Published online 23 May 2011 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* ***W283–W287***
doi:10.1093/nar/gkr311

PredUs: a web server for predicting protein interfaces using structural neighbors

Qiangfeng Cliff Zhang[1], Lei Deng[1,2], Markus Fisher[1], Jihong Guan[2], Barry Honig[1] and Donald Petrey[1,*]

[1]Department of Biochemistry and Molecular Biophysics, Center for Computational Biology and Bioinformatics, Howard Hughes Medical Institute, Columbia University, 1130 St. Nicholas Avenue, Room 815, New York, NY 10032, USA and [2]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China

Received February 25, 2011; Revised April 1, 2011; Accepted April 18, 2011

ABSTRACT

We describe PredUs, an interactive web server for the prediction of protein–protein interfaces. Potential interfacial residues for a query protein are identified by 'mapping' contacts from known interfaces of the query protein's structural neighbors to surface residues of the query. We calculate a score for each residue to be interfacial with a support vector machine. Results can be visualized in a molecular viewer and a number of interactive features allow users to tailor a prediction to a particular hypothesis. The PredUs server is available at: http://wiki.c2b2.columbia.edu/honiglab_public/index.php/Software:PredUs.

INTRODUCTION

Prediction of the potential locations at which proteins interact with other proteins is essential to understanding their function and has been successfully exploited in many applications, including identification of an approximate binding mode in protein–protein docking, as a guide in site-directed mutagenesis and in the identification of pharmacological targets. Approaches to interface prediction typically depend on the recognition of differences in the properties of amino acids (e.g. residue hydrophobicity and sequence conservation) in surface patches that interact with other molecules, as compared to other surface residues (1–6).

'Template-based' prediction, in which an interface for a given query protein is inferred based on some similarity to another protein or set of proteins with known interfaces has been less extensively used. This is especially true of remote similarities which may be due to the lack of data about conservation of the location of binding sites in remote neighbors. Recently, we reported a comprehensive analysis of the degree to which the location of a protein interface is conserved in sets of proteins that share varying degrees of similarities (7). Our results showed that while, in general, interface conservation is most significant among close neighbors, it is still significant even for remote structural neighbors. Based on this observation, we implemented a template-based protein interface prediction method and tested it on a docking benchmark and a set of CAPRI targets. Our method offered the best combination of prediction precision and recall among all methods tested, including PINUP (8), cons-PPISP (9) and ProMate (10), which were suggested to be the top three standalone protein interface prediction programs in a recent comparative study of six interface prediction methods (4).

Here we describe PredUs, an interactive web server using this template-based protein interface prediction method. Given a query protein structure as input, we 'map' interaction sites of structural neighbors involved in a complex to residues on the surface of the query. Based on the mapped contacting frequencies, we calculate a score for residues to be interfacial. In the version of our method implemented on the server we use a support vector machine (SVM) to calculate the score, which shows superior performance compared to the original score based on logistic regression (7) on the same benchmarks.

PredUs ALGORITHMS

Given a protein structure, we first find its structural neighbors using the structural alignment program Ska (11). We use a PSD [protein structure distance, a measure of structural similarity (12)] cutoff of 0.6, which allows detection of both close and remote relationships. Structures that are involved in a PQS [Protein Quaternary Structures, (13)] or PDB [Protein Data Bank, (14)] complex are kept and

*To whom correspondence should be addressed. Tel: (212) 851 4651; Fax: (212) 851 4650; Email: dsp18@columbia.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

ranked by structural alignment score, (15), which reflects a combination of structural similarity and alignment length.

An interface from a structural neighbor is 'mapped' to the query by placing any interacting partners of the structural neighbor in the coordinate system of the query, using the transformation that relates the structural neighbor to the query. If a heavy atom of a query residue is within 5.0 Å of an interacting partner after the transformation, we increment a counter associated with this residue with the sequence identity between the query and the structural neighbor. This is repeated for each structural neighbor ordered according to its structural alignment score. To avoid over counting of highly similar interfaces, we cluster PQS/PDB chains using cd-hit (16) at 40% sequence identity cutoff. If two structural neighbors belong to a single cluster and their interacting partners also belong to a single cluster, only the structural neighbor with the higher structural alignment score will be considered. We sum the weighted contact frequencies at each residue of the query after interfaces of all structural neighbors have been mapped [see reference (7) for details].

In the current version of the PredUs server, we use a SVM to predict whether or not a surface residue is in an interface. The SVM is implemented with the package *libsvm* 3.0 (17) using radial basis function as the kernel. For each surface residue, we define a patch that includes the residue and its 14 spatially nearest surface residues. The contacting frequencies (*freq*) and solvent accessible surface areas (*ASA*) of the residues in the surface patch and the maximum contacting frequency of residues of the entire protein constitute a feature profile of length 31, i.e. $[freq_{max}, freq_0, freq_1,\ldots, freq_{14}, ASA_0, ASA_1,\ldots, ASA_{14}]$. These profiles are used as the input to the SVM and are mapped to vectors of a high-dimensional space using the kernel function. The SVM attempts to construct a hyperplane in that space that separates the vectors associated with interfacial residues from those that are non-interfacial. The interfacial score reflects the distance above (positive score) or below (negative score) this hyperplane. The higher the score the more likely a given residue is to be in an interface. By default, PredUs predicts all residues with positive score to be interfacial, but this cutoff is adjustable by the user.

PredUs FEATURES

Input to the PredUs web server can be a protein structure file in PDB format, or a PDB code. PredUs will check the validity of the input structure, and once confirmed, submit it for prediction. Users can submit multiple structures, and provide a job title or email address to facilitate retrieval of results.

As a unique feature, PredUs allows users to specify the structure of the binding partner. Once users provide another structure file or PDB code as 'Partner Structure', PredUs will predict the interface specifically used in the binding of the provided partner by only mapping the interfaces between structural neighbors of the query protein and structural neighbors of the partner.

A typical prediction takes a few minutes and almost all complete in no more than 30 min. The output consists of a list of residues and their associated score to be in an interface for each submitted structure which can be downloaded in text format. Individual predictions can be visualized in the molecular viewer AstexViewer (18) by following the 'View Structure' link. Surface residues are rendered in different colors according to their predicted interfacial score (Figure 1).

Another unique feature of PredUs is that users can tailor a prediction to a particular hypothesis following the 'Interactive prediction' link. Figure 2 shows structure-based sequence alignments between the query protein (on the top) and its structural neighbors on which the prediction is based. Below the alignment are tools that allow users to filter structural neighbors based on functional information including GO terms (19), or SCOP (20), PFAM (21) and InterPro (22) categories. It is well known that proteins can interact with different partners at distinct regions of their surfaces and these different interfaces can be associated with different functions (23). By default, however, PredUs will map all interfaces of structural neighbors of a query protein without regard to sequence or functional relationships. Hence default predictions are indications of all possible places where the query may interact with other proteins and may initially be overly broad. Restricting the set of structural neighbors via filters to include only close sequence neighbors, for example, or remote homologs that are associated with a specific function should in many cases produce a more accurate prediction.

On this page, users can also reorder the set of structural neighbors using different ranking operators shown above the alignments. Structural neighbors can be ranked based on four scores: structural alignment score, the default; PSD; RMSD (root mean square deviation, based on aligned residues) and SID (sequence identity). With the different operators, users can compare predicted interfacial residues to real interfacial residues in structural neighbors ranked by different similarity measurements.

The query protein can be further analyzed in our protein function annotation server MarkUs (24) provided by the link 'MarkUs Annotation'. Interfaces predicted by PredUs can be examined in MarkUs and comparatively studied with other functional properties like ligand binding sites, enzymatic active sites and other residue and surface features, across a wide range of sequence and structural similarities.

PredUs BENCHMARKS

We used protein docking benchmark dataset of 188 chains in training and testing PredUs. As an independent test, we also used a set of CAPRI targets that contains 56 chains in both bound and unbound forms. Please see reference (7) for a detailed description of the datasets.

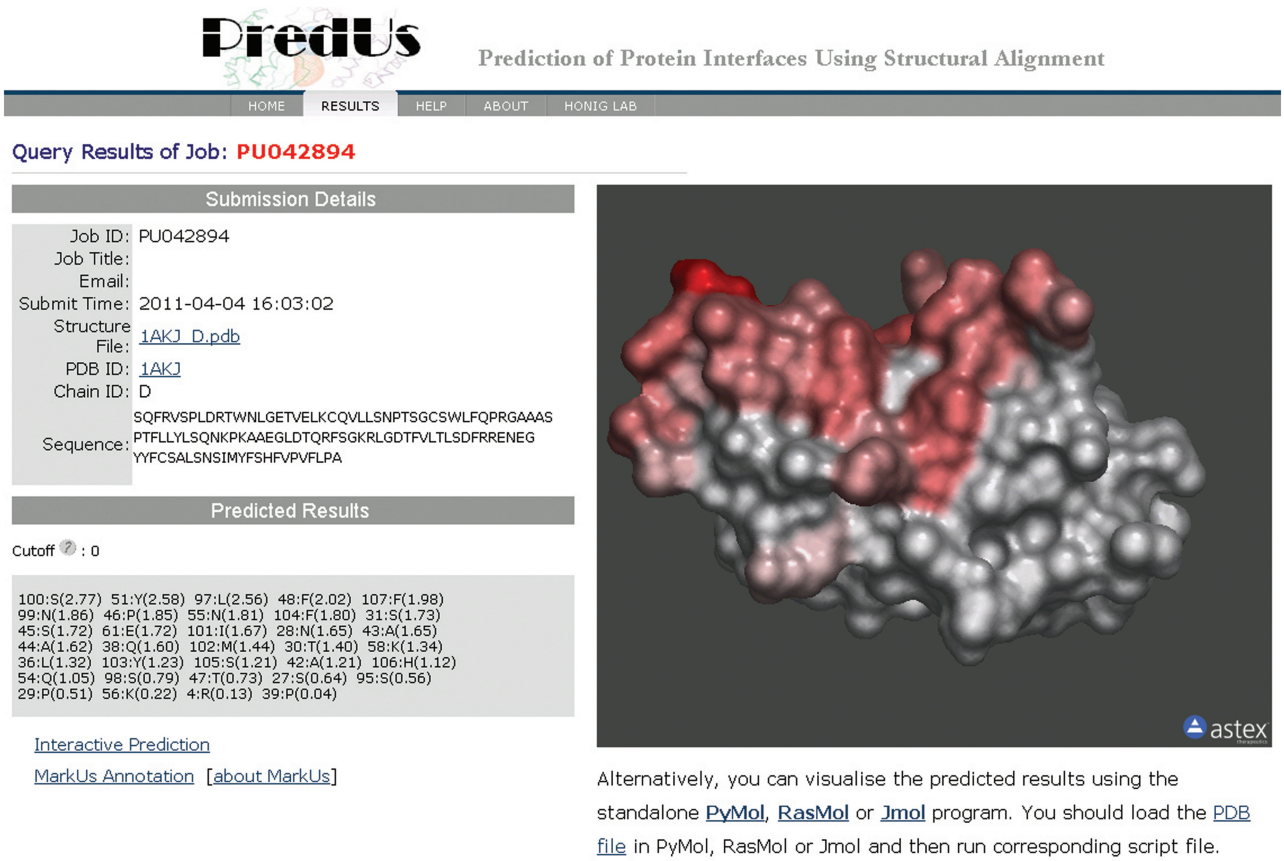

Figure 1. PredUs prediction output. The left of the figure shows the submission details and prediction results. All residues with interfacial score higher than zero are shown with scores in parentheses following residue number (in the PDB structure file) and residue name. On the right is the submitted structure with its molecular surface rendered in colors according to residue interfacial score. Residues of score higher than zero are shown from light red to red as the score increases.

To assess the predictions, we calculated a variety of quantities:

$$\text{Precision} = TP/(TP+FP)$$

$$\text{Recall} = TP/(TP+FN)$$

$$\text{Accuracy} = (TP+TN)/(TP+FP+TN+FN)$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision}+\text{Recall}}$$

Here TP, FP, TN, FN are true positive, false positive, true negative, false-negative predictions; MCC is the Matthews correlation coefficient. We also drew the receiver operating characteristic (ROC) curve and calculated the area under the curve (AUC).

We used 10-fold cross validation to test PredUs on the protein docking benchmark dataset. We tested the prediction performance of the SVM in terms of AUC value using different surface patch sizes ranging from 3 to 25 and found that the best performance was achieved with a 15-residue patch. No structural and functional filters were applied in benchmarking. All quantities except AUC were calculated using an interfacial score cutoff of zero (in principle, a score higher than zero means the residue is more likely to be in an interface). These are also default settings in the PredUs server.

As shown in Table 1, PredUs can achieve a high prediction precision and recall at the same time and achieves superior performance compared to our original study (7) as a result of the use of the SVM classifier. In the current version of PredUs, we achieve a precision and recall of 50 and 58%, compared to 44 and 46% using the original scoring scheme. Here and in the following test of CAPRI targets, we only compare with the original algorithm, which had been shown to offer the best combination of precision and recall among other methods we tested, including PINUP, cons-PPISP and ProMate (7).

The SVM classifier trained on the whole docking benchmark set was applied to the CAPRI test sets. The results are summarized in Table 1 and the performance was again improved [prediction precision and recall are 43/43% and 53/54% versus 42/40% and 42/45% in the original prediction for bound/unbound targets, respectively (7)].

DISCUSSION

PredUs predicts protein interfaces by mapping binding sites from structural neighbors. In contrast to methods

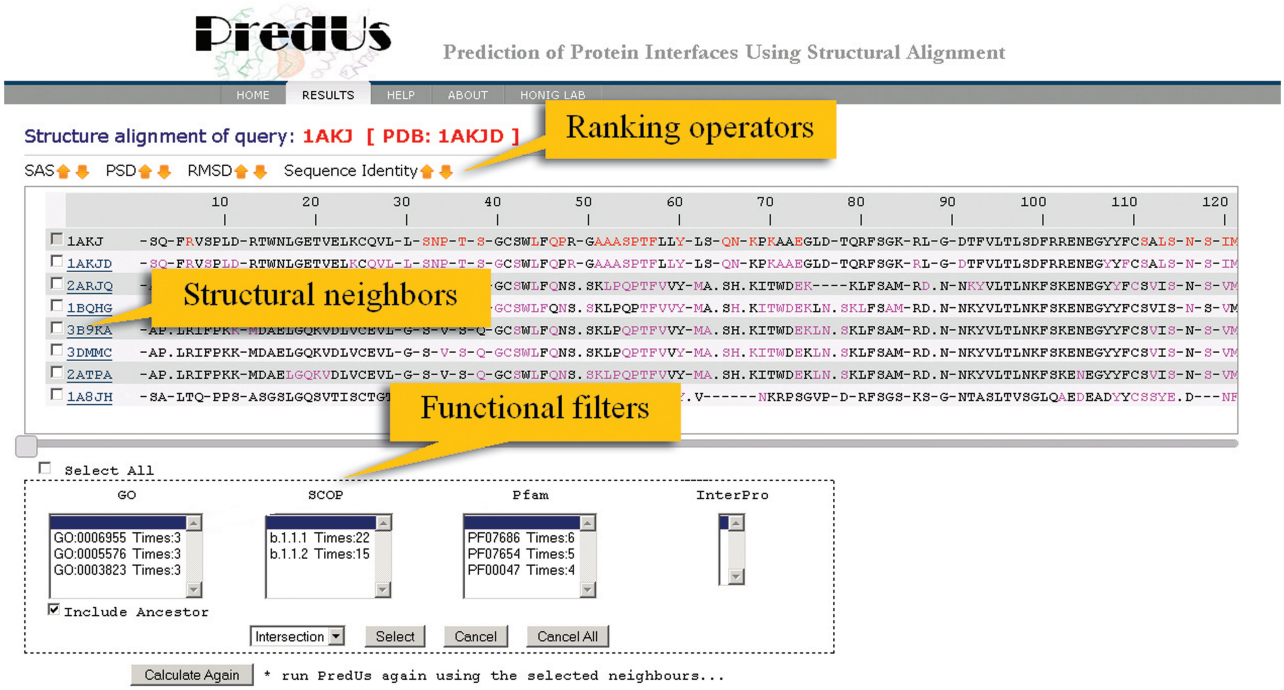

Figure 2. PredUs interactive prediction. The figure shows the structure-based sequence alignments of a query protein and its structural neighbors. Predicted interfacial residues in the query sequence are colored in red and the actual interfacial residues in the structural neighbors are indicated in purple. Functional terms populated in the set of structural neighbors are shown below the alignments. These can be used as functional filters to generate function-specific predictions by clicking the 'Calculate Again' button. Gaps are shown as dashes. For brevity, insertions of more than one residue with respect to the query are shown as dots.

Table 1. PredUs prediction performance averages on the docking benchmark dataset (DKBM3) and CAPRI bound/unbound targets

Dataset	Precision (%)	Recall (%)	Accuracy (%)	AUC	MCC	F1
10-fold cross-validation						
DKBM3	50.3	57.5	72.6	0.739	0.345	0.530
Independent test						
CAPRI bound	43.0	53.0	72.1	0.713	0.290	0.474
CAPRI unbound	43.3	53.6	73.2	0.729	0.304	0.479

Quantities in each column are defined in the description of the PredUs benchmarks in the main text.

Table 2. PredUs prediction performance averages when using structure neighbors from the same and different SCOP groupings on the docking benchmark dataset

Prediction methods	Cases	Precision average (%)	Recall average (%)
PredUs(server)	185	50.3	57.5
PredUs(original)	185	43.6	45.7
Family	141	50.8	33.8
Superfamily	147	45.9	36.2
Fold	153	41.8	38.7

Quantities in each column are defined in the description of the PredUs benchmarks in the main text.

based on residue properties, such as hydrophobicity and conservation, an advantage of this type of direct mapping is that it allows the identification of interfacial residues that are less distinctive in terms of such properties. This can be seen from the much higher recalls of the PredUs server than other protein interface prediction methods [Table 1 and reference (7)]. This type of mapping also seems to be insensitive to conformational changes that may occur upon binding, as can be seen from the small difference between the performances of PredUs on the bound and unbound CAPRI targets (Table 1).

The choice of structural neighbors is an important issue affecting the performance of template-based approaches and it might be expected that restricting the set of structural neighbors to closely related sequence homologs may produce more biologically relevant results. We have shown previously (7) that while such a limitation improves predictive accuracy it decreases the recall at the same time. As seen in Table 2, a general trend is that the number of cases for which we can make predictions and also the prediction recall improves as more remote neighbors are include with little sacrifice in precision. Consequently, the prediction strategy implemented in PredUs is to use the widest range of structural neighbors by default, since this appears to provide the best indication of the possible binding sites on a given protein. To limit the set of structural neighbors to those that a user thinks might be more biologically relevant, they can then apply the different evolutionary, structural and functional filters or specify a binding partner, as well as directly compare actual

interfacial residues in the structural neighbors to the predictions.

A limitation of PredUs is that, for every query protein, structural neighbors in a complex are required to make predictions. By exploiting remote structural homology, however, this limitation is small with only ~5% the proteins in our benchmark having no structural neighbors with binding partners, and this percentage should continue to decrease as more protein–protein complexes are characterized structurally.

PredUs has been set up for half a year and has been tested extensively. In an application of genome-wide modeling of protein–protein interactions, we have used it to predict interfaces for all proteins with structural information in the yeast and human proteomes.

FUNDING

National Institutes of Health [GM030518, GM094597 and CA121852]; National Natural Science Foundation of China [60873040]; Shuguang Scholar Program of Shanghai Education Development Foundation. Funding of open access charge: Howard Hughes Medical Institute.

Conflict of interest statement. None declared.

REFERENCES

1. Tsai,C.J., Lin,S.L., Wolfson,H.J. and Nussinov,R. (1996) Protein-protein interfaces: architectures and interactions in protein-protein interfaces and in protein cores. Their similarities and differences. *Crit. Rev. Biochem. Mol. Biol.*, **31**, 127–152.
2. Jones,S. and Thornton,J.M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol.*, **272**, 121–132.
3. Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein-protein recognition sites. *J. Mol. Biol.*, **285**, 2177–2198.
4. Zhou,H.X. and Qin,S. (2007) Interaction-site prediction for protein complexes: a critical assessment. *Bioinformatics*, **23**, 2203–2209.
5. de Vries,S.J. and Bonvin,A.M. (2008) How proteins get in touch: interface prediction in the study of biomolecular complexes. *Curr. Protein Pept. Sci.*, **9**, 394–406.
6. 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.
7. Zhang,Q.C., Petrey,D., Norel,R. and Honig,B.H. (2010) Protein interface conservation across structure space. *Proc. Natl Acad. Sci. USA*, **107**, 10896–10901.
8. Liang,S., Zhang,C., Liu,S. and Zhou,Y. (2006) Protein binding site prediction using an empirical scoring function. *Nucleic Acids Res.*, **34**, 3698–3707.
9. Chen,H.L. and Zhou,H.X. (2005) Prediction of interface residues in protein-protein complexes by a consensus neural network method: test against NMR data. *Proteins*, **61**, 21–35.
10. Neuvirth,H., Raz,R. and Schreiber,G. (2004) ProMate: a structure based prediction program to identify the location of protein-protein binding sites. *J. Mol. Biol.*, **338**, 181–199.
11. Petrey,D. and Honig,B. (2003) GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol.*, **374**, 492–509.
12. Yang,A.S. and Honig,B. (2000) An integrated approach to the analysis and modeling of protein sequences and structures. I. Protein structural alignment and a quantitative measure for protein structural distance. *J. Mol. Biol.*, **301**, 665–678.
13. Henrick,K. and Thornton,J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
14. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
15. Kolodny,R., Koehl,P. and Levitt,M. (2005) Comprehensive evaluation of protein structure alignment methods: scoring by geometric measures. *J. Mol. Biol.*, **346**, 1173–1188.
16. Li,W.Z. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
17. Chang,C.-C. and Lin,C.-J. (2001) LIBSVM, a library for support vector machines, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
18. Hartshorn,M.J. (2002) AstexViewer: a visualisation aid for structure-based drug design. *J. Comput. Aided Mol. Des.*, **16**, 871–881.
19. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
20. Lo Conte,L., Ailey,B., Hubbard,T.J.P., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res.*, **28**, 257–259.
21. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–222.
22. Apweiler,R., Attwood,T.K., Bairoch,A., Bateman,A., Birney,E., Biswas,M., Bucher,P., Cerutti,L., Corpet,F., Croning,M.D. *et al.* (2001) The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.*, **29**, 37–40.
23. Keskin,Z., Gursoy,A., Ma,B. and Nussinov,R. (2008) Principles of protein-protein interactions: what are the preferred ways for proteins to interact? *Chem. Rev.*, **108**, 1225–1244.
24. Petrey,D., Fischer,M. and Honig,B. (2009) Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc. Natl Acad. Sci. USA*, **106**, 17377–17382.