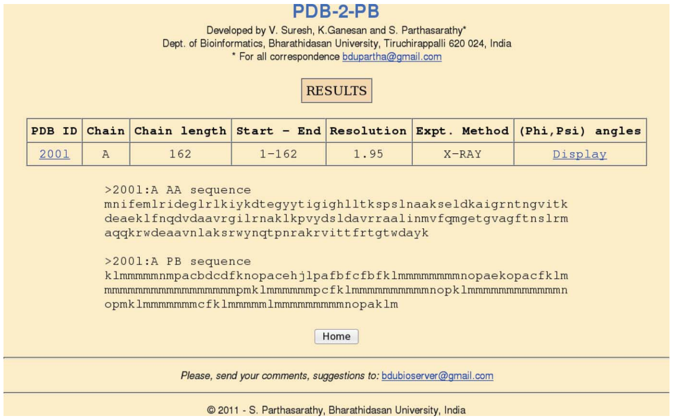
3. Description of the PDB-2-PB database

3.1. Search using a given PDB code

The server allows the user to query the PDB-2-PB database using a given PDB code. This results in the display of the following information for all the chains present in the given PDB entry: (i) PDB code, (ii) chain name, (iii) AA sequence, (iv) PB sequence, (v) chain length, (vi) sequence start and end positions, (vii) resolution (if any), and (viii) the experimental method used. There is a 'Display' link to extract the corresponding dihedral angles (φ and ψ). Moreover, each chain entry is provided with a PDB link to access more structural information. If the given PDB entry is not stored in our database, an error message will be displayed. As an example, Fig. 1 displays the PB sequence for PDB entry 200l (Baldwin *et al.*, 1996), along with its corresponding chain information and AA sequence, extracted by this server.

3.2. Search using a given parameter list

The server also allows the user to extract multiple entries as output from the PDB-2-PB database using a given parameter list. The various options available to make the parameter list are (i) experimental methods (X-ray, NMR and other), (ii) PDB entries with a maximum number of chains, (iii) PDB chain entries with a maximum number of residues and (iv) resolution of PDB entries (only applicable for X-ray). After the parameters have been given, there is a display page to confirm the user parameter list. If the parameter list is correct, the user can click the 'Search' button to extract the multiple PDB chain entries from the PDB-2-PB database. The number of chain entries in each page and the total number of chain entries extracted for the given parameter list will be displayed at the top of the results table. For each chain entry, the user will get similar information as output as mentioned in §3.1. In order to get the (i) dihedral angles (φ and ψ) and (ii) AA and PB sequences, the user can use the 'Display' links. Fig. 2 displays the 'Results' page extracted by this server using the default parameter list: (i) experimental method X-ray, (ii) PDB entries with number of chains $1 \le c \le 2$, (iii) PDB



Developed by V. Suresh, K.Ganesan and S. Parthasarathy*
Dept. of Bioinformatics, Bharathidasan University, Tiruchirappalli 620 024, India
* For all correspondence bdupartha@gmail.com

RESULTS

PDB ID	Chain	Chain length	Start - End	Resolution	Expt. Method	(Phi,Psi) angles
200l	A	162	1-162	1.95	X-RAY	Display

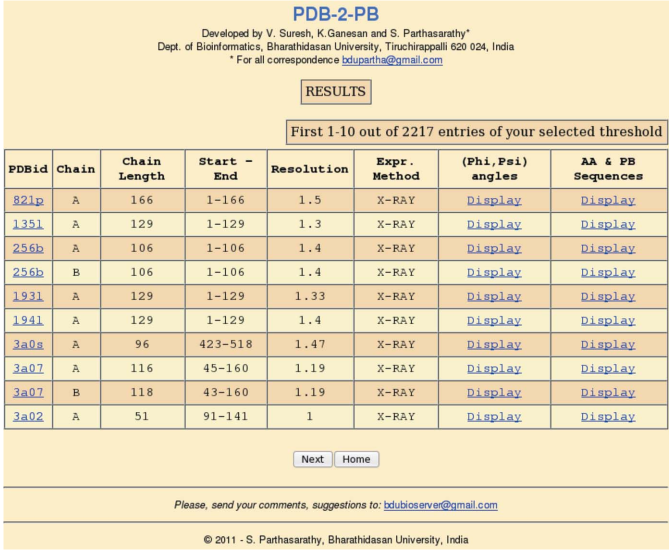
```
>200l:A AA sequence
mnifemlrideglrlkiykdtegyytigighlltkspslnaakseldkaigrntngvitk
deaeklfnqdvdaavrgilrnaklkpvydsldavrraalinmvfqmgetgvagftnslrm
aqqkrwdeaavnlaksrwynqtpnrakrvittfrtgtwdayk

>200l:A PB sequence
klmmmmmnmpacbdcdfknopacehjlpafbfcfbfklmmmmmmmmnopaekopacfklm
mmmmmmmmmmmmmmmmmpmklmmmmmmpcfklmmmmmmmmmmnopklmmmmmmmmmmmmn
opmklmmmmmmmcfklmmmmmlmmmmmmmmmnopaklm
```

Please, send your comments, suggestions to: bdubioserver@gmail.com

© 2011 - S. Parthasarathy, Bharathidasan University, India

Figure 1
'Results' page, displaying an example PB sequence of a chain, along with its corresponding chain information and AA sequence.



PDB-2-PB

Developed by V. Suresh, K.Ganesan and S. Parthasarathy*
Dept. of Bioinformatics, Bharathidasan University, Tiruchirappalli 620 024, India
* For all correspondence bdupartha@gmail.com

RESULTS

First 1-10 out of 2217 entries of your selected threshold

PDBid	Chain	Chain Length	Start - End	Resolution	Expr. Method	(Phi,Psi) angles	AA & PB Sequences
821p	A	166	1-166	1.5	X-RAY	Display	Display
1351	A	129	1-129	1.3	X-RAY	Display	Display
256b	A	106	1-106	1.4	X-RAY	Display	Display
256b	B	106	1-106	1.4	X-RAY	Display	Display
1931	A	129	1-129	1.33	X-RAY	Display	Display
1941	A	129	1-129	1.4	X-RAY	Display	Display
3a0s	A	96	423-518	1.47	X-RAY	Display	Display
3a07	A	116	45-160	1.19	X-RAY	Display	Display
3a07	B	118	43-160	1.19	X-RAY	Display	Display
3a02	A	51	91-141	1	X-RAY	Display	Display

Please, send your comments, suggestions to: bdubioserver@gmail.com

© 2011 - S. Parthasarathy, Bharathidasan University, India

Figure 2
'Results' page of PDB chain entries from the PDB-2-PB database for an example user-supplied parameter list.

chain entries with number of residues $5 \le n \le 300$, and (iv) resolution of PDB entries $0.68 \le r \le 1.5$.

4. Database implementation and access

The PDB entries were downloaded from the wwPDB by using a shell script developed in Linux. Extraction of coordinates, calculation of dihedral angles and encoding of PB letters were performed with Perl scripts. The database component of PDB-2-PB is implemented with MySQL under the Fedora 9 Linux operating system. The 'Home', 'Query' and 'Results' pages of PDB-2-PB were designed using HTML. Validation of input parameters and the search using the parameter list within MySQL are carried out with a PHP script. The database can be freely accessed at http://bioinfo.bdu.ac.in/~pb/. Users of PDB-2-PB are requested to cite this article or the URL of our database in their research work. Furthermore, PDB-2-PB will be updated periodically. Comments and suggestions can be e-mailed to Dr S. Parthasarathy (bdubioserver@gmail.com).

5. Advantages of PDB-2-PB

The PDB-2-PB database gives only the curated chain entries for structure comparison studies in terms of PB sequences. The dihedral angles used to convert the PB letters are also given, along with corresponding amino acid and protein block letter codes. This information will be useful for users intending to carry out further structural and conformational analysis. Furthermore, the search using parameter selection is very helpful for users wishing to short-list PDB entries based on their research requirements. Our server will be updated as and when a new snapshot of the PDB is released by the wwPDB.

6. Limitations

The current release of the PDB-2-PB database (version 1.0) covers only 57% of the PDB entries (*i.e.* 42 607 out of 74 297 entries) available in the wwPDB. The remaining entries were not considered because of missing records, as mentioned in §2. Moreover, the new release of PDB-2-PB requires manual processing to select the best NMR model from OLDERADO (Kelley & Sutcliffe, 1997).

References

Baldwin, E., Xu, J., Hajiseyedjavadi, O., Baase, W. A. & Matthews, B. W. (1996). *J. Mol. Biol.* **259**, 542–559.

Berman, H., Henrick, K., Nakamura, H. & Markley, J. L. (2007). *Nucleic Acids Res.* **35**, D301–D303.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Brevern, A. G. de (2005). *In Silico Biol.* **5**, 283–289.

Brevern, A. G. de, Etchebest, C. & Hazout, S. (2000). *Proteins*, **41**, 271–287.

Gibrat, J. F., Madej, T. & Bryant, S. H. (1996). *Curr. Opin. Struct. Biol.* **6**, 377–385.

Jones, T. A. & Thirup, S. (1986). *EMBO J.* **5**, 819–822.

Joseph, A. P., Agarwal, G., Mahajan, S., Gelly, J. C., Swapna, L. S., Offmann, B., Cadet, F., Bornot, A., Tyagi, M., Valadié, H., Schneider, B., Etchebest, C., Srinivasan, N. & de Brevern, A. G. (2010). *Biophys. Rev.* **2**, 137–147.

Kelley, L. A. & Sutcliffe, M. J. (1997). *Protein Sci.* **6**, 2628–2630.

Levitt, M. (1992). *J. Mol. Biol.* **226**, 507–533.

Lu, G. (2000). *J. Appl. Cryst.* **33**, 176–183.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Offmann, B., Tyagi, M. & de Brevern, A. G. (2007). *Curr. Bioinf.* **3**, 165–202.

Singh, A. P. & Brutlag, D. L. (1997). *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **5**, 284–293.

Suresh, V., Ganesan, K. & Parthasarathy, S. (2012). *Protein Pept. Lett.* In the press.

Tyagi, M., de Brevern, A. G., Srinivasan, N. & Offmann, B. (2008). *Proteins*, **71**, 920–937.

Tyagi, M., Gowri, V. S., Srinivasan, N., de Brevern, A. G. & Offmann, B. (2006). *Proteins*, **65**, 32–39.

Tyagi, M., Sharma, P., Swamy, C. S., Cadet, F., Srinivasan, N., de Brevern, A. G. & Offmann, B. (2006). *Nucleic Acids Res.* **34**, W119–W123.

Unger, R., Harel, D., Wherland, S. & Sussman, J. L. (1989). *Proteins*, **5**, 355–373.