W428–W434 *Nucleic Acids Research, 2012, Vol. 40, Web Server issue* *Published online 11 June 2012*
doi:10.1093/nar/gks527

PBSword: a web server for searching similar protein–protein binding sites

Bin Pang[1], Xingyan Kuang[1], Nan Zhao[1], Dmitry Korkin[1,2] and Chi-Ren Shyu[1,2,*]

[1]Informatics Institute and [2]Department of Computer Science, University of Missouri, Columbia, MO, USA

Received March 3, 2012; Revised May 9, 2012; Accepted May 11, 2012

ABSTRACT

PBSword is a web server designed for efficient and accurate comparisons and searches of geometrically similar protein–protein binding sites from a large-scale database. The basic idea of PBSword is that each protein binding site is first represented by a high-dimensional vector of 'visual words', which characterizes both the global and local shape features of the binding site. It then uses a scalable indexing technique to search for those binding sites whose visual words representations are similar to that of the query binding site. Our system is able to return ranked results of binding sites in short time from a database of 194 322 domain–domain binding sites. PBSword supports query by protein ID and by new structures uploaded by users. PBSword is a useful tool to investigate functional connections among proteins based on the local structures of binding site and has potential applications to protein–protein docking and drug discovery. The system is hosted at http://pbs.rnet.missouri.edu.

INTRODUCTION

Determining similar protein–protein binding site (PBS) plays an important role in understanding protein–protein interaction mechanisms (1) and has a potential impact on protein function prediction, protein–protein docking, drug discovery and evolutionary studies (2–6). As the size of data repositories of protein–protein interfaces continue to grow, numerous databases have been developed to organize and classify the interaction data at different subunit levels (chain or domain) (7,8). Some recent examples of databases include SCOPPI (9), PIBASE (10), IntAct (11), DOMMINO (12), iPfam (13) and SCOWLP (14). These databases can usually provide basic services of looking up a specific protein–protein interface according to the identification of protein or a group of interfaces based on the classification of protein family [e.g. SCOP (15) or CATH (16)]. In this case, similar binding sites are retrieved in terms of overall sequence and fold similarity of protein, which might not fulfill requirements of the binding site comparison and functional annotation, as a similar fold does not necessarily imply a similar function, and proteins of different folds may acquire similar functions. In contrast with the overall fold similarities, local structure of the binding site is highly possible to be connected to the functions of the protein (17,18). However, as the structure comparison of a query binding site against a large-scale database of binding sites can be very challenging and time consuming. It is a pressing need for the community to have an access to advanced services of searching similar functional sites for a newly discovered or existing protein based on the local patterns of binding site.

One of the most important tasks in constructing such a structure-based search engine is to develop an efficient and accurate method for comparison of binding sites. Early research works mainly used global sequence and structure alignment tools. However, these tools usually concentrated on the similarities of entire protein and may ignore the local structures of binding site. To overcome this issue, one cluster of approach is based on the alignment of local structures or functional groups on the protein surface [e.g. iAlign (19) and I2ISiteEngine (20)] to provide accurate comparison between binding sites. This approach is usually computationally expensive. To accelerate this process, another cluster aims to compare binding sites with extracted features [e.g. distance distribution (21) or moment invariant (22)] of surfaces or structure without explicit alignments. These methods, mainly designed for protein-ligand binding sites (23), have not been extensively evaluated on the datasets of protein–protein binding sites, which are known to have some unique characteristics, such as relatively large and planar surfaces (1).

To meet the challenges of efficiency and accuracy requirements, PBSword is developed to provide the community a web server for searching similar protein binding sites in terms of 'visual words'. The basic idea of PBSword is originated from the classic method developed in

*To whom correspondence should be addressed. Tel: +1 573 882 3884; Fax: +1 573 884 8709; Email: shyuc@missouri.edu

information retrieval area for comparing the similarity of documents based on the word frequency profiles, which has been successfully applied in web search engines. In PBSword server, we further extend the text comparison method and propose a novel approach, which integrates frequency of visual words as well as local spatial relationships among them, to represent the protein binding sites. By loading the visual words representations of database binding sites into a scalable indexing tree, PBSword server can achieve high-throughput while preserving reasonably high precision of binding site comparison.

The key features of PBSword server include the following: (i) The binding site comparison method introduces a novel feature extraction algorithm and online database indexing; (ii) the database of binding site is based on the interactions between domains which are defined using the latest SCOP version (24); (iii) for each retrieved binding site from the database, a 3-dimensional (3D) view of structure and surface, as well as physicochemical properties are presented; (iv) the efficiency has been significantly enhanced to meet the requirements of large-scale protein binding site database searching.

MATERIALS AND METHODS

The system architecture of PBSword server, as shown in Figure 1, contains four modules: (i) database management and preprocessing; (ii) query interfaces; (iii) search engine and (iv) retrieval results visualization. A system tutorial can be viewed at the PBSword website.

Database management and preprocessing

The database of PBSword contains domain–domain binding sites of known protein structures. The structural data are extracted from Protein Data Bank (PDB) (25). If a PDB entry has more than one structure model, the first model is used in the database's current implementation. For domain assignment, the most recent release (June 2009) of manually curated SCOP database is used. For each PDB structure, each pair of determined subunits (i.e. domains) is analysed to determine whether they interact with each other using the following definition. If any atom of a residue in one protein subunit is within 6 Å of any atom of a residue in another protein subunit, the two residues are determined as the contact pair residues.

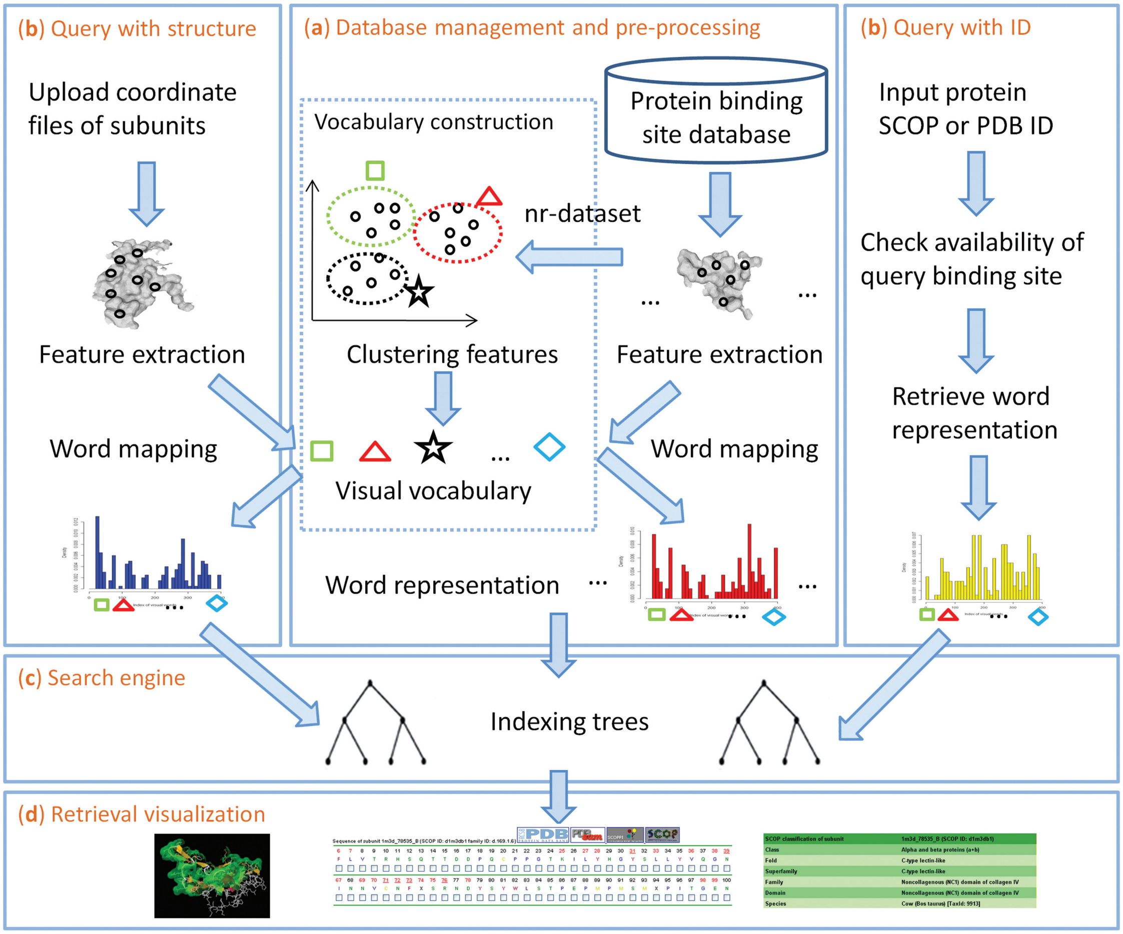


Figure 1. PBSword system architecture. (**a**) The database management and preprocessing module is responsible for feature extraction, visual vocabulary construction and word representation of the database binding sites, which can be performed offline. (**b**) The query interface modules provide friendly interfaces in an Internet browser to allow users to input protein ID or upload protein structure. (**c**) The search engine module organizes the word representation of the database binding site into indexing tree and returns *n* nearest neighbors for a query binding site in real-time. (**d**) The retrieval visualization module shows 3D structure/surface view, sequence and properties of retrieved binding sites.

Currently, the entire PBSword database contains 194 322 redundant binding sites selected from 3123 SCOP families. Two nonredundant (nr) databases, denoted as NR40 and NR60, are constructed using sequence similarity of 40% and 60%, respectively.

The workflow of database preprocessing consists of the following three steps (see top-middle block of Figure 1). First, we select feature points from each database binding site surface and extract corresponding geometric features. Second, a visual vocabulary is built by clustering a large number ($\sim 7 \times 10^5$) of feature point descriptors collected from nr dataset. The nr dataset is selected from the entire database by applying a cutoff of 40% sequence identity for each SCOP family using Cluster Database at High Identity with Tolerance (CD-HIT) (26). The clustering method is k-means and each feature cluster is represented by a representative, which is regarded as a visual word and used to form the final vocabulary. The size of vocabulary is determined by k, which is set to 1000 in the PBSword server. Third, according to its descriptor, each feature point from the database binding site surface is associated with the nearest visual word from the vocabulary. This allows each binding site to be represented by the corresponding distribution of visual words. It is noted that the aforementioned processes for the database binding sites are performed offline. Owing to page limitations, interested readers are referred to our article of algorithm for further details and discussions (27).

Query interfaces

There are two types of query methods, 'query by structure' and 'query by ID', as shown in the top-left and top-right blocks of Figure 1, respectively. Using an Internet browser, a user can upload a new protein structure in PDB format or provide a protein ID contained in a PBSword database to find similar protein binding sites. The target database could be (i) redundant; (ii) NR40 or (iii) NR60.

For the query by structure search, we follow the similar steps as the database binding sites to extract its features, map the features to the nearest visual word and generate the visual word representation. The word representation of the query binding site is then sent to the search engine.

For the query by protein ID search, users can provide (i) SCOP IDs for the interacting subunits or (ii) PDB ID and chain ID for the subunit under investigation. For the second option, chain ID of interacting partner is optional. In that case, PBSword will search the database to find matched binding site and allow user to select one from the matched list. After the query binding site is selected, the corresponding word representation is then sent to the search engine.

Search engine

When the redundant PBSword database is selected as target, the online binding site search is performed on two customized indexing trees to avoid time-consuming one-by-one feature similarity calculation for the two query methods, namely query by ID and query by 3D structure. In this case, the query protein binding site can be represented by a data point in the visual word (or feature) space populated by the database binding sites as mentioned in the previous two subsections. Thus, searching similar binding site from the database is analogous to the identification of n nearest neighbors in the feature space. Such a search can be completed in log(N) time, where N is the total number of binding sites in the database. When the NR40 or NR60 database is selected, binding site similarity and associated z-score are calculated for each database binding site.

Retrieval results visualization

The visualization of retrieval result includes seven parts: (i) structure and surface display, (ii) ranked list, (iii) sequence, (iv) SCOP classification, (v) properties of binding site, (vi) properties of each binding site residue and (vii) property statistics of a SCOP family. The properties of binding site mainly include accessible surface area (ASA), polarity, hydrophobicity, hydrogen bonds, planarity and gap index, which are originally defined in (28,29). For completeness, we briefly introduce the calculation of these properties as follows. The ASA of binding site, ASA_{bs}, is calculated as follows:

$$ASA_{bs} = ASA_1 - ASA_2,$$

where ASA_1 and ASA_2 are the ASAs of the subunit before and after its interacting partner presents, respectively. The ASA is calculated using the NACCESS (30), an implementation of method proposed in (31). A residue is defined as binding site residue if it loses 1.0 Å^2 of ASA after subunit partner presents. The polarity of binding site is defined as follows (28):

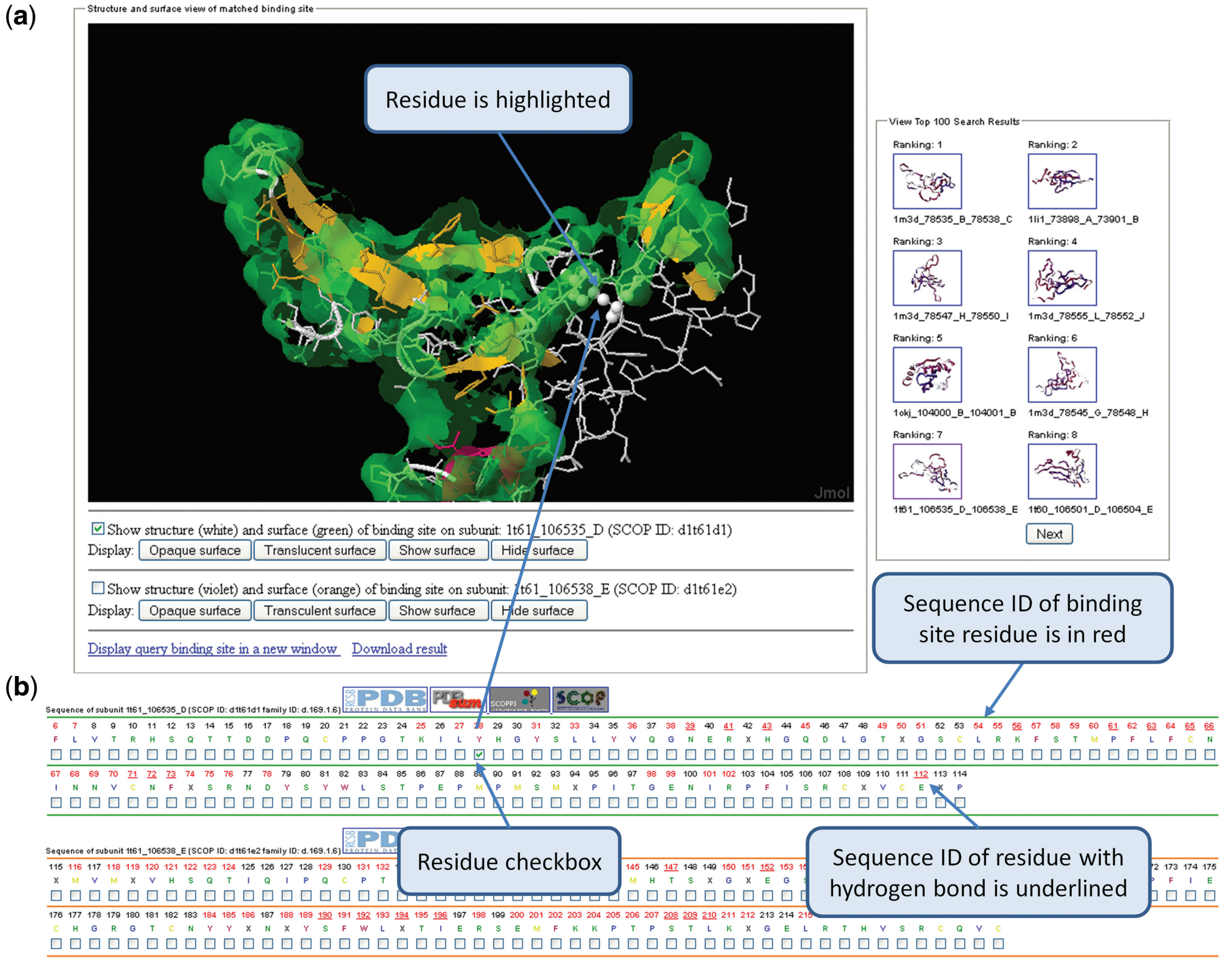
$$\text{polarity} = \left(ASA_{polar}/ASA_{bs}\right) \times 100$$

where ASA_{polar} represents the difference of ASA of polar atoms before and after interacting partner presents. The hydrophobicity is measured using method proposed in (29). The number of hydrogen bonds is calculated using the program HBPLUS (32). The planarity is defined as the root mean squared deviation between all binding site atoms and a best-fit plane through all the binding site atoms, which is calculated using the PRINCIP program from the SURFNET package (33). The gap index is defined as follows (28):

$$\text{gap index} = \text{gap volume}/ASA_{bs}$$

where gap volume is a measure of the closeness of the interface between the two subunits and calculated using the SURFNET package (33).

The retrieval results for an example query binding site 1m3d_78535_B_78538_C are shown in the top-left panel of Figure 2a. In PBSword, we use the identifier same as (22) to name each binding site: <PDB-ID>_<SCOP-domain of the binding site>_<Chain-ID of the binding site>_<SCOP-domain of the binding partner>_<Chain-ID of the binding partner>. Accordingly, each subunit is defined as <PBD_ID>_<SCOP-domain ID>_<Chain-ID>. For each query, a set of 100 top-ranked binding sites is returned to the user, eight at a time for


(a)
Structure and surface view of matched binding site
Residue is highlighted
Jmol
View Top 100 Search Results
Ranking: 1
1m3d_78535_B_78538_C
Ranking: 2
1li1_73898_A_73901_B
Ranking: 3
1m3d_78547_H_78550_I
Ranking: 4
1m3d_78555_L_78552_J
Ranking: 5
1okj_104000_B_104001_B
Ranking: 6
1m3d_78545_G_78548_H
Ranking: 7
1t61_106535_D_106538_E
Ranking: 8
1t60_106501_D_106504_E
Next
Display: Opaque surface Translucent surface Show surface Hide surface
Display: Opaque surface Transculent surface Show surface Hide surface
Display query binding site in a new window
Download result
Sequence ID of binding site residue is in red
(b)
PDB
PDBsum
SCOPPI
SCOP
Residue checkbox
Sequence ID of residue with hydrogen bond is underlined

☑ Show structure (white) and surface (green) of binding site on subunit: 1t61_106535_D (SCOP ID: d1t61d1)

□ Show structure (violet) and surface (orange) of binding site on subunit: 1t61_106538_E (SCOP ID: d1t61e2)

Sequence of subunit 1t61_106535_D (SCOP ID: d1t61d1 family ID: d.169.1.6)

```
6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66
F L V T R H S Q T T D D P Q C P P G T K I L Y H G Y S L L Y V Q G N E R X H G Q D L G T X G S C L R K F S T M P F L F C N
67 68 69 70 71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114
I N N V C N F X S R N D Y S Y W L S T P E P M P M S M X P I T G E N I R P F I S R C X V C E X P
```

Sequence of subunit 1t61_106538_E (SCOP ID: d1t61e2 family ID: d.169.1.6)

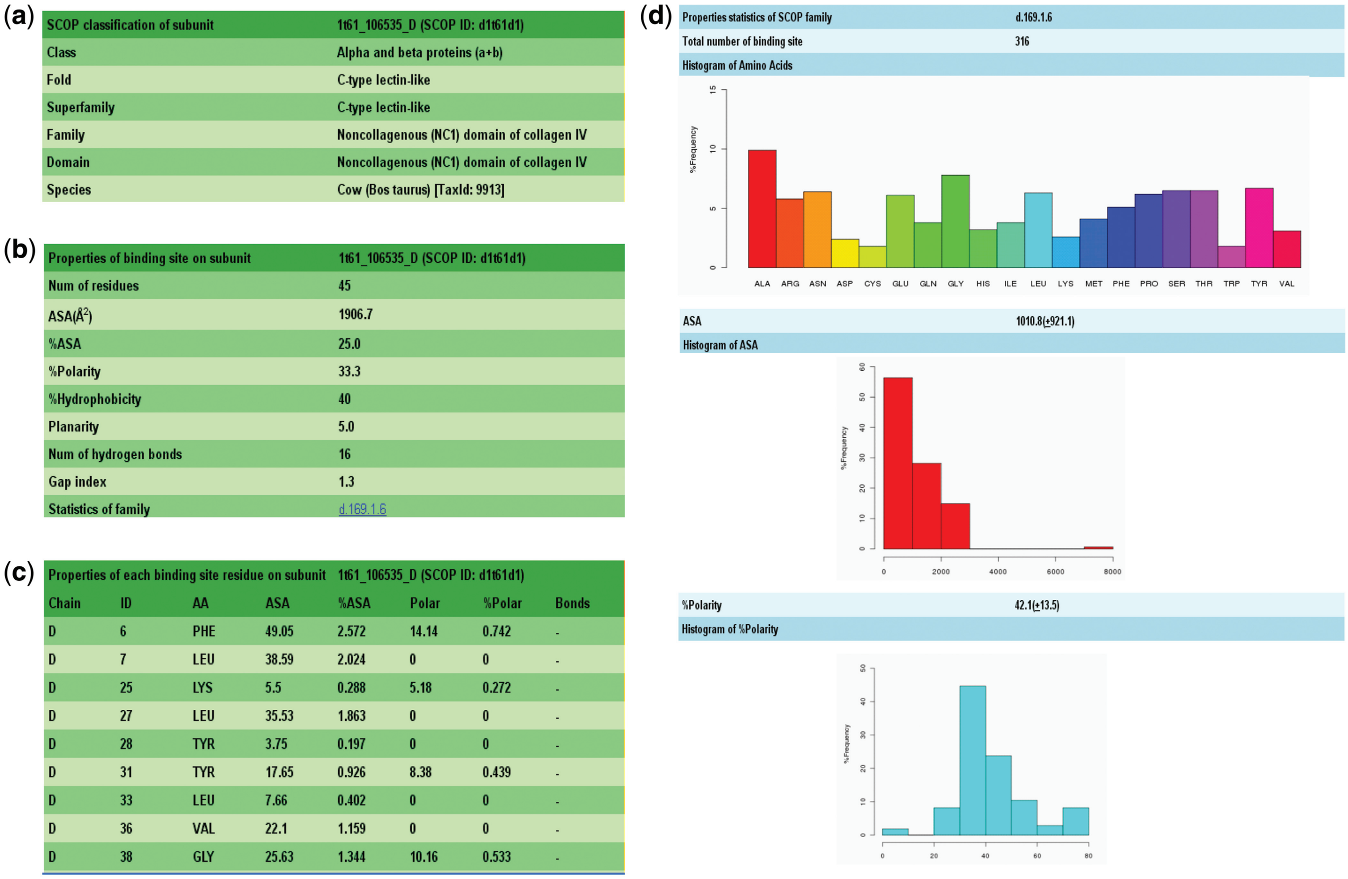
```
115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132       145 146 147 148 149 150 151 152 153       173 174 175
X M V M X V H S Q T I Q I P Q C P T       M H T S X G X E G       F I E
176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215
C H G R G T C N Y Y X N X Y S F W L X T I E R S E M F K K P T P S T L K X G E L R T H V S R C Q V C
```

Figure 2. PBSword retrieval results visualization. (**a**) The top-left panel shows a 3D structure and surface view of a selected result protein-binding site from the ranked list in the top-right panel. Users can click on the buttons and checkboxes in the top-left panel to select binding site and its partner as well as display modes of surface. (**b**) The sequence panel shows sequence information of subunit pairs. Each column in the panel corresponds to an amino acid of protein subunit, which consists of three rows. First row represents residue sequence number. For binding site residue, its number is in red font. For the residue with intermolecular hydrogen bonds, it numbers is underlined. Second row is the residue name. The third row is residue check box. By clicking a checkbox, corresponding residues will be shown in the top-left structure view.

each page. To visualize the search results, a 3D structure and surface view of the top-retrieval result is displayed to the user. The user can select any of the ranked results from the top-right panel. The top-left panel in Figure 2a presents the structure and surface view of the top-ranked result, 1t61_106535_D_106538_E, which is generated by clicking on the thumbnail image on the top-right panel. In addition, the users can (i) select to show/hide structures of two subunits by clicking the checkboxes and (ii) specify different display themes of binding site, such as opaque/translucent surface, by clicking on the buttons. The ranked list of protein binding sites can be downloaded from the result pages.

The sequence panel (Figure 2b) shows the sequence information of a subunit and its partner. For easy identification, the binding site residues are shown in red font and the residues with intermolecular hydrogen bond are underlined. The users can use the 'residue checkbox' under the residue to interact with the 3D structure view shown in Figure 2a. Clicking on the 'residue checkbox' will highlight one designated residue. Hyperlinks pointing to the protein's corresponding entry in PDB, PDBSum (34), SCOPPI (9) and SCOP (24) are also provided.

The SCOP classification panel (Figure 3a) shows the description of corresponding SCOP class, fold, superfamily, family and species for two subunits. The properties of binding site and its interacting partner are shown in Figure 3b, including the number of binding site residues, ASA, percentage of ASA, percentage of polarity, percentage of hydrophobicity, planarity, number of hydrogen bonds and gap index. By clicking on the hyperlink of SCOP family at the row 'Statistics of family', user can view the histogram and summary statistics of each property by SCOP family (see Figure 3d). The properties of each binding site residue, shown in Figure 3c, include ASA of all atoms and polar atoms for a specific residue and percentage of ASA against the entire binding site ASA, as well as the number of intermolecular hydrogen bonds. The family statistics panel (Figure 3d) shows the statistics summary of properties of binding sites belonging to a SCOP family, including total number of binding sites, amino acids compositions, as well as the mean



SCOP classification of subunit	1t61_106535_D (SCOP ID: d1t61d1)
Class	Alpha and beta proteins (a+b)
Fold	C-type lectin-like
Superfamily	C-type lectin-like
Family	Noncollagenous (NC1) domain of collagen IV
Domain	Noncollagenous (NC1) domain of collagen IV
Species	Cow (Bos taurus) [TaxId: 9913]

Properties of binding site on subunit	1t61_106535_D (SCOP ID: d1t61d1)
Num of residues	45
ASA(Å^2)	1906.7
%ASA	25.0
%Polarity	33.3
%Hydrophobicity	40
Planarity	5.0
Num of hydrogen bonds	16
Gap index	1.3
Statistics of family	d.169.1.6

Properties of each binding site residue on subunit				1t61_106535_D (SCOP ID: d1t61d1)			
Chain	ID	AA	ASA	%ASA	Polar	%Polar	Bonds
D	6	PHE	49.05	2.572	14.14	0.742	-
D	7	LEU	38.59	2.024	0	0	-
D	25	LYS	5.5	0.288	5.18	0.272	-
D	27	LEU	35.53	1.863	0	0	-
D	28	TYR	3.75	0.197	0	0	-
D	31	TYR	17.65	0.926	8.38	0.439	-
D	33	LEU	7.66	0.402	0	0	-
D	36	VAL	22.1	1.159	0	0	-
D	38	GLY	25.63	1.344	10.16	0.533	-

Properties statistics of SCOP family	d.169.1.6
Total number of binding site	316
Histogram of Amino Acids	

ASA	1010.8(±921.1)
Histogram of ASA	

%Polarity	42.1(±13.5)
Histogram of %Polarity	

Figure 3. PBSword retrieval results of binding site properties. (**a**) The SCOP classification panel shows the subunit's classification, including class, fold, superfamily, family and species. (**b**) The site properties panel shows values of various physicochemical properties of binding sites, including number of residues, ASA, percentage of ASA, percentage of polarity, percentage of hydrophobicity, planarity, number of hydrogen bonds and gap index. In addition, users can click on the hyperlink of SCOP family at the row "Statistics of family" to view the statistics summary of these properties for those binding sites belonging to same SCOP family. (**c**) The residue properties panel shows detailed properties for each binding site residue, including ASA, percentage of ASA, ASA of polar atoms, percentage of polar residue and number of hydrogen bonds. (**d**) The family statistics panel shows the statistics summary of binding sites from a SCOP family, including number of binding sites, amino acids compositions, as well as the mean (standard deviation) and histogram of binding site properties. The properties include ASA, percentage of polarity, percentage of hydrophobicity, planarity, hydrogen bonding and gap index. Here, hydrogen bonding is defined as the number of hydrogen bonds per 100 Å^2 ASA. Owing to the page limitation, we have only shown subset of each panel.

(standard deviation) and histogram of binding site properties. The properties include ASA, percentage of polarity, percentage of hydrophobicity, planarity, hydrogen bonding and gap index. In this panel, hydrogen bonding is defined as the number of hydrogen bonds per 100 Å^2 ASA.

For a search with query ID, PBSword retrieval results can be generated in real-time. For the query with protein structures, however, the system will usually take minutes to generate surface and extract features, which is dependent on the size of the query binding site. Our system provides the following two options for the users: (i) PBsword server will return a session ID for the query along with an estimated execution time after the query protein structure has been uploaded. The user can then bookmark the link of the session ID and check the resulting page a few minutes later (ii) If the user is willing to provide an email address when the query protein structure is uploaded, PBSword server will send ranked results to the user's email account.

Performance evaluation

We applied the PBSword algorithm to SCOPPI binding site classification and compared its performance with a feature-based method, moment invariants (MI) (22), and an alignment-based method, iAlign (19), on an nr database of 2819 protein binding sites selected from SCOPPI 1.69 (22). Our experimental results show that PBSword algorithm can achieve comparable classification accuracy with iAlign and improve accuracy of MI by 36% on the nr dataset. Simultaneously, PBSword algorithm exhibits a significant efficiency improvement over the alignment-based method. For example, PBSword algorithm takes 0.31 second for a one-against-all search on the nr dataset, whereas iAlign spends 1016 seconds on a complete scan. In PBSword server, the efficiency has been further enhanced by using the indexing trees to organize the visual words representations of database, which can efficiently retrieve top 100 best matched sites for a query binding site without exhaustively performing one-against-all comparisons over all the 194 322 binding sites in the database.

DISCUSSION

Searching similar protein binding sites from a large-scale dataset is extremely important for various biological applications. The PBSword web server presented in this article comes equipped with an efficient and accurate search engine with a user-friendly interface and an informative retrieval result visualization design. Our server can return retrieval results in short time while preserving high accuracy. It is expected that this web server will be beneficial to the life sciences community by revealing functional and evolutionary connections between proteins based on the local similarity of binding site.

We finally emphasize that PBSword, as a feature-based method for comparing similarity of binding sites, is not designed to be a replacement of existing alignment-based methods (e.g. iAlign). Instead, it works as a complementary approach to the structure comparison methods and offers an efficient way to filter out dissimilar binding sites.

ACKNOWLEDGEMENTS

The authors are grateful to the researchers and groups who made the following software packages and databases available for us to use in PBSword: the Jmol package (http://jmol.sourceforge.net/) for generating view of proteins, the SCOP for constructing database, MSMS/NACESS/HBPLUS/SURFNET software for generating surface and properties of binding sites and the PDB (http://www.pdb.org) for maintaining the tertiary structures. The authors acknowledge funding from National Science Foundation.

FUNDING

National Science Foundation [DBI-0845196 to D.K.]; Shumaker Endowment in Bioinformatics (to B.P., X. K. and N.Z.). Funding for open access charge: University of Missouri Shumaker Endowment in Bioinformatics.

Conflict of interest statement. None declared.

REFERENCES

1. Bahadur,R. and Zacharias,M. (2008) The interface of protein-protein complexes: analysis of contacts and prediction of interactions. *Cell. Mol. Life Sci.*, **65**, 1059–1072.
2. Bradford,J., Needham,C., Bulpitt,A. and Westhead,D. (2006) Insights into protein–protein interfaces using a Bayesian Network Prediction method. *J. Mol. Biol.*, **362**, 365–386.
3. Henschel,A., Kim,W. and Schroeder,M. (2006) Equivalent binding sites reveal convergently evolved interaction motifs. *Bioinformatics*, **22**, 550–555.
4. Tuncbag,N., Gursoy,A., Guney,E., Nussinov,R. and Keskin,O. (2008) Architectures and functional coverage of protein-protein interfaces. *J. Mol. Biol.*, **381**, 785–802.
5. Wu,C.H., Huang,H., Nikolskaya,A., Hu,Z. and Barker,W.C. (2004) The iProClass integrated database for protein functional analysis. *Comput. Biol. Chem.*, **28**, 87–96.
6. Zhao,N., Pang,B., Shyu,C.R. and Korkin,D. (2011) Structural similarity and classification of protein interaction interfaces. *PLoS One*, **6**, e19554.
7. Tuncbag,N., Kar,G., Keskin,O., Gursoy,A. and Nussinov,R. (2009) A survey of available tools and web servers for analysis of protein-protein interactions and interfaces. *Brief Bioinform.*, **10**, 217–232.
8. De Las Rivas,J. and Fontanillo,C. (2010) Protein-protein interactions essentials: key concepts to building and analyzing interactome networks. *PLoS Comput. Biol.*, **6**, e1000807.
9. Winter,C., Henschel,A., Kim,W.K. and Schroeder,M. (2006) SCOPPI: a structural classification of protein-protein interfaces. *Nucleic Acids Res.*, **34**, D310–D314.
10. Davis,F.P. and Sali,A. (2005) PIBASE: a comprehensive database of structurally defined protein interfaces. *Bioinformatics*, **21**, 1901–1907.
11. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
12. Kuang,X., Han,J.G., Zhao,N., Pang,B., Shyu,C.R. and Korkin,D. (2012) DOMMINO: a database of macromolecular interactions. *Nucleic Acids Res.*, **40**, D501–D506.
13. Finn,R.D., Marshall,M. and Bateman,A. (2005) iPfam: visualization of protein-protein interactions in PDB at domain and amino acid resolutions. *Bioinformatics*, **21**, 410–412.
14. Teyra,J., Doms,A., Schroeder,M. and Pisabarro,M.T. (2006) SCOWLP: a web-based database for detailed characterization and visualization of protein interfaces. *BMC Bioinformatics*, **7**, 104.
15. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
16. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
17. Yin,S., Proctor,E.A., Lugovskoy,A.A. and Dokholyan,N.V. (2009) Fast screening of protein surfaces using geometric invariant fingerprints. *Proc. Natl. Acad. Sci. U S A.*, **106**, 16622–16626.
18. Gao,M. and Skolnick,J. (2010) Structural space of protein-protein interfaces is degenerate, close to complete, and highly connected. *Proc. Natl. Acad. Sci. U S A.*, **107**, 22517–22522.
19. Gao,M. and Skolnick,J. (2010) iAlign: a method for the structural comparison of protein-protein interfaces. *Bioinformatics*, **26**, 2259–2265.
20. Shulman-Peleg,A., Mintz,S., Nussinov,R. and Wolfson,H. (2004) In: Jonassen,I. and Kim,J. (eds), *Algorithms in Bioinformatics*, Vol. 3240. Springer, Berlin/Heidelberg, pp. 194–205.
21. Das,S., Kokardekar,A. and Breneman,C.M. (2009) Rapid comparison of protein binding site surfaces with property encoded shape distributions. *J. Chem. Inf. Model*, **49**, 2863–2872.
22. Sommer,I., Muller,O., Domingues,F., Sander,O., Weickert,J. and Lengauer,T. (2007) Moment invariants as shape recognition technique for comparing protein binding sites. *Bioinformatics*, **23**, 3139–3146.
23. Das,S., Krein,M.P. and Breneman,C.M. (2010) PESDserv: a server for high-throughput comparison of protein binding site surfaces. *Bioinformatics*, **26**, 1913–1914.
24. 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.
25. Berman,H.M. (2008) The Protein Data Bank: a historical perspective. *Acta Crystallogr. A*, **64**, 88–95.
26. Li,W., Jaroszewski,L. and Godzik,A. (2001) Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, **17**, 282–283.
27. Pang,B., Zhao,N., Korkin,D. and Shyu,C.-R. (2012) Fast protein binding site comparisons using visual words representation. *Bioinformatics*, **28**, 1345–1352.
28. Jones,S., Marin,A. and Thornton,J.M. (2000) Protein domain interfaces: characterization and comparison with oligomeric protein interfaces. *Protein Eng.*, **13**, 77–82.

29. Jones,S. and Thornton,J.M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.
30. Hubbard,S.J. and Thornton,J.M. (1993), 'NACCESS', computer program.
31. Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–400.
32. McDonald,I.K. and Thornton,J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.*, **238**, 777–793.
33. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330, 307–308.
34. Laskowski,R.A. (2009) PDBsum new things. *Nucleic Acids Res.*, **37**, D355–D359.