W278–W282 *Nucleic Acids Research, 2011, Vol. 39, Web Server issue* *Published online 26 May 2011*
doi:10.1093/nar/gkr389

Phosfinder: a web server for the identification of phosphate-binding sites on protein structures

Luca Parca, Iolanda Mangone, Pier Federico Gherardini, Gabriele Ausiello* and Manuela Helmer-Citterich

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

Received February 11, 2011; Revised April 18, 2011; Accepted May 3, 2011

ABSTRACT

Phosfinder is a web server for the identification of phosphate binding sites in protein structures. Phosfinder uses a structural comparison algorithm to scan a query structure against a set of known 3D phosphate binding motifs. Whenever a structural similarity between the query protein and a phosphate binding motif is detected, the phosphate bound by the known motif is added to the protein structure thus representing a putative phosphate binding site. Predicted binding sites are then evaluated according to (i) their position with respect to the query protein solvent-excluded surface and (ii) the conservation of the binding residues in the protein family. The server accepts as input either the PDB code of the protein to be analyzed or a user-submitted structure in PDB format. All the search parameters are user modifiable. Phosfinder outputs a list of predicted binding sites with detailed information about their structural similarity with known phosphate binding motifs, and the conservation of the residues involved. A graphical applet allows the user to visualize the predicted binding sites on the query protein structure. The results on a set of 52 apo/holo structure pairs show that the performance of our method is largely unaffected by ligand-induced conformational changes. Phosfinder is available at http://phosfinder.bio.uniroma2.it.

INTRODUCTION

Several key reactions in a cell involve proteins interacting with the phosphate moiety, either as an isolated phosphate ion or as part of a phosphorylated ligand. The phosphate group has been observed to interact with more than half of the known proteins (1). Moreover many phosphate binding proteins are involved in pathways whose malfunction causes important human diseases (2,3). The binding of the phosphate group usually gives a significant contribution to the overall binding energy in the interaction between proteins and phosphate-containing ligands (4). The ability to bind the phosphate group evolved multiple times, as evidenced by its occurrence in several non-homologous protein families. However some recognition motifs such as the P-loop (5) and the Rossmann-type fold (6) are extremely frequent. Several methods for the prediction of ligand binding sites are available as web servers. Tools like 3DLigandSite (7), ProBiS (8,9) and SITE HOUND-web (10) use information derived from protein structures to predict binding sites irrespective of the interacting ligand.

Other web-based methods are focused on the prediction of binding sites for specific classes of ligands. For instance MetalDetector (11), predicts metal binding sites using only the sequence of the protein. Similarly ProteDNA (12) is a DNA binding site predictor based on the analysis of the sequence of known transcription factors that also takes into account the alignment of the predicted secondary structure elements.

There is also a small number of web servers devoted to the prediction of binding sites for specific ligands using structural information. RNABindR (13) predicts RNA binding sites using a Naive Bayes classifier trained on solved RNA–protein complexes; PEPSITE (14) predicts peptide binding sites using spatial position-specific scoring matrices that describe the preferred protein environment of each amino acid in the peptide.

Given the importance of the phosphate group in several biological processes (see above), we developed Pfinder (15), the only available method for the prediction of phosphate binding sites in protein structures. This method is based on the observation that the same phosphate binding structural motifs occur in evolutionarily unrelated

*To whom correspondence should be addressed. Tel: +39 06 72594324; Fax: +39 06 2023500; Email: gabriele.ausiello@uniroma2.it

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

proteins, irrespective of the identity of the ligand as a whole (16,17). Our approach therefore consists in using a previously constructed dataset of phosphate binding motifs (16) to scan a structure of interest with the Superpose3D (18) structural comparison algorithm. The residues in the query structure that match one of these motifs are predicted as phosphate binding. Moreover the phosphate group is placed on the query protein according to its position in the template motif. The predictions are then filtered to exclude those found in the interior of the protein. Residues which are not conserved in the family of the query protein are also discarded. In the present work, we describe Phosfinder (http://phosfinder.bio.uniroma2.it), a web server interface for the Pfinder method that makes it accessible to a broader audience.

METHODS

The Phosfinder server is based on the Pfinder method (15) for the prediction of phosphate binding sites in protein structures. Pfinder uses the Superpose3D structural comparison software (18,19) to scan a structure of interest against a data set of 215 phosphate binding motifs identified in a previous work (16). Each one of these motifs is composed of at least three amino acids binding a phosphate group, either in its ion form or as part of a bigger ligand, and it is present in at least two different SCOP folds (20). The structural comparison is governed by two parameters: the root mean square deviation (RMSD, a measure of geometric similarity) between corresponding residues (the Cα and side-chain geometric centroid are considered for this calculation), and the BLOSUM62 (21) substitution value of paired residues.

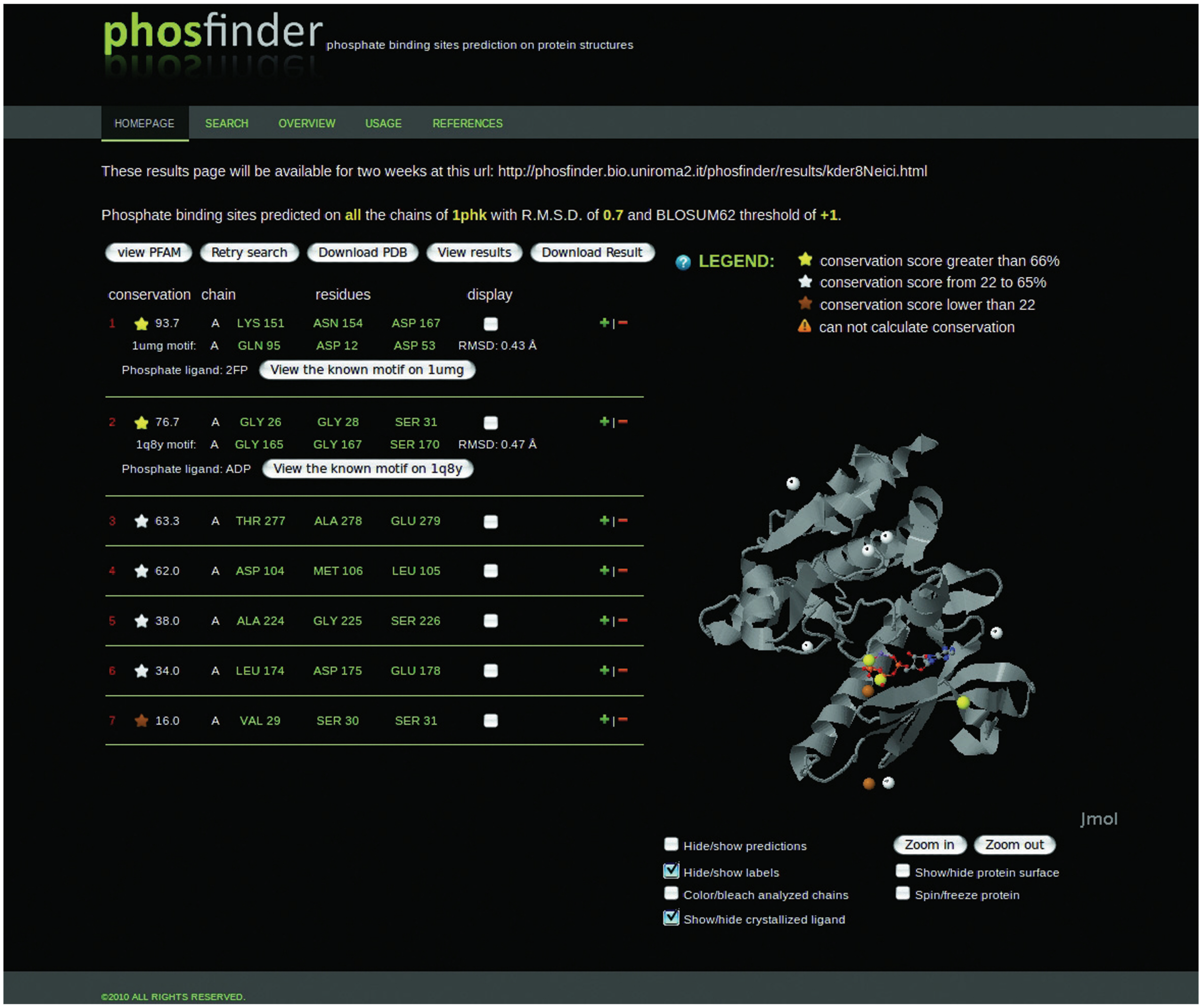
Whenever a known phosphate binding motif matches with residues of the query protein, its bound phosphate group is roto-translated onto the query structure. The phosphate group represents the predicted phosphate position and the protein residues in the match represent the corresponding inferred phosphate binding site. Predictions that result in the phosphate group being placed inside the protein solvent-excluded surface are discarded since they are unlikely to represent real binding sites. The remaining sites are then clustered with a hierarchical-clustering (centroid-linkage) procedure. For each cluster, the predicted phosphate binding position that is closer to the cluster centroid is retained as representative of the cluster. Finally a conservation score is assigned to each predicted site using the available PFAM (22) multiple alignments. For each protein residue, the percentage of similar (BLOSUM62 score ≥ 1) residues in the corresponding multiple alignment column is calculated. In order to normalize and compare values from different multiple alignments, the percentile corresponding to each value with respect to the distributions of values in the alignment is calculated. This percentile corresponds to the residue conservation score. The conservation score of the predicted phosphate binding site is calculated as the average of the conservation scores of its constituent residues. The core programs of Phosfinder are written in Python, C and C^{++} and are linked to the web interface using CGI.

Usage of Phosfinder

Phosfinder takes as input a protein structure and optionally a chain identifier. The structure can be provided as a PDB code (23), or as a user-submitted PDB format file. Using the advanced search, the user can specify different parameters: (i) the RMSD threshold of the structural comparison (ranging from 0.6Å to 0.9Å); (ii) the BLOSUM62 substitution threshold (that can be set to −1, 0 or +1, making the search less or more stringent); and (iii) the conservation value (ranging from 0 to 100, the default value is set to 66) used to rank the predictions. The user can also upload a phosphate binding motif that is not comprised in the selected data set. The server outputs a list of predicted phosphate binding sites ranked and colored according to the calculated conservation score (Figure 1). Detailed information about the structural matches between the query protein residues and the known phosphate binding motifs are also displayed; these include the number and position of the residues in the protein, the RMSD of the structural match, the conservation score and a button that allows the user to inspect the known phosphate binding motif, involved in the structural match, in its original protein. The phosphate binding sites are also displayed in a Jmol (an open-source Java-based viewer for 3D chemical structures available at http://www.jmol.org/) graphical applet: the query protein structure is shown, in ribbon style, together with the predicted sites, represented as spheres and colored according to the conservation score (Figure 2). The user can highlight the analyzed chains, show/hide the predictions, the prediction labels, the protein surface and any ligand bound by the protein. A group of buttons allows the user to view in detail the PFAM alignment of the protein and to retrieve the results as a parsable text file. A PDB-formatted file can be downloaded, containing the query protein structure together with the coordinates of the predicted phosphate binding sites. The website also includes help pages that guide the user with worked examples (a complete and interactive output page is given as example). The Phosfinder web site is freely available at http://phosfinder.bio.uniroma2.it and does not require any registration.

Experimental results

Phosfinder was trained on a set composed of 59 high-quality, non-redundant (30% sequence identity level), structures of proteins binding nucleotides and other non-nucleotide phosphorylated ligands in a 2:1 proportion (as in the whole Protein Data Bank). Nucleotide binding proteins are taken from the work of Zhao *et al.* (24). Non-nucleotide phosphorylated ligands are randomly chosen from a set of 1273 phosphate-containing ligands occurring in <10 PDB structures in the whole Protein Data Bank. For each of these ligands, a protein structure is chosen at random from those that bind it. Identical chains (e.g. dimers) are grouped leaving only one as representative. The training phase allowed us to find


phosfinder
phosphate binding sites prediction on protein structures
HOMEPAGE
SEARCH
OVERVIEW
USAGE
REFERENCES
view PFAM
Retry search
Download PDB
View results
Download Result
View the known motif on 1umg
View the known motif on 1q8y
Jmol
Zoom in
Zoom out
©2010 ALL RIGHTS RESERVED.

These results page will be available for two weeks at this url: http://phosfinder.bio.uniroma2.it/phosfinder/results/kder8Neici.html

Phosphate binding sites predicted on **all** the chains of **1phk** with R.M.S.D. of **0.7** and BLOSUM62 threshold of **+1**.

	conservation	chain	residues			display	
1	★ 93.7	A	LYS 151	ASN 154	ASP 167	□	+|−
	1umg motif:	A	GLN 95	ASP 12	ASP 53	RMSD: 0.43 Å	
	Phosphate ligand: 2FP						
2	★ 76.7	A	GLY 26	GLY 28	SER 31	□	+|−
	1q8y motif:	A	GLY 165	GLY 167	SER 170	RMSD: 0.47 Å	
	Phosphate ligand: ADP						
3	★ 63.3	A	THR 277	ALA 278	GLU 279	□	+|−
4	★ 62.0	A	ASP 104	MET 106	LEU 105	□	+|−
5	★ 38.0	A	ALA 224	GLY 225	SER 226	□	+|−
6	★ 34.0	A	LEU 174	ASP 175	GLU 178	□	+|−
7	★ 16.0	A	VAL 29	SER 30	SER 31	□	+|−

LEGEND:
- ★ conservation score greater than 66%
- ★ conservation score from 22 to 65%
- ★ conservation score lower than 22
- ⚠ can not calculate conservation

□ Hide/show predictions
☑ Hide/show labels
□ Color/bleach analyzed chains
☑ Show/hide crystallized ligand
□ Show/hide protein surface
□ Spin/freeze protein

Figure 1. An example Phosfinder output page. Seven phosphate binding sites have been predicted on the *Oryctolagus cuniculus* phosphorylase kinase (PDB code 1phk) with 0.7Å and +1 as the RMSD and BLOSUM62 thresholds, respectively. The extensible table, on the left, reports all the predicted phosphate binding sites in detail, ranked according to their conservation score. A color scheme (top-right) helps the user to visually discriminate binding sites with a score higher than a specified threshold (in this case the threshold has a value of 70). The top-ranked phosphate binding site has a high-conservation score (97.0). The prediction derives from a structural match involving three residues of the query protein and a known phosphate binding motif belonging to the *Sulfobulus tokodaii* fructose-1,6-bisphosphatase (PDB code 1umg) with an RMSD of 0.43Å. A Jmol graphical applet displays the query protein structure (grey) in ribbon style with the predicted binding sites represented as spheres and colored according to the conservation score (gold, silver and bronze). The top-ranked and highly conserved prediction is colored in gold and is located extremely close to the γ-phosphate of the crystallized ATP molecule. Two groups of buttons, located above the results table and below the applet, respectively, allow the user to view/download the results, and to interact with the 3D visualization of the protein with the predicted binding sites.

optimal values for the method parameters: the RMSD, BLOSUM62 substitution value and conservation thresholds. The best results were obtained with an RMSD threshold of 0.9 Å, 1 as the BLOSUM62 substitution threshold and 66 for the conservation score. This combination of parameters resulted in the identification of at least one correct prediction in 69% of the training proteins, with an average of 3.7 ± 0.4 false positives per structure.

The method was then tested on an independent set of 52 proteins, culled from the LigASite database (25), which bind phosphorylated ligands. Each protein was evaluated both in its apo and holo form. All the test proteins (i) are checked for redundancy with the training set (at 30% of sequence identity); (ii) do not contain mutations; (iii) bind a phosphorylated ligand. The method obtained similar results on both sets. Pfinder identified at least one correct prediction in 63% of the holo and in 62% of the apo structures with an average of 4.8 ± 0.7 false positives per structure in both cases. Previous works showed that binding site residues have lower B-factors (26) and even though they can not be located in extremely stable regions of the proteins they can not be disordered either (27). This is encouraging for our method because it implies that when we compare two binding sites of similar shape and

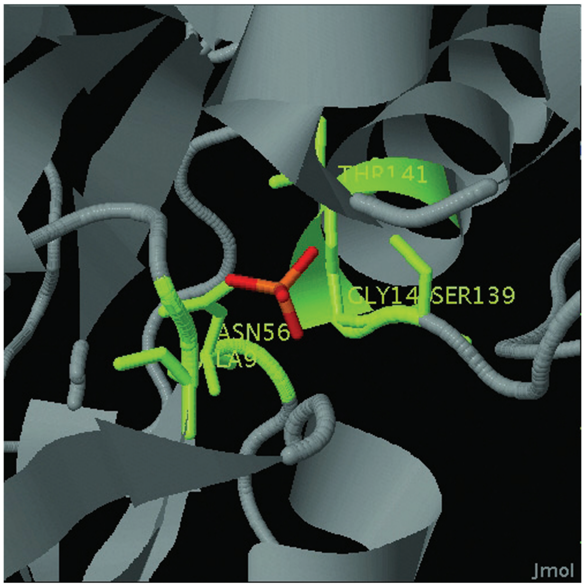


Figure 2. The graphical Jmol applet showing a known phosphate binding motif on a phosphate binding protein from *Escherichia coli* (PDB code 1ixi). This motif is composed of five amino acids binding a dihydrogenphosphate ion: Ala9, Asn56, Ser139, Gly140 and Thr141. This visualization is accessed from the results table and shows in more detail the original structure from which the motif responsible for the prediction was derived.

in the same conformation (i.e. both apo or holo), there should not be too many differences in the coordinates of the residues due to thermal motion. Moreover, it has been observed (28,29) that the variability in terms of binding site conformation is correlated with the size and flexibility of the ligand. Therefore, we investigated whether the performance of Phosfinder, in terms of the results on the holo versus apo conformation of the same protein, is correlated with the size or the number of hydrogen bond donors and acceptors of the cognate ligand. We found that this is not the case. We first considered the average distance between the best prediction and the position of the phosphate in the crystal. This measure does not differ significantly between the holo and apo sets (paired Wilcoxon test $P > 0.1$). Moreover the difference between the distance of the best prediction in the holo and apo structure of the same protein has only very weak correlations with the size of the ligand (−0.15), and the number of hydrogen bond donors (−0.15) and acceptors (−0.19). There are also no significant differences between the average size of the ligand or the number of hydrogen bond donors and acceptors for the apo proteins for which no correct prediction is made versus the ones for which phosfinder correctly predicts the location of the phosphate (Wilcoxon test $P > 0.9$ for size, >0.3 for number of H-bond acceptors, >0.5 for H-bond donors).

We think that the performance of our method is not affected when analyzing large, flexible ligands because we only consider small structural motifs mostly composed of three residues only. Therefore, even if the overall shape of the binding pocket varies when the ligand is bound, the local conformation of small groups of residues is mostly preserved.

Moreover, we also showed that our data sets are not biased in favor of nucleotide binding folds since 34 out of 59 training proteins and 35 out of 52 test proteins have a fold that is not a wide-spread nucleotide binding folds such as the Rossmann-type folds and the P-loop-containing nucleotide hydrolases. We also demonstrated that 105 out of 111 structures have at least one correct prediction made by a phosphate binding motif from a non-common nucleotide binding fold. This means that our set of template phosphate binding motifs is not biased in favor of a particular group of folds and that phosphate binding sites can be identified by motifs belonging to the different folds.

SUMMARY

Phosfinder is a web server for the prediction of phosphate binding sites in protein structures. Given the biological importance of the phosphate group Phosfinder represents a valuable resource for structural biologists. The web server provides a user-friendly version of the Pfinder method, enriched with the possibility to visualize the predicted phosphate binding sites on the query structure. Moreover, the web server incorporates a new feature: the possibility for the user to upload his own phosphate binding motifs and to search them in the query structure. One of the main advantages of Phosfinder is that it predicts the actual coordinates where the phosphate group is located, as opposed to a generic surface region. Moreover the specific amino acids that bind phosphate are also predicted. Such precise information can be used to guide drug design and molecular docking experiments. The analysis of apo/holo structure pairs (15) shows that the performance of Phosfinder is almost completely unaffected by ligand-induced conformational changes. Therefore, Phosfinder can be applied to structures of unknown function that have been crystallized without a ligand.

FUNDING

AIRC (to M.H.C.); FIRB ‘Futuro in ricerca’ Project (number E81J10000030001 to G.A.). Funding for open access charge: AIRC.

Conflict of interest statement. None declared.

REFERENCES

1. Hirsch,A.K., Fischer,F.R. and Diederich,F. (2007) Phosphate recognition in structural biology. *Angew. Chem. Int. Ed. Engl.*, **46**, 338–352.
2. Traxler,P. and Furet,P. (1999) Strategies toward the design of novel and selective protein tyrosine kinase inhibitors. *Pharmacol. Ther.*, **82**, 195–206.
3. Gitlin,J.D. (2003) Wilson disease. *Gastroenterology*, **125**, 1868–1877.
4. Ji,H.F., Kong,D.X., Shen,L., Chen,L.L., Ma,B.G. and Zhang,H.Y. (2007) Distribution patterns of small-molecule ligands in the protein universe and implications for origin of life and drug discovery. *Genome Biol.*, **8**, R176.

5. Saraste,M., Sibbald,P.R. and Wittinghofer,A. (1990) The P-loop–a common motif in ATP- and GTP-binding proteins. *Trends Biochem. Sci.*, **15**, 430–434.
6. Kleiger,G. and Eisenberg,D. (2002) GXXXG and GXXXA motifs stabilize FAD and NAD(P)-binding Rossmann folds through C(alpha)-H. O hydrogen bonds and van der waals interactions. *J. Mol. Biol.*, **323**, 69–76.
7. Wass,M.N., Kelley,L.A. and Sternberg,M.J. (2010) 3DLigandSite: predicting ligand-binding sites using similar structures. *Nucleic Acids Res.*, **38**, W469–W473.
8. Konc,J. and Janezic,D. (2010) ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, **26**, 1160–1168.
9. Konc,J. and Janezic,D. (2010) ProBiS: a web server for detection of structurally similar protein binding sites. *Nucleic Acids Res.*, **38**, W436–W440.
10. 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.
11. Lippi,M., Passerini,A., Punta,M., Rost,B. and Frasconi,P. (2008) MetalDetector: a web server for predicting metal-binding sites and disulfide bridges in proteins from sequence. *Bioinformatics*, **24**, 2094–2095.
12. Chu,W.Y., Huang,Y.F., Huang,C.C., Cheng,Y.S., Huang,C.K. and Oyang,Y.J. (2009) ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors. *Nucleic Acids Res.*, **37**, W396–W401.
13. Terribilini,M., Sander,J.D., Lee,J.H., Zaback,P., Jernigan,R.L., Honavar,V. and Dobbs,D. (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, **35**, W578–W584.
14. Petsalaki,E., Stark,A., Garcia-Urdiales,E. and Russell,R.B. (2009) Accurate prediction of peptide binding sites on protein surfaces. *PLoS Comput. Biol.*, **5**, e1000335.
15. Parca,L., Gherardini,P.F., Helmer-Citterich,M. and Ausiello,G. (2010) Phosphate binding sites identification in protein structures. *Nucleic Acids Res.*
16. Ausiello,G., Gherardini,P.F., Gatti,E., Incani,O. and Helmer-Citterich,M. (2009) Structural motifs recurring in different folds recognize the same ligand fragments. *BMC Bioinformatics*, **10**, 182.
17. Gherardini,P.F., Ausiello,G., Russell,R.B. and Helmer-Citterich,M. (2010) Modular architecture of nucleotide-binding pockets. *Nucleic Acids Res.*, **38**, 3809–3816.
18. 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.
19. Ausiello,G., Via,A. and Helmer-Citterich,M. (2005) Query3d: a new method for high-throughput analysis of functional residues in protein structures. *BMC Bioinformatics*, **6(Suppl 4)**, S5.
20. 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.
21. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.
22. 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–D222.
23. Rose,P.W., Beran,B., Bi,C., Bluhm,W.F., Dimitropoulos,D., Goodsell,D.S., Prlic,A., Quesada,M., Quinn,G.B., Westbrook,J.D. *et al.* (2011) The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res.*, **39**, D392–D401.
24. Zhao,S., Morris,G.M., Olson,A.J. and Goodsell,D.S. (2001) Recognition templates for predicting adenylate-binding sites in proteins. *J. Mol. Biol.*, **314**, 1245–1255.
25. 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.
26. Yuan,Z., Zhao,J. and Wang,Z.X. (2003) Flexibility analysis of enzyme active sites by crystallographic temperature factors. *Protein Eng.*, **16**, 109–114.
27. Luque,I. and Freire,E. (2000) Structural stability of binding sites: consequences for binding affinity and allosteric effects. *Proteins*, **41(Suppl 4)**, 63–71.
28. Kahraman,A., Morris,R.J., Laskowski,R.A. and Thornton,J.M. (2007) Shape variation in protein binding pockets and their ligands. *J. Mol. Biol.*, **368**, 283–301.
29. Weng,Y.Z., Chang,D.T., Huang,Y.F. and Lin,C.W. (2011) A study on the flexibility of enzyme active sites. *BMC Bioinformatics*, **12(Suppl 1)**, S32.