W308–W313 *Nucleic Acids Research, 2013, Vol. 41, Web Server issue* *Published online 3 June 2013*
doi:10.1093/nar/gkt457

webPDBinder: a server for the identification of ligand binding sites on protein structures

Valerio Bianchi, Iolanda Mangone, Fabrizio Ferrè, Manuela Helmer-Citterich* and Gabriele Ausiello

Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Via della Ricerca Scientifica snc, 00133 Rome, Italy

Received February 18, 2013; Revised April 30, 2013; Accepted May 4, 2013

ABSTRACT

The webPDBinder (http://pdbinder.bio.uniroma2.it/PDBinder) is a web server for the identification of small ligand-binding sites in a protein structure. webPDBinder searches a protein structure against a library of known binding sites and a collection of control non-binding pockets. The number of similarities identified with the residues in the two sets is then used to derive a propensity value for each residue of the query protein associated to the likelihood that the residue is part of a ligand binding site. The predicted binding residues can be further refined using conservation scores derived from the multiple alignment of the PFAM protein family. webPDBinder correctly identifies residues belonging to the binding site in 77% of the cases and is able to identify binding pockets starting from holo or apo structures with comparable performances. This is important for all the real world cases where the query protein has been crystallized without a ligand and is also difficult to obtain clear similarities with bound pockets from holo pocket libraries. The input is either a PDB code or a user-submitted structure. The output is a list of predicted binding pocket residues with propensity and conservation values both in text and graphical format.

INTRODUCTION

The identification of ligand-binding sites is a crucial part in functional annotation of proteins, which benefits enormously from the knowledge of the protein 3D structure. It can provide clues about the molecular function of a protein, even in cases in which the relationship between molecular and biological function is not clear and where there is scarce sequence similarity between the protein of interest and available annotated proteins.

Indeed, a large fraction of the protein structures deposited in Protein Data Bank (PDB) (1) remains with unknown function (2). This happens because methods that are commonly used to transfer functional annotations from homologous proteins, i.e. BLAST (3) and Dali (4), are not able to capture from the sequence all the information needed to infer the function, especially when the global sequence similarity falls in the twilight zone (below 25% sequence identity) (5).

As the protein global fold is more conserved than its sequence in protein families (6), structure-based methods outperform sequence-based methods in functional annotations (7) when they operate in the twilight zone.

The increase in size of solved protein structures with poorly characterized biochemical functions or molecular interactions has led to the need for computational methods able to detect and characterize functional sites on protein structures (8). Functional site detection is also important for targeting specific pockets in structure-based and fragment-based drug design (9,10), drug discovery (11) and molecular docking (12,13).

The ligand-binding site detection procedure is generally divided in two steps: the identification of the location of an appropriate cavity on the protein surface and the prediction of a suitable ligand that could fit into it. The latter step is a challenging task and generally requires extremely high computational resources. Accordingly, a variety of algorithms have been developed to identify ligand-binding pockets in protein structures to limit the search space in molecular docking pipelines. The available methods use geometric criteria (14–17), energy functions (18–21) or other types of characteristics, such as surface accessibility, the net charge on the protein residues in a protein as a function of pH, sequence conservation and so forth (22–27).

*To whom correspondence should be addressed. Tel: +39 06 7259 4324; Fax: +39 06 2023500; Email: citterich@uniroma2.it
Present address:
Valerio Bianchi, Center for Genomic Science of IIT@SEMM, Istituto Italiano di Tecnologia, via Adamello 16, 20139 Milan, Italy.

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

Obviously, when the structure of a homologue of the protein of interest is available, the identification of binding sites can take advantage from the transferred structural information. Different methods have been published such as 3DLigandSite (28), FINDSITE (29) and Firestar (30). They achieve better performances in identifying the correct location of binding sites by superposing homologue(s) of the protein of interest, whenever available, onto the query protein structure, to determine the location of the ligand binding site(s) and/or the residues involved in binding (31). However, the major limitation of these methods is that they cannot be used when no annotated homologues is available.

The performance of binding site prediction methods differs according to whether the analysis is performed on apo (ligand-unbound) or holo (ligand-bound) structures because proteins often undergo conformational changes on ligand binding. In general, most of these methods correctly identify the location of the binding site in 70–95% of the cases if the protein analyzed is in the bound conformation. In contrast, the same analysis performed on the apo structures achieves a success rate ranging from 50 to 75% (32,33).

Although knowledge-based approaches have been developed that use these data to dock ligands onto proteins (34–36), no method exists that uses available protein-ligand complexes to predict binding sites on a query structure, irrespective of the nature of the ligand. For this reason, we have developed the PDBinder algorithm (37), a knowledge-based method based on the observation that unrelated binding sites often share small structural motifs that bind the same chemical fragments, irrespective of the type of ligand they are able to bind. The method correctly identifies residues belonging to the binding site in 77% of the cases. In particular, we have proved it to be the best available method on both holo and apo protein structures, obtaining for the latter performances similar to those for proteins in their bound conformation. These results are extremely important for all the real world cases where the query protein has been crystallized without a ligand and is also difficult to obtain clear similarities with bound pockets from holo pocket libraries. In this web server version, the method has been improved by the addition of a novel conservation scoring function with an improvement of 3% in positive predictive value and 13% in sensitivity both in the apo and holo test set. The interface of webPDBinder is designed to facilitate the input process and to obtain a clear and user-friendly graphical output that makes the method accessible to a broad audience.

METHODS AND EXPERIMENTAL RESULTS

The webPDBinder is based on the PDBinder method for the identification of protein structure residues in contact with a putative ligand. PDBinder searches for small (three residues) similarities between a query structure and two libraries of binding and non-binding residues using Superpose3D (38), a fast local structure comparison method.

The binding and non-binding residue datasets are derived from a subset of the PDB protein structures obtained from the BLASTClust (39) sequence clusters at 30% sequence identity. Binding pockets were defined by selecting all the residues having an atom closer than 3.5 Å to any atom of a ligand. This distance threshold was determined during the training of the method. In total, the ligand-binding data set was composed of 1896 binding pockets comprising 25 905 residues and the non-binding one of 423 556 residues not interacting with any ligand.

When the residues are compared with those of the query structure, they are considered similar by Superpose 3D if they can be superimposed with a Root Mean Square Deviation (RMSD) lower than a given threshold, using a two point (c-alpha and side chain centroid) amino acid representation. The ratio of similarities identified in the two sets (binding and not-binding) is then used to derive a propensity value for each residue belonging to the protein of interest. Surface areas with high propensity values denote the position of the predicted binding site.

The PDBinder method has been tested on a data set of 239 holo and apo protein structures. The method achieved an average sensitivity of 30%, an average specificity of 98% and a precision of 41% on holo protein structures. Using the apo test set, the method achieved average sensitivity of 25%, specificity of 98% and precision of 37%. The ability of the classifier in correctly identifying binding residues from non-binding ones is 77% in both holo and apo protein structures.

An improved version of the algorithm that takes in account the sequence conservation is optionally available on the web server for all those cases where a PFAM domain (40) can be associated to the query structure. A residue conservation score is derived from the available PFAM multiple alignments in two steps. (i) The percentage of similar residues (BLOSUM62 scores $\geq$1) in each alignment columns is computed. (ii) The score is normalized across differently overall conserved PFAM families, by using for each residue the percentile score of its conservation versus the distribution of conservation scores of the whole protein.

This modified version of the algorithm has been trained on the same training data set of the PDBinder method (1356 high-quality non-redundant protein structures). We were able to assign a conservation score to 1237 of these proteins, and we identified a combination of conservation threshold (58) and propensity value threshold (0.125) that obtained the best performance. The application of these new thresholds on the data set produced a 5% increase on the Positive Predictive Value of the method while leaving all the other values almost unchanged (data not shown).

After the identification of the best combination of conservation and propensity value thresholds, we tested the new version of PDBinder on the LigASite (41) test set used by the older version of PDBinder (239 holo and apo protein structures). The results (Table 1) show that the combination of conservation and propensity value increases the performance of PDBinder both on bound

Table 1. Results for PDBinder and the modified version in webPDBinder (PDBinder + conservation score) on the original test set of 239 holo and apo protein structures (LigASite) in terms of Sensitivity (SENS), Specificity (SPEC), Positive predictive value (PPV) and Matthew's Correlation Coefficient (MCC)

Method	SENS	SPEC	PPV	MCC	Dataset
PDBinder	0.295	0.983	0.413	0.313	HOLO
PDBinder + cons	0.430	0.968	0.433	0.384	
PDBinder	0.251	0.984	0.372	0.271	APO
PDBinder + cons	0.378	0.969	0.400	0.342	

Table 2. PDBinder results after the removal of protein structures at different thresholds of sequence identity in their binding pockets

Sequence identity threshold (%)	SENS	SPEC	PPV	MCC
5	0.271	0.977	0.441	0.300
10	0.269	0.979	0.450	0.304
15	0.267	0.980	0.456	0.305
20	0.266	0.981	0.461	0.306

The sequence identity threshold refers to the maximum percentage of sequence identity between the query protein-binding pocket and the binding pockets of each protein of the binding and non-binding residues data set. The results report performances in terms of Sensitivity (SENS), Specificity (SPEC), Positive Predictive Value (PPV) and Matthew's Correlation Coefficient (MCC).

and unbound protein structures. We obtained a slight increase both in Positive Predictive Value and Specificity, an increase of 13.5 and 12.7% in Sensitivity, 0.071 and 0.071 in Matthew's Correlation Coefficient, respectively, on holo and apo protein structures.

We tested the method in conditions that mimic the prediction for protein structures without known homologues in the PDB. We performed a leave-one-out experiment using protein structures from the data set used in the original PDBinder manuscript, comprising 1356 high-quality protein structures. For each protein analyzed, we used for comparison only residues from the binding and non-binding data sets that belong to protein structures sharing less than a fixed threshold of sequence identity with the query protein binding pocket. Table 2 reports the performances of PDBinder for different thresholds of binding pockets sequence identity, in terms of Sensitivity, Specificity, Positive Predictive Value and Matthew's Correlation Coefficient. The results show that also in the worst scenario in which we are able to use only structures that share <5% sequence identity with the query protein in the binding pocket, the method identifies ~27% of the binding site residues, without loosing in precision (Positive Predictive Value).

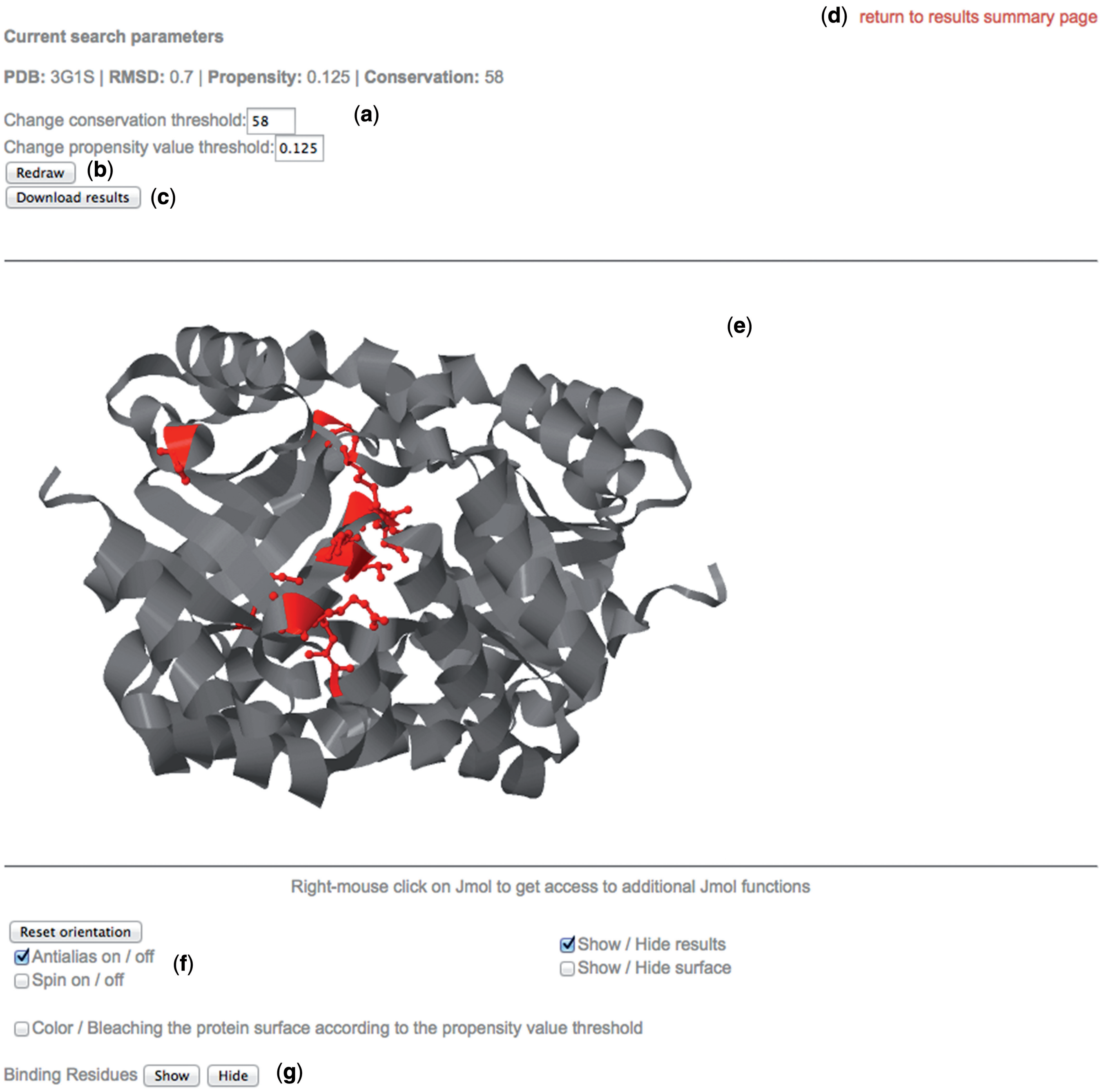
Input and output

To submit a job to webPDBinder, the user can: (i) input a PDB ID (or a list of PDB IDs) in the textbox of the 'Run your job' section or (ii) upload a PDB formatted file in the 'Upload your file' section. In both modes by using the advanced parameters, the user can choose to run a more or less stringent search. The user can change the RMSD threshold used in the structural comparison phase as well as the Propensity value and Conservation score used in the identification of the binding residues. The RMSD thresholds can be set to 0.5, 0.6 and 0.7, the Propensity threshold can be set anywhere in the range from 0 to 1 and the conservation threshold can range from 0 to 100. Default parameters as specified in 'Methods and Experimental Results' section are 0.7 for RMSD, 0.125 for the Propensity value and 58 for Conservation score. To use the old version of PDBinder (without the sequence conservation) the propensity and the conservation thresholds must be fixed to 0.143 and 0, respectively. The user can also provide an email address to which links to the results will be sent, although providing an email address is not mandatory. An auto-refresh intermediate page will report when every submitted job is finished.

The 'Results page' (Figure 1) contains a summary of the input data, reporting the query protein and parameters used in the analysis (a). Buttons are available that can be used to download the results in a parsable text file (c), or go back to the Summary page (d). Using the 'Redraw' button (b), the user can change the propensity value and the conservation score thresholds and view the new results without the need to re-run the whole search. The analyzed structure is shown using the Jmol Java applet, an open-source Java-based viewer for 3D chemical structures (http://www.jmol.org/). By default, the protein structure is represented as a gray ribbon, whereas predicted binding residues are represented in balls and sticks mode and colored in red (e). Under the Jmol applet interface, some shortcuts options are available to modify the default visualization (f) together with a button for the visualization of a table with the predicted binding residues (g). This table (Figure 2) shows the residues identified in a binding site and for each prediction information is reported about the residue name (a), number (a) and chain (c), the Propensity Value achieved (d) and, if available, the residue Conservation Score (e) in its PFAM family. The user can also highlight binding residues on the Jmol structure by checking the associated residue's radio button (f). A Usage guide is provided to the user, which describes every step graphically and by means of simple instructions. The web site is freely available at http://pdbinder.bio.uniroma2.it/PDBinder and does not require registration.

CONCLUSIONS

PDBinder is a web server for the prediction of ligand-binding pockets on protein structures. The identification of ligand-binding sites is a difficult task when there is scarce sequence similarity between the protein of interest and available annotated proteins or when the similar structures are only crystallized in their apo form. The analysis of apo/holo structure pairs shows that the



Current search parameters

PDB: 3G1S | RMSD: 0.7 | Propensity: 0.125 | Conservation: 58

Change conservation threshold: 58

Change propensity value threshold: 0.125

Right-mouse click on Jmol to get access to additional Jmol functions

☑ Antialias on / off

□ Spin on / off

☑ Show / Hide results

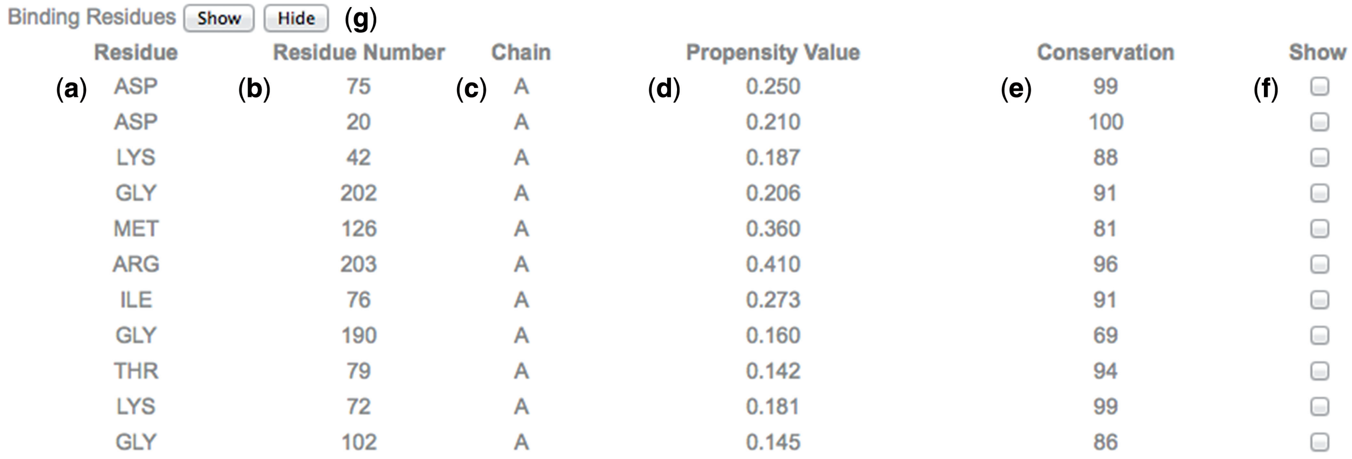
□ Show / Hide surface

□ Color / Bleaching the protein surface according to the propensity value threshold

Figure 1. Example of a webPDBinder results page. The figure shows the prediction made on the Orotidine 5′-monophosphate decarboxylase from Methanobacterium thermoautotrophicum (PDB code 3G1S) with the default parameters (an RMSD threshold of 0.7 Å, a Propensity value of 0.125 and a conservation score of 58%). In the upper part, a summary of the parameters used in the search is reported (**a**), together with a button to download a parsable result file (**c**), a button to go back to the Summary page (**d**) and a button to re-submit a new job after changing parameters (**b**). In the Java Applet, the predicted residues are colored in red and displayed as ball and sticks, whereas the query protein is showed in ribbon style and colored in gray (**e**). In the bottom part, buttons are available to change the Jmol visualization options (**f**) and to view a list of the predicted residues (**g**).

performance of PDBinder is almost similar in both cases. This could be explained by the fact that, even though the ligand induces some rearrangements in the overall structure of the binding site, the local conformation of small sets of residues, which is the level of detail relevant for PDBinder, does not vary much upon binding. Therefore, PDBinder can be applied to all those structures of unknown function that lack homologue(s) and have been crystallized without the ligand.

This web server provides a user-friendly version of the PDBinder method, enriched with a new parameter, the conservation value and improved performance. It gives an interactive and easy to use interface to visualize the predicted binding sites on the query structure



Residue	Residue Number	Chain	Propensity Value	Conservation	Show
ASP	75	A	0.250	99	□
ASP	20	A	0.210	100	□
LYS	42	A	0.187	88	□
GLY	202	A	0.206	91	□
MET	126	A	0.360	81	□
ARG	203	A	0.410	96	□
ILE	76	A	0.273	91	□
GLY	190	A	0.160	69	□
THR	79	A	0.142	94	□
LYS	72	A	0.181	99	□
GLY	102	A	0.145	86	□

Figure 2. The list of the binding site predicted residues shown in the result page. For each prediction, information is reported about the residue name (**a**), number (**b**) and chain (**c**), the Propensity Value achieved (**d**) and, if available, the residue Conservation Score (**e**) in its PFAM family. Users can highlight binding residues on the Jmol structure by checking the relative residue's radio button (**f**).

directly on the web, without the need to install locally the program.

FUNDING

AIRC [IG 10298 to M.H.C.]; FIRB 'Futuro in ricerca' [RBFR08ZSXY to G.A.]. Funding for open access charge: FIRB 'Futuro in ricerca' [RBFR08ZSXY].

Conflict of interest statement. None declared.

REFERENCES

1. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D Biol. Crystallog.*, **58**, 899–907.
2. Nadzirin,N. and Firdaus-Raih,M. (2012) Proteins of Unknown function in the Protein Data Bank (PDB): an inventory of true uncharacterized proteins and computational tools for their analysis. *Int. J. Mol. Sci.*, **13**, 12761–12772.
3. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
4. Holm,L. and Sander,C. (1995) Dali: a network tool for protein structure comparison. *Trends Biochem. Sci.*, **20**, 478–480.
5. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng.*, **12**, 85–94.
6. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
7. Kristensen,D.M., Ward,R.M., Lisewski,A.M., Erdin,S., Chen,B.Y., Fofanov,V.Y., Kimmel,M., Kavraki,L.E. and Lichtarge,O. (2008) Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. *BMC Bioinformatics*, **9**, 17.
8. Gherardini,P.F. and Helmer-Citterich,M. (2008) Structure-based function prediction: approaches and applications. *Brief. Funct. Genomic. Proteomic.*, **7**, 291–302.
9. Mortier,J., Rakers,C., Frederick,R. and Wolber,G. (2012) Computational tools for in silico fragment-based drug design. *Curr. Top. Med. Chem.*, **12**, 1935–1943.
10. Sgrignani,J. and Magistrato,A. (2012) First-Principles Modeling of Biological Systems and Structure-Based Drug-Design. *Curr. Comput. Aided Drug Des.*, **9**, 15–34.
11. Ou-Yang,S.-S., Lu,J.-Y., Kong,X.-Q., Liang,Z.-J., Luo,C. and Jiang,H. (2012) Computational drug discovery. *Acta Pharmacol. Sin.*, **33**, 1131–1140.
12. Ghersi,D. and Sanchez,R. (2009) Improving accuracy and efficiency of blind protein-ligand docking by focusing on predicted binding sites. *Proteins*, **74**, 417–424.
13. Hermann,J.C., Marti-Arbona,R., Fedorov,A.A., Fedorov,E., Almo,S.C., Shoichet,B.K. and Raushel,F.M. (2007) Structure-based activity prediction for an enzyme of unknown function. *Nature*, **448**, 775–779.
14. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.*, **13**, 323–330, 307–308.
15. Laskowski,R.A., Luscombe,N.M., Swindells,M.B. and Thornton,J.M. (1996) Protein clefts in molecular recognition and function. *Protein Sci.*, **5**, 2438–2452.
16. Capra,J.A., Laskowski,R.A., Thornton,J.M., Singh,M. and Funkhouser,T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLoS Comput. Biol.*, **5**, e1000585.
17. Tan,K.P., Varadarajan,R. and Madhusudhan,M.S. (2011) DEPTH: a web server to compute depth and predict small-molecule binding cavities in proteins. *Nucleic Acids Res.*, **39**, W242–W248.
18. Hernandez,M., Ghersi,D. and Sanchez,R. (2009) SITEHOUND-web: a server for ligand binding site identification in protein structures. *Nucleic Acids Res.*, **37**, W413–W416.
19. Morita,M., Nakamura,S. and Shimizu,K. (2008) Highly accurate method for ligand-binding site prediction in unbound state (apo) protein structures. *Proteins*, **73**, 468–479.
20. Laurie,A.T.R. and Jackson,R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, **21**, 1908–1916.
21. Ngan,C.H., Hall,D.R., Zerbe,B., Grove,L.E., Kozakov,D. and Vajda,S. (2012) FTSite: high accuracy detection of ligand binding sites on unbound protein structures. *Bioinformatics*, **28**, 286–287.
22. Amitai,G., Shemesh,A., Sitbon,E., Shklar,M., Netanely,D., Venger,I. and Pietrokovski,S. (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, **344**, 1135–1146.
23. Ondrechen,M.J., Clifton,J.G. and Ringe,D. (2001) THEMATICS: a simple computational predictor of enzyme function from structure. *Proc. Natl Acad. Sci. USA*, **98**, 12473–12478.
24. Ota,M., Kinoshita,K. and Nishikawa,K. (2003) Prediction of catalytic residues in enzymes based on known tertiary structure, stability profile, and sequence conservation. *J. Mol. Biol.*, **327**, 1053–1064.

25. Mehio,W., Kemp,G.J.L., Taylor,P. and Walkinshaw,M.D. (2010) Identification of protein binding surfaces using surface triplet propensities. *Bioinformatics*, **26**, 2549–2555.
26. Elcock,A.H. (2001) Prediction of functionally important residues based solely on the computed energetics of protein structure. *J. Mol. Biol.*, **312**, 885–896.
27. Neuvirth,H., Heinemann,U., Birnbaum,D., Tishby,N. and Schreiber,G. (2007) ProMateus–an open research approach to protein-binding sites analysis. *Nucleic Acids Res.*, **35**, W543–W548.
28. Wass,M.N., Kelley,L.A. and Sternberg,M.J.E. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.
29. Brylinski,M. and Skolnick,J. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc. Natl Acad. Sci. USA*, **105**, 129–134.
30. López,G., Valencia,A. and Tress,M.L. (2007) firestar–prediction of functionally important residues using structural templates and alignment reliability. *Nucleic Acids Res.*, **35**, W573–W577.
31. Roche,D.B., Buenavista,M.T. and McGuffin,L.J. (2012) FunFOLDQA: a quality assessment tool for protein-ligand binding site residue predictions. *PLoS One*, **7**, e38219.
32. Zhang,Z., Li,Y., Lin,B., Schroeder,M. and Huang,B. (2011) Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, **27**, 2083–2088.
33. Huang,B. (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS*, **13**, 325–330.
34. Kasahara,K., Kinoshita,K. and Takagi,T. (2010) Ligand-binding site prediction of proteins based on known fragment-fragment interactions. *Bioinformatics*, **26**, 1493–1499.
35. Ramensky,V., Sobol,A., Zaitseva,N., Rubinov,A. and Zosimov,V. (2007) A novel approach to local similarity of protein binding sites substantially improves computational drug design results. *Proteins*, **69**, 349–357.
36. Verdonk,M.L., Cole,J.C. and Taylor,R. (1999) SuperStar: a knowledge-based approach for identifying interaction sites in proteins. *J. Mol. Biol.*, **289**, 1093–1108.
37. Bianchi,V., Gherardini,P.F., Helmer-Citterich,M. and Ausiello,G. (2012) Identification of binding pockets in protein structures using a knowledge-based potential derived from local structural similarities. *BMC Bioinformatics*, **13(Suppl. 4)**, S17.
38. Gherardini,P.F., Ausiello,G. and Helmer-Citterich,M. (2010) Superpose3D: a local structural comparison program that allows for user-defined structure representations. *PloS ONE*, **5**, e11988.
39. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **41**, D8–D20.
40. Punta,M., Coggill,P.C., Eberhardt,R.Y., Mistry,J., Tate,J., Boursnell,C., Pang,N., Forslund,K., Ceric,G., Clements,J. *et al.* (2012) The Pfam protein families database. *Nucleic Acids Res.*, **40**, D290–D301.
41. Dessailly,B.H., Lensink,M.F., Orengo,C.A. and Wodak,S.J. (2008) LigASite–a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res.*, **36**, D667–D673.