Published online 16 November 2006 *Nucleic Acids Research, 2007, Vol. 35, Database issue* ***D229–D231***
doi:10.1093/nar/gkl922

Phospho3D: a database of three-dimensional structures of protein phosphorylation sites

Andreas Zanzoni*, Gabriele Ausiello, Allegra Via, Pier Federico Gherardini and Manuela Helmer-Citterich

Centre for Molecular Bioinformatics, Department of Biology, University of Rome 'Tor Vergata', Rome 00133, Italy

Received August 11, 2006; Revised October 13, 2006; Accepted October 17, 2006

ABSTRACT

Phosphorylation is the most common protein post-translational modification. Phosphorylated residues (serine, threonine and tyrosine) play critical roles in the regulation of many cellular processes. Since the amount of data produced by screening assays is growing continuously, the development of computational tools for collecting and analysing experimental data has become a pivotal task for unravelling the complex network of interactions regulating eukaryotic cell life. Here we present Phospho3D, http://cbm.bio.uniroma2.it/phospho3d, a database of 3D structures of phosphorylation sites, which stores information retrieved from the phospho.ELM database and is enriched with structural information and annotations at the residue level. The database also collects the results of a large-scale structural comparison procedure providing clues for the identification of new putative phosphorylation sites.

INTRODUCTION

The phosphorylation of specific protein residues is a crucial event in the regulation of several cellular processes, operating on activation, deactivation or recognition of the target protein. A great deal of eukaryotic proteins (∼30% in the human genome) undergo this reversible post-translational modification (1). Phosphorylation on serine/threonine or tyrosine residues is accomplished by protein kinases (PKs), one of the largest protein families, comprising 1.5–2.5% of all eukaryotic genes (2).

Although the amount of data produced in various screening assays is steadily growing (3–6), experimental identification of phosphoproteins and the determination of individual phosphorylation sites remains a difficult and time-consuming task. Hence, the implementation of computational tools proves to be very useful for collecting and analysing experimental data.

Several sequence-based methods to predict phosphorylation sites were developed using different computational approaches such as regular expressions with context-based rules (7), position-specific scoring matrices (PSSMs) (8), artificial neural networks (9,10), support vector machines (SVMs) (11,12), hidden Markov models (13) and iterative statistical methods (14). All these methods are based on the hypothesis that the sequence surrounding the phosphorylated residue represents the main determinant for kinase specificity. They are reasonably accurate and work well with a number of specific kinases. However, the specificity determinants and rules remain elusive for a large number of protein kinases that display a number of substrates sharing little or no sequence similarity in the known phosphopeptides. We propose that, at least in some cases, the rules of kinase specificity may reside in the presence of structural determinants which only occasionally overlap with sequence *consensi* and which might be independent of the residue order in protein sequences.

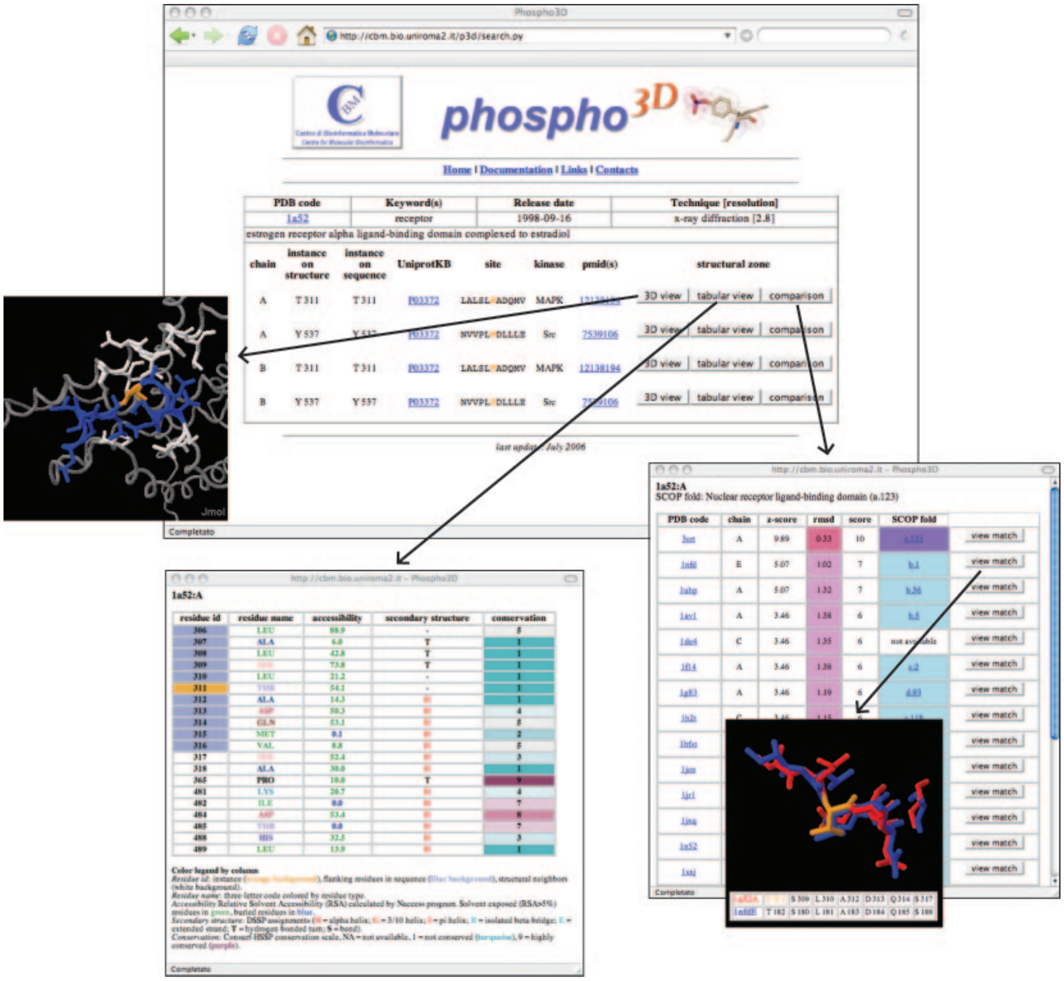
Here we describe Phospho3D, a database of 3D structure of phosphorylation sites. It collects information retrieved from the phospho.ELM database (15) and is enriched with structural information and diverse annotations at the residue level. In addition, the database stores the results of a large-scale local structural comparison which suggest functional annotation of phosphorylation sites by 3D similarity. Cases of significant structural similarity between phosphorylation sites may indicate that they are phosphorylated by the same kinase.

DATABASE CONSTRUCTION AND CONTENT

The Phospho3D database was constructed by collecting data from the phospho.ELM database which gathers experimentally verified phosphorylation sites manually extracted from the literature. The phospho.ELM dataset used in this work (version 4.0) contains 5314 phosphorylation sites, or instances, belonging to 1805 different sequences.

The correspondence between phospho.ELM sequences and the Protein Data Bank (PDB) chains was established via the Seq2Struct resource (16), an exhaustive collection of

*To whom correspondence should be addressed. Tel: +39 067 259 4314; Fax: +39 062 023 500; Email: andreas@cbm.bio.uniroma2.it



Phospho3D
http://cbm.bio.uniroma2.it/p3d/search.py
Centro di Bioinformatica Molecolare
Centre for Molecular Bioinformatics
phospho3D
Jmol
Completato

Home | Documentation | Links | Contacts

PDB code	Keyword(s)	Release date	Technique [resolution]
1a52	receptor	1998-09-16	x-ray diffraction [2.8]

estrogen receptor alpha ligand-binding domain complexed to estradiol

chain	instance on structure	instance on sequence	UniprotKB	site	kinase	pmid(s)	structural zone
A	T 311	T 311	P03372	LALSL ADQMV	MAPK	12138194	3D view | tabular view | comparison
A	Y 537	Y 537	P03372	NVVPL DLLLE	Src	7539106	3D view | tabular view | comparison
B	T 311	T 311	P03372	LALSL ADQMV	MAPK	12138194	3D view | tabular view | comparison
B	Y 537	Y 537	P03372	NVVPL DLLLE	Src	7539106	3D view | tabular view | comparison

last update: July 2006

http://cbm.bio.uniroma2.it – Phospho3D

1a52:A

residue id	residue name	accessibility	secondary structure	conservation
306	LEU	88.9	-	5
307	ALA	6.0	T	1
308	LEU	42.8	T	1
309	SER	73.8	T	1
310	LEU	21.2	-	1
311	THR	54.1	-	1
312	ALA	14.3	H	1
313	ASP	50.3	H	4
314	GLN	53.1	H	5
315	MET	0.1	H	2
316	VAL	8.8	H	5
317	SER	52.4	H	3
318	ALA	30.0	H	1
365	PRO	10.0	T	9
481	LYS	20.7	H	4
482	ILE	0.0	H	7
484	ASP	53.4	H	8
485	THR	0.0	H	7
488	HIS	32.5	H	3
489	LEU	13.9	H	1

Color legend by column
Residue id: instance (orange background), flanking residues in sequence (lilac background), structural neighbors (white background).
Residue name: three-letter code colored by residue type.
Accessibility: Relative Solvent Accessibility (RSA) calculated by Naccess program. Solvent exposed (RSA>5%) residues in green, buried residues in blue.
Secondary structure: DSSP assignments (H = alpha helix; G = 3/10 helix; I = pi helix; B = isolated beta-bridge; E = extended strand; T = hydrogen bonded turn; S = bend).
Conservation: Consurf-HSSP conservation scale, NA = not available, 1 = not conserved (turquoise), 9 = highly conserved (purple).

http://cbm.bio.uniroma2.it – Phospho3D

1a52:A
SCOP fold: Nuclear receptor ligand-binding domain (a.123)

PDB code	chain	z-score	rmsd	score	SCOP fold	
3ert	A	9.89	0.33	10	a.123	view match
1nfd	E	5.07	1.02	7	b.1	view match
1uhp	A	5.07	1.32	7	b.36	view match
1avl	A	3.46	1.38	6	b.5	view match
1de4	C	3.46	1.35	6	not available	view match
1f14	A	3.46	1.38	6	c.2	view match
1g83	A	3.46	1.19	6	d.93	view match
1h2t	C	3.46	1.15	6	c.118	view match
1h6o						view match
1jcn						view match
1jrl						view match
1jxq						view match
1n52						view match
1sxj						view match

1a52A	T 311	S 309	L 310	A 312	D 313	Q 314	S 317
1nfdE	T 182	S 180	L 181	A 183	D 184	Q 185	S 188

Figure 1. In the central panel a list of instances for the PDB file 1A52 is shown. For each of them, users can visualize the corresponding *zone* via the Jmol viewer, the annotation at the residue level and the results of the large-scale local structural comparison. For each structural match the score, the *Z*-score, and the rmsd are reported along with the SCOP fold (27) of the matching PDB files.

annotated links between SwissProt-TrEMBL and PDB sequences. Links are based on sequence alignment using pre-established highly reliable thresholds. From a list of 4530 sequence–structure links (for further details see website documentation), only the ones having the phosphorylable residue in the alignment region were retained, this resulting in 2726 instances (166 unique phospho.ELM instances on 1219 protein chains).

The basic information stored in Phospho3D consists of the instance, its flanking sequence (10 residues) and any residue whose distance from the instance does not exceed 12 Å thus defining a 3D neighbourhood which we define as *zone*.

For each *zone*, annotation at the residue level is provided, namely solvent accessibility supplied by the NACCESS program (17), secondary structure assignment given by the DSSP program (18) and residue conservation as from the Consurf-HSSP database (19).

Users can also retrieve information extracted from the phospho.ELM dataset; for instance, the Medline reference PMID and, when available, the kinase(s) that phosphorylate(s) the given site.

In addition, for each *zone* the results of a large-scale local structural comparison versus a representative dataset of PDB (20) protein chains from eukaryotic organisms are also given. The comparison was carried out using the Query3D sequence/fold independent algorithm (21). Structural matches are assessed by two criteria: structural similarity and biochemical similarity. The structural similarity demands that matching residues have a root mean square deviation (r.m.s.d.) lower than a given threshold, whereas the biochemical similarity is evaluated using a Dayhoff substitution matrix (22). The score of the match is the number of matching residues which fulfil the similarity criteria. The significance of the score is evaluated by calculating the *Z*-score over the score distribution of the query *zone* comparison to the whole dataset.

THE WEB INTERFACE

The Phospho3D database can be searched by kinase name, by PDB identification code or keyword. A browsing function has been also implemented.

The information returned to the user consists of a brief description of the PDB structure(s) which fulfil the search criterion and of a list of instances presented along with associated information (Figure 1). For each instance, 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; a list of 3D matches identified by local structural comparison. Each match can be visualized using Jmol. A tabular view of the matching residues is also presented (Figure 1).

CONCLUSION AND FUTURE PERSPECTIVES

The Phospho3D database is a useful tool for the analysis of the structural features of experimentally verified phosphorylation sites. Moreover, it provides the results of a large-scale local structural comparison between the *zones* and a representative set of eukaryotic protein chains. The results of such a comparison identify new putative phosphorylation sites and suggest the kinase(s) responsible for phosphorylation.

Phospho3D will be regularly updated as soon as the new Phospho.ELM datasets are released. The annotations will be integrated as a feature in the pdbFun server (23). We are also planning to identify and annotate those sites which are recognized by protein phosphatases and phosphoresidues-binding modules (24–26).

The Phospho3D dataset (annotations at the residue level and structural comparison results) is available upon request.

ACKNOWLEDGEMENTS

The authors would like to thank Francesca Diella for support and suggestions and for providing the Phospho.ELM dataset. We acknowledge the support of AIRC, Telethon (GGP04273), a PNR 2001–2003 (FIRB art.8) and a PNR 2003–2007 (FIRB art.8). Funding to pay the Open Access publication charges for this article was provided by AIRC.

Conflict of interest statement. None declared.

REFERENCES

1. Cohen,P. (2002) The origins of protein phosphorylation. *Nature Cell. Biol*., **4**, E127–E130.
2. Manning,G., Plowman,G.D., Hunter,T. and Sudarsanam,S. (2002) Evolution of protein kinase signaling from yeast to man. *Trends Biochem. Sci*., **27**, 514–520.
3. Salomon,A.R., Ficarro,S.B., Brill,L.M., Brinker,A., Phung,Q.T., Ericson,C., Sauer,K., Brock,A., Horn,D.M., Schultz,P.G. *et al*. (2003) Profiling of tyrosine phosphorylation pathways in human cells using mass spectrometry. *Proc. Natl Acad. Sci. USA*, **100**, 443–448.
4. Shu,H., Chen,S., Bi,Q., Mumby,M. and Brekken,D.L. (2004) Identification of phosphoproteins and their phosphorylation sites in the WEHI-231 B lymphoma cell line. *Mol. Cell. Proteomics*, **3**, 279–286.
5. Brill,L.M., Salomon,A.R., Ficarro,S.B., Mukherji,M., Stettler-Gill,M. and Peters,E.C. (2004) Robust phosphoproteomic profiling of tyrosine phosphorylation sites from human T cells using immobilized metal affinity chromatography and tandem mass spectrometry. *Anal. Chem*., **76**, 2763–2772.
6. Beausoleil,S.A., Jedrychowski,M., Schwartz,D., Elias,J.E., Villen,J., Li,J., Cohn,M.A., Cantley,L.C. and Gygi,S.P. (2004) Large-scale characterization of HeLa cell nuclear phosphoproteins. *Proc. Natl Acad. Sci. USA*, **101**, 12130–12135.
7. Puntervoll,P., Linding,R., Gemund,C., Chabanis-Davidson,S., Mattingsdal,M., Cameron,S., Martin,D.M., Ausiello,G., Brannetti,B., Costantini,A. *et al*. (2003) ELM server: a new resource for investigating short functional sites in modular eukaryotic proteins. *Nucleic Acids Res*., **31**, 3625–3630.
8. Obenauer,J.C., Cantley,L.C. and Yaffe,M.B. (2003) Scansite 2.0: proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*., **31**, 3635–3641.
9. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol*., **294**, 1351–1362.
10. Iakoucheva,L.M., Radivojac,P., Brown,C.J., O'Connor,T.R., Sikes,J.G., Obradovic,Z. and Dunker,A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res*., **32**, 1037–1049.
11. 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.
12. Kim,J.H., Lee,J., Oh,B., Kimm,K. and Koh,I. (2004) Prediction of phosphorylation sites using SVMs. *Bioinformatics*, **20**, 3179–3184.
13. Huang,H.D., Lee,T.Y., Tzeng,S.W., Wu,L.C., Horng,J.T., Tsou,A.P. and Huang,K.T. (2005) Incorporating hidden Markov models for identifying protein kinase-specific phosphorylation sites. *J. Comput. Chem*., **26**, 1032–1041.
14. Schwartz,D. and Gygi,S.P. (2005) An iterative statistical approach to the identification of protein phosphorylation motifs from large-scale data sets. *Nat. Biotechnol*., **23**, 1391–12398.
15. Diella,F., Cameron,S., Gemund,C., Linding,R., Via,A., Kuster,B., Sicheritz-Ponten,T., Blom,N. and Gibson,T.J. (2005) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, **5**, 79.
16. Via,A., Zanzoni,A. and Helmer-Citterich,M. (2005) Seq2Struct: a resource for establishing sequence-structure links. *Bioinformatics*, **21**, 551–553.
17. Hubbard,S. and Thornton,J.M. (1993) NACCESS Computer Program. Department of Biochemistry and Molecular Biology, University College, London.
18. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.
19. 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.
20. 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.
21. 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*, **4**, S5.
22. Dayhoff,M.O., Schwartz,R.M. and Orcutt,B.C. (1978) A model of evolutionary change in proteins. *Atlas Prot. Seq. Struct*., **5**, 345–352.
23. Ausiello,G., Zanzoni,A., Peluso,D., Via,A. and Helmer-Citterich,M. (2005) pdbFun: mass selection and fast comparison of annotated PDB residues. *Nucleic Acids Res*., **33**, W133–W137.
24. Espanel,X., Huguenin-Reggiani,M. and Van Huijsduijnen,R.H. (1998) The SPOT technique as a tool for studying protein tyrosine phosphatase substrate specificities. *Protein Sci*., **11**, 2326–2334.
25. Walchli,S., Espanel,X., Harrenga,A., Rossi,M., Cesareni,G. and van Huijsduijnen,R.H. (2004) Probing protein-tyrosine phosphatase substrate specificity using a phosphotyrosine-containing phage library. *J. Biol. Chem*., **279**, 311–318.
26. Yaffe,M.B. and Smerdon,S.J. (2004) The use of *in vitro* peptide-library screens in the analysis of phosphoserine/threonine-binding domain structure and function. *Annu. Rev. Biophys. Biomol. Struct*., **33**, 225–244.
27. 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.