D268–D271 *Nucleic Acids Research, 2011, Vol. 39, Database issue* *Published online 21 October 2010*
doi:10.1093/nar/gkq936

Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites

Andreas Zanzoni[1,*], Daniel Carbajo[2], Francesca Diella[3,4], Pier Federico Gherardini[5], Anna Tramontano[2], Manuela Helmer-Citterich[5,*] and Allegra Via[2,*]

[1]Institute for Research in Biomedicine, Joint IRB-BSC program in Computational Biology, c/ Baldiri Reixac 10, 08028 Barcelona, Spain, [2]Biocomputing group, Department of Biochemical Sciences 'A. Rossi Fanelli', Sapienza University of Rome, P.le Aldo Moro 5, Rome, Italy, [3]European Molecular Biology Laboratory, Postfach 10.2209, 69012 Heidelberg, [4]Biobyte solutions GmbH, 69126 Heidelberg, Germany and [5]Centre for Molecular Bioinformatics, Department of Biology, University of Rome Tor Vergata, Italy

Received August 14, 2010; Revised September 17, 2010; Accepted September 29, 2010

ABSTRACT

Phospho3D is a database of three-dimensional (3D) structures of phosphorylation sites (P-sites) derived from the Phospho.ELM database, which also collects information on the residues surrounding the P-site in space (3D *zones*). The database also provides the results of a large-scale structural comparison of the 3D *zones* versus a representative dataset of structures, thus associating to each P-site a number of structurally similar sites. The new version of Phospho3D presents an 11-fold increase in the number of 3D sites and incorporates several additional features, including new structural descriptors, the possibility of selecting non-redundant sets of 3D structures and the availability for download of non-redundant sets of structurally annotated P-sites. Moreover, it features P3Dscan, a new functionality that allows the user to submit a protein structure and scan it against the 3D *zones* collected in the Phospho3D database. Phospho3D version 2.0 is available at: http://www.phospho3d.org/.

INTRODUCTION

During recent years, there has been an increasing interest in the structural features of protein phosphorylation sites (P-sites). This fact can be ascribed to the steadily growing experimentally verified P-sites provided by high-throughput mass spectrometry-based proteomics techniques [e.g. (1)]. The simultaneous availability of an increasing number of three-dimensional (3D) structures is making it possible to infer the structural context for a significant number of P-sites. In order to identify the structural determinants of kinase specificity, some authors tried to characterize the 3D environment of P-sites with the aim of pinpointing specific structural features (2–4). Both phosphorylation sites databases, also reporting structural data, and interesting systematic structural analyses of P-sites have recently appeared in the literature [for a review, see ref. (5)]. PHOSIDA (1), a phosphorylation sites database, includes the predicted accessibility and secondary structure of each P-site. The mtcPTM (6) database stores homology models for proteins and protein domains that contain phosphorylated residues. Finally, many of the P-site predictors incorporating 3D-context information (1,4,7–10) display an improvement in performance with respect to predictors using sequence information only.

So far the structural attributes stored in P-site databases and incorporated in P-site predictors are essentially of two types: accessibility and secondary structure.

In 2007, we presented Phospho3D, a database of reliably predicted 3D structures of protein phosphorylation sites (11), derived from the Phospho.ELM database (12) and enriched with structural annotation at the residue level, including accessibility, secondary structure and residue conservation as from the Consurf-HSSP database (13). Phospho3D also stored and annotated the sequence flanking the P-site (10 residues) and the *zone*, i.e. the 3D region defined by the set of residues at a distance not exceeding 12 Å from the phospho-instance.

Since then, the number of Phospho.ELM instances increased ~4-fold, raising from 5314 (in 1805 proteins) to 42 474 (in 8718 proteins), and more than 26 000 structures were added to the Protein Data Bank (PDB) (14).

*To whom correspondence should be addressed. Tel: +34 93403 9689; Fax: +34 93403 9954; Email: andreas.zanzoni@irbbarcelona.org
Correspondence may also be addressed to Manuela Helmer-Citterich. Tel: +39 067259 4324; Fax: +39 06202 3500; Email: citterich@uniroma2.it
Correspondence may also be addressed to Allegra Via. Tel: +39 064991 0957; Fax: +39 06444 0062; Email: allegra.via@uniroma1.it

Here we introduce Phospho3D version 2.0, which—besides an eleven-fold increase in the number of Phospho.ELM unique instances mapped onto 3D structures (compared to version 1.0)—incorporates several new features, including additional structural descriptors of P-sites, the possibility of browsing the database selecting non-redundant sets of 3D structures, the availability for download of many non-redundant sets of structurally annotated P-sites—aimed at serving as reliable benchmark datasets for predictors' training and test—and P3Dscan, a new functionality that allows the user to submit a protein structure and scan it against the 3D P-site *zones* collected in the Phospho3D database.

DATABASE CONTENTS

The updated Phospho3D database was constructed by collecting data from the latest release of the Phospho.ELM database (Version 9.0, August 2010), which currently stores about 42 500 experimentally verified phosphorylation sites in 8718 substrate proteins, both manually extracted from the literature and obtained from mass spectrometry-based proteomics experiments. The correspondence between Phospho.ELM sequences and PDB chains was based on sequence alignment using at least 98% sequence identity. P-sites in gapped regions of the alignment were discarded.

This resulted in 5387 mapped instances (1770 unique Phospho.ELM instances on 2158 protein chains—897 Ser, 338 Thr, 535 Tyr).

Notice that P-sites derived from mass spectrometry (MS) experiments should be taken with caution. In fact, due to the current procedures for MS data deposition, it is difficult to systematically detect if a phospho-instance was identified in physiologically abnormal conditions (e.g. in proteins extracted from oncogenic tissues or that do not undergo phosphorylation, such as hemoglobin) (15). In order to help users detect such potentially problematic cases, we reported—for each P-site—the nature of the original experiment (low- or high-throughput) and the corresponding literature reference (PMID). Moreover, we encourage users to carefully analyze the structural context of P-sites, which might be indicative of problems in the original data. One example is represented by the Tyr phosphorylation site mapped to position 133 of the human hemoglobin subunit beta (UniProtKB:P68871), for which Phospho3D stores 43 PDB structures. In most of the reported structures, the solvent accessibility of Y133 is zero and it is never >3.5%. This structural information suggests that the original data might not be reliable.

The basic information stored in Phospho3D consists of the P-site instance, its flanking sequence (10 residues) and the P-site 3D *zone*, i.e. the set of residues in a 12 Å *radius* surrounding the P-site in space. For each residue in the *zone* Phospho3D 2.0 stores the following structural descriptors: secondary structure and solvent accessibility (in Å^2) as defined by DSSP (16); percentage solvent accessibility, obtained by normalizing the DSSP solvent accessibility by the maximum accessibility value for each residue as determined in ref. (17); B-factor, computed as specified in ref. (18); occurrence in a cavity together with the rank and volume of the cavity calculated with the SURFNET program (19); the depth index DPX (20) and the protrusion index CX (21), obtained using the PSAIA software (22); the CONSURF conservation score extracted from ConSurfDB (23); the disorder probability provided by DisEMBL (24) according to three different criteria: (i) loops/coils as defined by DSSP, (ii) hot loops, i.e. loops with a high degree of mobility as determined by temperature (B-) factor and (iii) missing coordinates in X-Ray structure as defined by REMARK-465 entries in PDB.

A detailed description of each structural attribute is reported in the website documentation.

Phospho3D 2.0 also provides information derived from Phospho.ELM, such as, when available, the kinase(s) phosphorylating a given P-site, and, for each *zone*, the results of a large-scale local structural comparison versus a non redundant (sequence identity ≤20%) dataset of 487 PDB X-ray protein chains with experimental resolution ≤1.5 Å extracted from eukaryotic organisms. The comparison is carried out using the new version of the algorithm (25) and the same criteria for assessing structural similarity used in the previous database version (11) although more stringent thresholds are applied in this case, as described in the website documentation. The database queries can now be performed on seven PDB non-redundant sets: the whole collection of P-sites, the set of P-sites found in non-identical structures (PDB100) and P-sites found in PDB structures belonging to five redundancy sets, ranging from PDB90 to PDB20, where the number corresponds to the maximum sequence identity shared by the protein chains in the redundancy set. These sets have been determined using the PISCES resource (26).

Additionally, the P-site annotations at the residue level are available for download on the Phospho3D website. These can serve as benchmark for P-site predictors' training and test and for analyses of P-site structural features.

Finally, Phospho3D 2.0 now links each entry to the corresponding Phospho.ELM instance and the kinase names to their UniProt ACs (27).

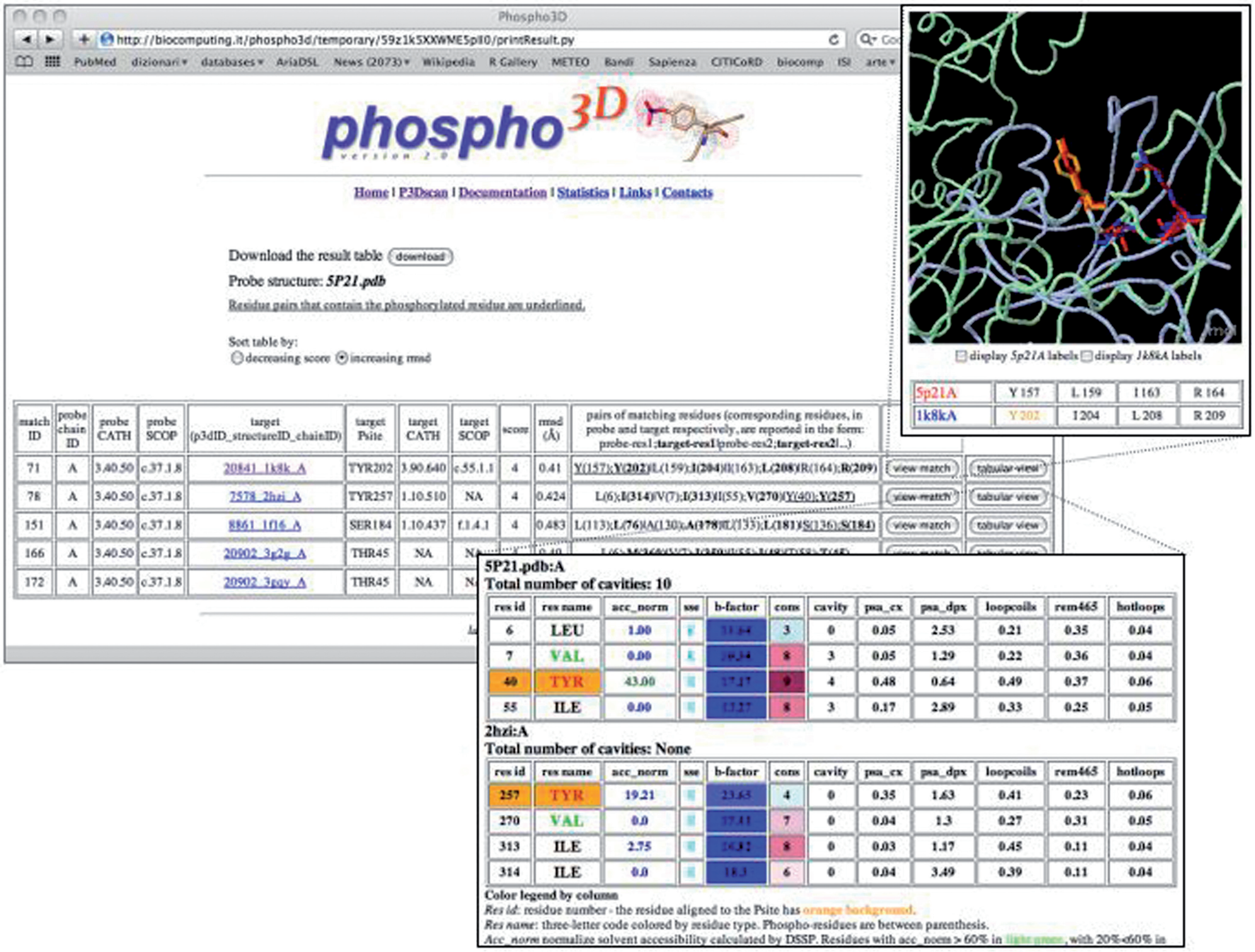
P3DSCAN

Phospho3D 2.0 provides a novel functionality that allows the user to upload a PDB-formatted structure and perform a local structural comparison against the 5387 *zones* (one for each Phospho.ELM mapped instance) stored in the database, aimed at identifying local structural similarities between the user query structure and one of the structural patches containing a P-site. In order to evaluate the structural context of each match, we provide its graphical display and a table reporting the structural information at the residue level of both the query and the target 3D matching patches. The comparison algorithm—that P3Dscan runs on-the-fly—is the same as the one used for the large-scale comparison whose results are stored in the database. The comparison results are also provided in text format for download.

THE WEB INTERFACE

Similarly to the previous version, Phospho3D 2.0 can be searched by kinase name, by PDB identification code or by keyword. In this new version, however, the user can additionally select a redundancy set in order to avoid retrieving identical or very similar P-sites. The data returned to the user consist of a brief description of the PDB structure(s) that fulfill the search criteria and a list of instances presented along with associated information. In particular, each instance is now linked to the corresponding Phospho.ELM entry. For each P-site, the user can select three options related to the surrounding structural zone: a graphical view using the Jmol Java Applet (http://www.jmol.org), a tabular view reporting the *zone* annotation at the residue level or a list of 3D matches identified by local structural comparison. Each match can be visualized using Jmol.

The P3Dscan webpage can be reached from the Phospho3D homepage. Users upload a PDB-formatted file, choose a redundancy set of 3D *zones* they want to scan against their structure and run the comparison by clicking the 'p3d scan' button. P3Dscan results are displayed in tabular format (Figure 1). The result table can be sorted by increasing match score or decreasing RMSD. Each line of the table reports the information of a single match. A match can be graphically visualized by clicking on the corresponding button. Moreover, the tabular view button links to a window displaying structural annotation at the residue level, both for the query and the target 3D patches. The Result Table can also be downloaded in text format.


Phospho3D
http://biocomputing.it/phospho3d/temporary/59z1k5XXWME5pll0/printResult.py
PubMed dizionari databases AriaDSL News (2073) Wikipedia R Gallery METEO Bandi Sapienza CITICoRD biocomp ISI arte
phospho 3D
version 2.0
Home | P3Dscan | Documentation | Statistics | Links | Contacts
jmol

Download the result table (download)

Probe structure: ***5P21.pdb***

Residue pairs that contain the phosphorylated residue are underlined.

Sort table by:
○ decreasing score ● increasing rmsd

match ID	probe chain ID	probe CATH	probe SCOP	target (p3dID_structureID_chainID)	target Psite	target CATH	target SCOP	score	rmsd (Å)	pairs of matching residues (corresponding residues, in probe and target respectively, are reported in the form: probe-res1;**target-res1**|probe-res2;**target-res2**|...)		
71	A	3.40.50	c.37.1.8	20841_1k8k_A	TYR202	3.90.640	c.55.1.1	4	0.41	Y(157);**Y(202)**|L(159);**I(204)**|I(163);**L(208)**|R(164);**R(209)**	view match	tabular view
78	A	3.40.50	c.37.1.8	7578_2hzi_A	TYR257	1.10.510	NA	4	0.424	L(6);**I(314)**|V(7);**I(313)**|I(55);**V(270)**|Y(40);**Y(257)**	view match	tabular view
151	A	3.40.50	c.37.1.8	8861_1f16_A	SER184	1.10.437	f.1.4.1	4	0.483	L(113);**L(76)**|A(130);**A(178)**|L(133);**L(181)**|S(136);**S(184)**	view match	tabular view
166	A	3.40.50	c.37.1.8	20902_3g2g_A	THR45	NA	NA	[illegible]	0.40	L(6);**M(360)**|V(7);**I(380)**|I(55);**I(48)**|D(58);**T(45)**	view match	tabular view
172	A	3.40.50	c.37.1.8	20902_3gqy_A	THR45	NA	N	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

☐ display *5p21A* labels ☐ display *1k8kA* labels

5p21A	Y 157	L 159	I 163	R 164
1k8kA	Y 202	I 204	L 208	R 209

5P21.pdb:A
Total number of cavities: 10

res id	res name	acc_norm	sse	b-factor	cons	cavity	psa_cx	psa_dpx	loopcoils	rem465	hotloops
6	LEU	1.00		31.64	3	0	0.05	2.53	0.21	0.35	0.04
7	VAL	0.00	E	20.34	8	3	0.05	1.29	0.22	0.36	0.04
40	TYR	43.00		17.17	9	4	0.48	0.64	0.49	0.37	0.06
55	ILE	0.00		13.27	8	3	0.17	2.89	0.33	0.25	0.05

2hzi:A
Total number of cavities: None

res id	res name	acc_norm	sse	b-factor	cons	cavity	psa_cx	psa_dpx	loopcoils	rem465	hotloops
257	TYR	19.21		23.63	4	0	0.35	1.63	0.41	0.23	0.06
270	VAL	0.0		37.41	7	0	0.04	1.3	0.27	0.31	0.05
313	ILE	2.75		16.82	8	0	0.03	1.17	0.45	0.11	0.04
314	ILE	0.0		18.3	6	0	0.04	3.49	0.39	0.11	0.04

Color legend by column
Res id: residue number - the residue aligned to the Psite has orange background.
Res name: three-letter code colored by residue type. Phospho-residues are between parenthesis.
Acc_norm normalize solvent accessibility calculated by DSSP. Residues with acc_norm > 60% in light green, with 20%<60% in

Figure 1. The P3Dscan output page for the crystal structure of the H-ras oncogene protein p21 (PDB 5P21). The table reports the list of matches between the query protein (probe) and the zone(s) collected in the database (target). SCOP and CATH annotation, when available, is reported for both the query structure and for each matching zone therefore the user can discriminate matches due to overall fold similarity from functionally interesting ones. The 3D zone (column 5) is linked to the corresponding Phospho3D database entry and the phospho-residue is explicitly reported in column 6. The score (column 9) corresponds to the number of paired residues in the match. The RMSD (column 10) is calculated on the matching amino acids. The pairs of residues participating in the match are reported in column 11. Upper right inset: graphical display (Jmol) of the probe and target structures. Residues participating in the match are in stick and the P-site is colored in orange. Bottom inset: tabular view: structural information at the residue level is reported for both the probe (user query structure) and target (3D zone) residues involved in the match.

STRUCTURAL ANALYSIS

We performed a large-scale structural analysis of the P-sites stored in Phospho3D and plotted the statistical distributions of each 3D attribute used to annotate the P-sites in the database. The analysis was carried out separately for each redundancy set. The distributions for the P-sites falling on non-identical structures (PISCES PDB100) can be found at http://www.phospho3d.org/stats.py#3.

CONCLUSIONS

The new version of Phospho3D stores a markedly increased number of structurally annotated P-sites. In addition, it incorporates new significant improvements, such as several new structural descriptors, non-redundant datasets and a tool, P3Dscan, for the analysis of uploaded protein structures.

We believe that this enhanced version of the database makes it possible to fully exploit available structural information on P-sites and use it to perform structural analyses and/or build P-site predictors.

Importantly, the Phospho3D update procedure is now completely automated, allowing regular and timely updates of the database with each new Phospho.ELM release.

ACKNOWLEDGEMENTS

Many thanks to Holger Dinkel (Phospho.ELM database) for technical support.

FUNDING

This work was supported by Istituto Pasteur—Fondazione Cenci Bolognetti, Roma; the 7th EC Framework Programme LEISHDRUG project [grant number 223414]; and a 'Juan de la Cierva' fellowship to A.Z. Funding for open access charge: 7th EC Framework Programme LEISHDRUG project.

Conflict of interest statement. None declared.

REFERENCES

1. Gnad,F., Ren,S., Cox,J., Olsen,J.V., Macek,B., Oroshi,M. and Mann,M. (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites. *Genome Biol.*, **8**, R250.
2. Fan,S.C. and Zhang,X.G. (2005) Characterizing the microenvironment surrounding phosphorylated protein sites. *Genomics Proteomics Bioinformatics*, **3**, 213–217.
3. Kitchen,J., Saunders,R.E. and Warwicker,J. (2008) Charge environments around phosphorylation sites in proteins. *BMC Struct. Biol.*, **8**, 19.
4. Durek,P., Schudoma,C., Weckwerth,W., Selbig,J. and Walther,D. (2009) Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins. *BMC Bioinformatics*, **10**, 117.
5. Via,A., Diella,F., Gibson,T.J. and Helmer-Citterich,M. (2011) From sequence to structural analysis in protein phosphorylation motifs. *Front. Biosci.*, **16**, 1261–1275.
6. Jimenez,J.L., Hegemann,B., Hutchins,J.R., Peters,J.M. and Durbin,R. (2007) A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database. *Genome Biol.*, **8**, R90.
7. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.
8. Brinkworth,R.I., Breinl,R.A. and Kobe,B. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl Acad. Sci. USA*, **100**, 74–79.
9. Plewczynski,D., Tkacz,A., Godzik,A. and Rychlewski,L. (2005) A support vector machine approach to the identification of phosphorylation sites. *Cell Mol. Biol. Lett.*, **10**, 73–89.
10. Plewczynski,D., Jaroszewski,L., Godzik,A., Kloczkowski,A. and Rychlewski,L. (2005) Molecular modeling of phosphorylation sites in proteins using a database of local structure segments. *J. Mol. Model.*, **11**, 431–438.
11. Zanzoni,A., Ausiello,G., Via,A., Gherardini,P.F. and Helmer-Citterich,M. (2007) Phospho3D: a database of three-dimensional structures of protein phosphorylation sites. *Nucleic Acids Res.*, **35**, D229–D231.
12. Diella,F., Gould,C.M., Chica,C., Via,A. and Gibson,T.J. (2008) Phospho.ELM: a database of phosphorylation sites – update 2008. *Nucleic Acids Res.*, **36**, D240–D244.
13. Glaser,F., Rosenberg,Y., Kessel,A., Pupko,T. and Ben-Tal,N. (2005) The ConSurf-HSSP database: the mapping of evolutionary conservation among homologs onto PDB structures. *Proteins*, **58**, 610–617.
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. Nichols,A.M. and White,F.M. (2009) Manual validation of peptide sequence and sites of tyrosine phosphorylation from MS/MS spectra. *Methods Mol. Biol.*, **492**, 143–160.
16. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
17. Miller,S., Janin,J., Lesk,A.M. and Chothia,C. (1987) Interior and surface of monomeric proteins. *J. Mol. Biol.*, **196**, 641–656.
18. Bartlett,G.J., Porter,C.T., Borkakoti,N. and Thornton,J.M. (2002) Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.*, **324**, 105–121.
19. Laskowski,R.A. (1995) SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph*, **13**, 323–330, 307–328.
20. Pintar,A., Carugo,O. and Pongor,S. (2003) DPX: for the analysis of the protein core. *Bioinformatics*, **19**, 313–314.
21. Pintar,A., Carugo,O. and Pongor,S. (2002) CX, an algorithm that identifies protruding atoms in proteins. *Bioinformatics*, **18**, 980–984.
22. Mihel,J., Sikic,M., Tomic,S., Jeren,B. and Vlahovicek,K. (2008) PSAIA - protein structure and interaction analyzer. *BMC Struct. Biol.*, **8**, 21.
23. Goldenberg,O., Erez,E., Nimrod,G. and Ben-Tal,N. (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res.*, **37**, D323–D327.
24. Linding,R., Jensen,L.J., Diella,F., Bork,P., Gibson,T.J. and Russell,R.B. (2003) Protein disorder prediction: implications for structural proteomics. *Structure*, **11**, 1453–1459.
25. 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.
26. Wang,G. and Dunbrack,R.L. Jr (2003) PISCES: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.
27. The Universal Protein Resource (UniProt) (2010). *Nucleic Acids Res.*, **38**, D142–D148.