Proc Indian Natn Sci Acad **81** No. 2 March 2015 pp. 493-507

DOI: 10.16943/ptinsa/2015/v81i2/48102

Research Paper

Getting Phosphorylated: Is it Necessary to be Solvent Accessible ?

NARENDRA KUMAR, NIKHIL PRAKASH DAMLE and DEBASISA MOHANTY*
Bioinformatics Center, National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi 110 067, India

(Received on 25 November 2013; Revised on 17 November 2014; Accepted on 12 December 2014)

The solvent accessibility of the Ser/Thr/Tyr containing peptide stretches in the substrate proteins of various kinases are likely to play a major role in substrate recognition by various kinases. Even though some of the computational tools make use of the solvent accessibility probabilities for prediction of phosphosites, no systematic analysis has been carried to investigate the solvent accessibilities of known phosphorylation sites. In this study, we have systematically analyzed the solvent accessibilities of serine, threonine and tyrosine containing phosphorylation sites in known substrate proteins by mapping them onto PDB structures, and compared with the accessibilities of sites which are not phosphorylated. The average relative solvent accessible area of phosphorylation site residues was found to be significantly more than their non-phosphorylated counterparts. The difference between phospho and non-phospho residues was statistically significant as judged by Wilcoxon test p-values of 2.20×10^{-16}, 5.07×10^{-6} and 2.34×10^{-8} for serine, threonine and tyrosine containing sites respectively. Nonetheless, there are several known phosphorylation sites whose relative accessible surface areas are lower than $10Å^2$, thus these sites are not accessible on the surface of the substrate protein. MD simulations on representative protein structures suggest that, thermal fluctuations at 300K can significantly enhance the accessibilities of such buried Ser/ Thr/ Tyr containing peptides, thus making them available for phosphorylation. Thus our analysis highlights the fact that, it would be in general difficult to set a deterministic criterion based on accessibility values for identifying potential phosphorylation sites.

Key Words: Protein Kinases; Phosphorylation; Surface Accessibility; Molecular Dynamics; Buried Phosphosites

Introduction

Protein kinases play an important role in relaying signals during various signal transduction processes in the cellular machinery by carrying out phosphorylation of specific Ser/Thr or Tyr residues on their cognate substrate proteins. Phosphorylation by kinases alters the functional state of a substrate protein using a number of different mechanisms which include conformational changes in the substrate protein, producing a docking/binding site on the substrate protein for a modular interaction domain, or by altering the existing binding site on the substrate in a manner which can inhibit binding of an interaction partner. Well known example of regulation of functional state of a substrate protein is tumour suppressor protein p53 which contains a number of phosphorylation sites in its transactivation domain (TAD), DNA binding domain (DBD), tetramerization and regulatory domains (Caspari, 2000). Fig. 1 shows the structure of DBD of p53 in complex with a DNA fragment based on the crystal structure 1TUP (Cho *et al*., 1994), where three chains of DBD are bound to the double helical DNA and inset highlights known phosphorylation sites. As can be seen replacement of nonpolar sidechains of Ser/ Thr by highly charged phosphate moiety can significantly alter the binding affinity of p53 DBD for DNA, thus providing a structural basis for

*Author for Correspondence: E-mail: deb@nii.res.in; Tel. : +91-11-26703749

regulation of p53 activity by phosphorylation and dephosphorylation. This example also highlights that, sites on substrate proteins where phosphorylation can affect binding with interaction partners need to be present on the surface of the proteins and are likely to be solvent accessible. Therefore, computational methods for prediction of phosphosites on putative substrate of kinases can take into account solvent accessibility of peptide stretches containing Ser/Thr/Tyr residues.

Various available programs for prediction of phosphorylation sites (Trost and Kusalik, 2011) by a given kinase break the substrate proteins into all possible Ser/Thr/Tyr containing peptides and then evaluate the scores for the phosphorylation of each of these peptides. The high scoring peptides above certain cut off value of the score are reported to be the most likely targets for phosphorylation. In the sequence based prediction methods (Blom *et al*., 2004; Obenauer *et al*., 2003; Xue *et al*., 2008), the score is computed in a probabilistic manner as per the probability of occurrence of various amino acids flanking the Ser/Thr/Tyr. On the other hand, structure based methods (Durek *et al*., 2009; Ellis and Kobe, 2011; Kumar and Mohanty, 2010) calculate the score based on the binding affinity of the each of the Ser / Thr/Tyr containing peptides in complex with the protein kinase in question. All these programs assume that, all the peptides containing the Ser/Thr/Tyr residues are equally accessible to the protein kinases for the phosphorylation reaction. The selectivity of the kinase for a target phosphorylation site is attributed exclusively to the residues flanking the phosphorylatable Ser/Thr/Tyr residues in the substrate protein. These residues, usually three on each side of the phosphorylation site make favourable contacts with the specificity determining residues in the protein kinase.

Although complementarities between the peptide residues and the kinase substrate binding pockets play a crucial role in the phosphorylation event, experimental evidences suggest that it is not the only factor determining the substrate selectivity. Other factors such as the proximity of the substrate to the protein kinase play an equally important role (Ubersax and Ferrell, 2007). This process, called substrate recruitment, is any mechanism, which brings substrate and kinase in close proximity and thus increases the chances of substrate-kinase complex formation. One of the mechanisms of substrate recruitment involves the interaction between docking motifs on the substrate and interaction domain of the kinase (Biondi and Nebreda, 2003). These motifs are located far apart from the phosphorylation site in the substrate protein and increase the affinity of the substrate for the kinase many fold. In some substrates, phosphorylation event increases the affinity of the substrate for the next phosphorylation in the same substrate; this is a recurring theme in the phophorylation by protein kinases. Localization of protein kinase in a specific sub cellular compartment provides a further layer of specificity. Sometimes kinases interact with the substrate through an intermediary of scaffold protein, which acts as a platform for both interacting partners. Such factors have been incorporated in the form of protein-protein interaction information from STRING (Szklarczyk *et al*., 2011) or other such databases has been incorporated for prediction of phosphorylation sites on a given substrate protein (Linding *et al*., 2007; Song *et al*., 2012). Similarly, accessibility of the Ser/Thr/Tyr containing peptide stretch on a substrate protein also plays a major role in the substrate recognition by kinases. Some of the peptides, which otherwise fulfil the requirements of the high affinity to protein kinase, might be buried and hence not accessible for the protein kinase for the transfer of phosphate group. Therefore, in the absence of a conformation change in the substrate upon recruitment, only a subset of peptides, which is spatially located on the surface of the substrate protein, can potentially be phosphorylated (Eisenhaber and Eisenhaber, 2007; Iakoucheva *et al*., 2004; Neuberger *et al*., 2007).

Thus, a number of factors help in maintaining the high level of substrate specificity that is observed in the cellular phosphorylation networks. However, it is very difficult to incorporate all these effects in the prediction programs. In principle, solvent accessibility of the peptides containing potential phosphorylation sites can be calculated; hence, incorporation of the surface accessibility terms in the

prediction algorithms might help in improving their prediction accuracy. However, none of the currently available computation programs for the substrate prediction with the exception of SCANSITE (Obenauer *et al*., 2003), make use of the surface accessibility of the peptides while making predictions.

Large-scale high throughput mass spectrometric experiments have discovered a huge number phosphorylated peptides which are substrates of protein kinases (Olsen *et al*., 2006). Conventional kinase assays and peptide library experiments have also contributed to the known phosphorylation sites of the kinases (Songyang *et al*., 1994). Phospho.ELM (Diella *et al*., 2004) is a database, which catalogues these sites. Although substrate data stored in this database has been used for the development, or benchmarking of a number of prediction programs, no information about the accessibility of these peptides in their substrate protein is available. This is in part because; the crystal structures for most of the substrate proteins have not yet been solved. Phospho3D (Zanzoni *et al*., 2011) is a database, which catalogues and stores the substrate protein whose structure has been solved along with the functional annotations at the phosphorylation site. It also stores the results of local structural alignment of substrates at the phosphorylation site. Although it provides the accessibility of phosphorylation sites, it is only for those few substrates whose structures are available in PDB. In another study, large-scale calculation of predicted values of solvent accessible area of known phosphorylation sites concluded that the phosphorylation sites are mostly accessible on the surface of the substrate protein, and are mostly found in the loop and hinge regions of the proteins (Gnad *et al*., 2007). Although these two studies have shed some light on the surface accessibility of the known site of phosphorylation, one of these has information about a very small number of substrates, while the prediction of accessibilities from sequence alone becomes a limitation of the second study. Therefore, before inclusion of the solvent accessible area terms in the prediction algorithms, it is necessary to carry out an exhaustive analysis of the solvent accessibilities of phosphorylation sites in the known substrate proteins.

In this study, we have attempted to investigate whether inclusion of solvent accessibility probabilities of putative substrate peptides can help in improvement of prediction accuracy. We have calculated the solvent accessibilities of phosphorylation sites in the crystal structures or homology models of known substrate proteins. The accessibilities of the known phosphorylation sites have been compared with the accessibilities of the Ser/Thr/Tyr containing peptides, which are not phosphorylated. Based on this analysis, we have attempted to estimate if statistically significant correlation exists between solvent accessibility and propensity for phosphorylation.

Materials and Methods

Dataset of Protein Kinase Substrates

Substrate proteins of protein kinases catalogued in phospho.ELM (Diella *et al*., 2004) database (version 5.0) were downloaded from UNIPROT database (http://www.uniprot.org). Information about the location of sites of phosphorylation on these proteins was also extracted. This version contains 13563 phosphorylation sites in 4422 protein sequences.

Identification of Structural Homologs

PDB (Berman *et al*., 2000) database 2007 was used for identifying the structural homologs of substrate proteins. This PDB release consisted of 107691 polypeptide chains from 45658 unique structures. Amino acid sequences of these polypeptide chains were downloaded from RCSB website. All PDB sequences were converted to a searchable database using the formatdb program of NCBI blast suite of softwares. For finding the structural homologs of the substrate proteins of various kinases, blast search was carried out against sequences in PDB with an e-value cutoff of 10^{-6}. The most significant BLAST hit for each substrate protein was selected as structural template. If the alignment between the template and the substrate protein showed a gap over the known phosphorylation site, the corresponding substrate sequence was removed from the data set. Fig. 1 shows a flowchart depicting the protocol for mapping of phosphorylation sites onto the PDB structures. The

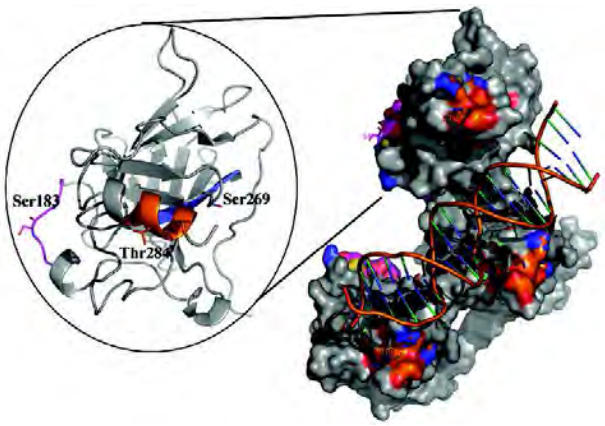

Fig. 1: The structure of DBD of p53 in complex with a DNA fragment based on the crystal structure 1TUP, where three chains of DBD are bound to the double helical DNA. The inset inside the circle highlights known phosphorylation sites of p53 DBD

PDB coordinate files corresponding to the structural templates of the substrate proteins were downloaded from PDB website.

Calculation of Solvent Accessible Area of Phosphorylation Sites

The fragments of the structural templates aligning with the heptameric sequences of the substrate proteins containing the phosphorylation sites were identified. The absolute and relative solvent accessible surface areas of these heptameric peptide fragments in the template structure were calculated using the NACCESS (Hubbard and Thorntan, 1993) program. The solvent accessible surface areas of all other Ser/Thr/Tyr containing heptapeptides, which are not phosphorylated, were also calculated. Apart from the solvent accessible surface areas of the heptameric peptide fragments, the relative and absolute accessible surface areas were also calculated for the central Ser/Thr/Tyr residues.

MD Simulations

Explicit solvent molecular dynamics simulations were carried out on the crystal structure (2C23 (Yang *et al*., 2006)) of human 14-3-3 beta protein using GROMACS4 (Hess, 2008) package and GROMOS96 43a1 force field (Van Gunsteren, 1996). The structure was initially solvated in explicit simple point charge extended (SPCE) water (Berendsen, 1987) in an octahedron box with edges in all directions lying at 10Å from the outer most atom of the protein. Neighbour list was updated every 10 steps with a short-range neighbour cut-off of 10Å. Long range electrostatics interactions were computed using Particle Mesh Ewald (PME) summation method (Darden, 1993; Essmann, 1995) applying cut-off of 10Å. Van der Waal energies were computed using twin range cut-off of 10Å/14Å. The overall charge on the system was neutralized by replacing solvent molecules with counter ions as necessary. All simulations were carried out using periodic boundary conditions in NPT ensemble. Constant temperature of 300K and constant pressure of 1 Bar were maintained using velocity rescale thermostat and Parinello-Rahman barostat (Parrinello, 1981). The isotropic pressure coupling was achieved through time constant 0.5 ps and compressibility of 4.5×10^{-5} bar^{-1}. The bonds and angles in SPCE water were constrained using LINCS algorithm (Hess, 1997). MD simulations were performed using a time step of 2 fs. The total system consisting of protein, solvent molecules and counter ions were subjected to water equilibration keeping the protein heavy atoms restrained for 40ps. Production MD run was carried out for 15ns.

Results

Structural Homologs of Substrate Proteins of Kinases

Fig. 2 shows a flowchart depicting the protocol for mapping of phosphorylation sites onto the PDB structures. Out of 4422 substrate protein sequences catalogued in the phospho.ELM database, structural homologues could be identified for 1425 proteins containing 2860 phosphorylation sites. The structural templates for these 1425 substrate proteins corresponded to 990 PDB structures. A careful examination of the sequence alignments of substrate proteins with the structures in PDB showed that, in some of the alignments, the region corresponding to the phosphorylation site contained a gap in the structure, or the Ser/Thr/Tyr residue was replaced by amino acids which cannot be phosphorylated. In such cases, the sequence of the substrate protein was

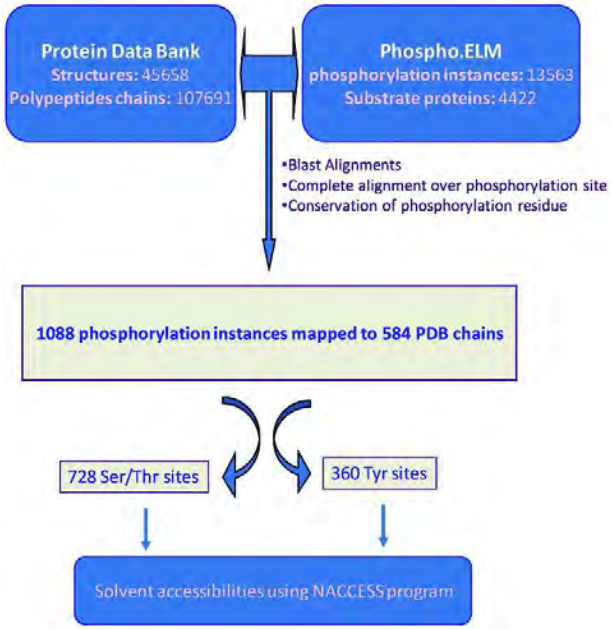

Fig. 2: Flowchart depicting calculation of solvent accessibilities of phosphorylation sites in the substrate proteins based on its crystal structure or structural homolog in PDB

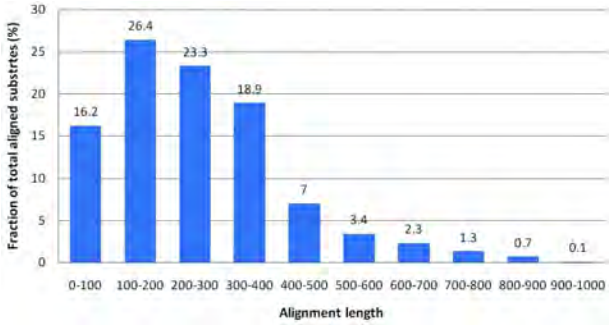

Fig. S1: The distribution of length of sequence alignments between kinase substrates and homologous proteins in PDB

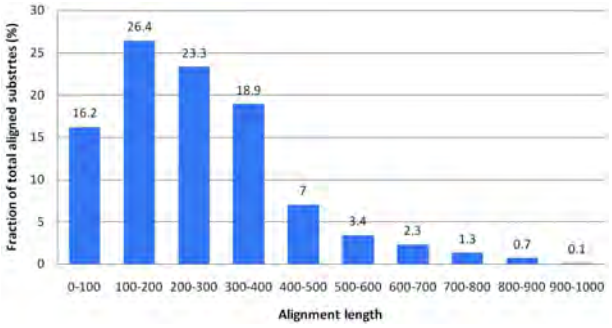

Fig. S2: Distribution of the sequence identities of the substrate proteins (of kinases) which showed significant alignment with proteins in PDB

removed from the data set. After applying this filter, the data set consisted of 1088 phosphorylation instances from 719 substrate proteins, which could be mapped on to 584 polypeptide chains from 571 PDB entries. These included 526, 202 and 360 instances containing serine, threonine and tyrosine as phosphorylation residues respectively. Figs. S1 and S2 show the distribution of the alignment length and percentage identity respectively. As can be seen from Fig. S1, 85.8% of all hits show an alignment over a length of 100 or more amino acids. Similarly, the percentage identities of the hits were also significant as about 44% of hits showed a percentage identity in the range of 90-100%, while 90% of the hits showed identity above 30% (Fig. S2). Thus, the structural templates have good homology with the substrate proteins over a significant length of alignment. Hence, these 719 substrate proteins are likely to have structures similar to those of the templates in PDB. Therefore, the solvent accessibilities of the Ser/Thr/Tyr containing peptide stretches in the substrate proteins can be estimated based on accessibility of the corresponding structural fragments in the identified PDB hits.

Solvent Accessibility of Phosphorylation Sites

Solvent accessible surface areas of the phosphorylation sites on the substrate proteins were calculated from their crystal structures when available, or from their structural homologs. The use of homologs for surface area calculation is appropriate as structures are known to be more conserved than sequences and the homologues have been identified based on significant sequence similarity. In addition, given the biological importance of the phosphorylation event, the phosphorylation sites are also likely to be conserved. Therefore, the probability of the finding the phosphorylation site on a substrate protein in same structural context as in the structural match is very high, and the surface areas calculated from it are likely to be true representative of the surface area in the actual substrate proteins. Seven amino acids long peptide stretches with phosphorylation site as the central residue were selected for the calculation of solvent accessible surface areas by NACCESS

computer program (Hubbard and Thorntan, 1993). For each substrate protein, apart from the phosphorylation site, the accessible surface areas of the all other Ser/Thr/Tyr containing potential sites, which are not phosphorylated, were also calculated. In every case, absolute as well as relative surface areas were calculated. The relative surface area values represent the surface accessibility in comparison to the accessibility of a residue in extended conformation when it is surrounded by alanine residue on both sides. Fig. S3 shows an example of the analysis of solvent accessibility of a representative query substrate protein 3-phosphoinisitide dependent protein kinase-1 whose solvent accessible area of phosphorylation site is calculated from the structural homolog with PDB ID 2BCJ. Fig. 3 shows the average solvent accessible surface areas for the central Ser/Thr/Tyr residues as well as the heptameric peptide fragment corresponding to sites, which are known to be phosphorylated, as well as for those, which are not phosphorylated. As can be seen, in general the average solvent accessible surface area of the phosphorylation site residues is higher compared to the accessibilities of Ser/Thr/Tyr containing peptides that are not phosphorylated. The statistical significance of the difference between the accessibilities of phosphorylation site residues and non-phosphorylated residues was judged by Wilcoxon rank sum test with continuity correction. The difference in average relative solvent accessible area between phospho and non-phospho residues was statistically significant as judged by p-values of 2.20×10^{-16}, 5.07×10^{-6} and 2.34×10^{-8} for Ser, Thr and Tyr residues respectively. Similarly, p-values for the difference between the average relative solvent accessible area of phospho-peptide and non-phospho peptides were 2.20×10^{-16}, 2.05×10^{-13} and 4.77×10^{-12} for Ser, Thr and Tyr containing peptides respectively. Fig. 3 also shows that the difference between the average solvent accessibility values for the absolute area was significant as judged by Wilcoxon test p-values. Thus, based on average accessibility values the phosphorylation site residues/peptides were found to be more exposed to the solvent. Fig. 4 shows a comparison of the distribution of absolute accessibilities of phosphorylation site containing peptides with their counterparts which are not phosphorylated. As can be seen from Fig. 4, at high accessibility values (> 400 Å^2) the percentage of phosphorylated peptides are higher than the percentage of peptides which are not phosphorylated, while the trend is reversed in the low accessibility range. However, a surprising observation from this

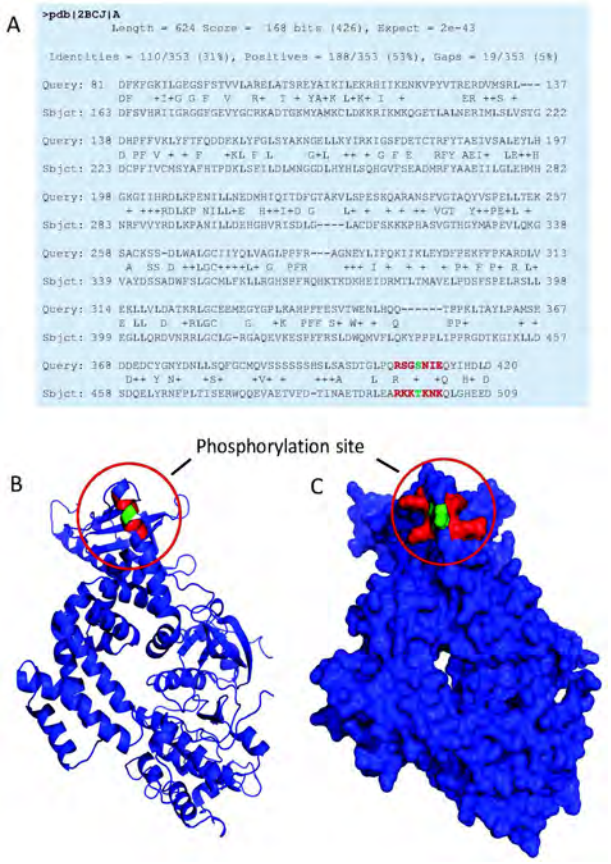

Fig. S3: An example of the analysis of solvent accessibility of a heptapeptide containing the site of phosphorylation in substrate protein of a kinase. Solvent accessibility was calculated by NACCESS program from the homologous protein identified by BLAST search in PDB. (A) Sequence alignment of query substrate 3-phosphoinisidite dependent protein kinase-1 (O15530) with the sequence of structural homologue guanine nucleotide binding protein: GPCR kinase 2 (PDB identifier: 2BCJ). The phosphorylation site residue has been shown in green and three residues on each side are shown in red. (B) Cartoon representation of 2BCJ. Phosphorylation site and the residues surrounding it are shown in green and red respectively. (C) Surface representation of 2BCJ. The peptide stretch corresponding the phosphorylation site and surrounding residues have been shown in the green and red.

Table 1: The solvent accessible surface areas of phosphorylation sites*

A	B	C	D	E	F	G	H	I	J
Catalase	P04040	230	eavYckf	1QQW	169.6	14.5	6.8	30.2	Abl, Abl2
Phosphoglycerate mutase	P18669	22	nrfSgwy	1YJX	185.2	14.7	12.6	16.3	PAK1
Proto-oncogene tyrosine	P12931	215	ggfYits	2H8H	253.4	44.0	20.7	31.3	Src
Serine/threonine-protein kinase 6	O14965	266	llgSage	2J4Z	282.9	88.4	11.9	34.2	Aurora A
Serine/threonine-protein kinase 6	O14965	278	fgwSvha	2J4Z	305.5	41.2	6.2	34.2	Aurora A
Glucose-6-phosphate isomerase	P06744	184	wyvSnid	1NUH	323.1	13.8	18	30.3	CK2 group
Casein kinase II subunit alpha	P68400	255	edlYdyi	1JWH	374.9	79.5	37.4	30.6	Lyn, Fgr
Dihydropteridine reductase	P09417	223	nrpSsgs	1HDR	381.3	86.0	75.9	27.5	CaM-KII group
E3 ubiquitin-protein ligase CBL	P22681	371	yelYcem	1FBV	393.7	13.2	40.9	29.4	EGFR, InsR
Retinoic acid receptor RXR-alpha	P19793	249	tetYvea	1LBD	402.5	32.9	34.9	41.5	MAP2K4
Phosphoglycerate mutase1	P18669	117	wrrSydv	1YJX	418.0	20.9	73.9	16.3	PAK1
C-Rel proto-oncogene protein	P16236	266	rrpSdqe	1GJI	419.4	86.0	73.8	30.6	PKA group
14-3-3 protein eta	Q04917	58	rrsSwrv	2C74	445.7	47.6	40.7	31.2	SDK1
Syntaxin-1A	Q16623	188	ssiSkqa	1DN1	487.4	13.6	11.7	38.0	DAPK group
Band 3 anion transport protein	P02730	303	maqSrge	1HYN	492.8	48.5	41.6	29.3	CK1 alpha
Retinoic acid receptor RXR-alpha	P19793	260	npsSpnd	1LBD	500.6	56.0	25.8	41.5	MAPK1, MAPK3, MAPK group
Interferon regulatory factor 3	Q14653	385	ggaSsle	1QWT	507.4	46.9	59.2	39.0	IKK-epsilon, TBK1
Interferon regulatory factor 3	Q14653	386	gasSlen	1QWT	518.1	97.6	30.3	39.0	IKK group
Serine/threonine-protein kinase 6	O14965	391	skpSncq	2J4Z	524.5	48.8	41.9	34.2	Aurora A
Catalase	P04040	385	vanYqrd	1QQW	526.9	54.9	66	30.2	Abl, Abl2
Serine/threonine-protein kinase 6	O14965	226	qklSkfd	2J4Z	574.7	69.0	48	34.2	Aurora A
Serine/threonine-protein kinase 6	O14965	287	srrTtlc	2J4Z	581.1	91.9	40.3	34.2	Aurora A
Tyrosine-protein kinase BTK	Q06187	550	ddeYtss	1K2P	615.6	52.1	83.8	36.7	Lyn, BTK
Proteasome subunit alpha type 3	P25788	242	akeSlke	1IRU	623.6	35.3	24.5	36.3	CK2 group
Eukaryotic translation initiation factor 4E	P07260	15	envSvdd	1AP8	748.6	113.3	97.2	43.0	CK2 group

*Solvent accessible surface areas in protein kinase substrates, which show 100%, match over more than 200 alignment length with a PDB structure. The absolute solvent accessible area of heptapeptide containing phosphorylation site and absolute and relative areas of phosphorylation site Ser/Thr/Tyr residues are calculated by NACCESS. (A) Substrate protein containing the phosphorylation site for protein kinase in column J. (B) Swissprot/TrEmbl code for the substrate protein. (C) Residue number of phosphorylation site Ser/Thr/Tyr in the protein sequence of substrate. (D) Heptapeptide sequence containing the phosphorylation site (Upper case). (E) PDB id of structural match of substrate protein. (F) Absolute solvent accessible area of heptapeptide (in column D) as calculated by NACCESS program. (G) Absolute solvent accessible area of phosphorylation site Ser/Thr/Tyr residue (Upper case in column D). (H) Relative solvent accessible area of phosphorylation site Ser/Thr/Tyr residue. (I) Average relative solvent accessible area of all the Ser/Thr/Tyr residues in the structural match. (J) Protein kinase responsible for phosphorylation of the substrate at the given site.

analysis is the occurrence of a significant number of phosphorylated peptides even at accessibility values below 200 Å^2.

Since the results shown in Fig. 4 is based on analysis of accessibility values in crystal structures homologous to various kinase substrates, we decided to analyze separately the accessibilities of phosphorylation sites in substrate proteins for which crystal structures were available. Some of the protein kinase substrates in our data set had structural hits from PDB with 100% match over alignment length of more than 200 amino acids. Table 1 lists 25 such proteins that showed 100% match with PDB entries. In 19 of these proteins, the peptides harbouring the phosphorylation site had accessibility values higher than 350 Å^2. However, for the remaining six proteins the known phosphorylation sites were less accessible

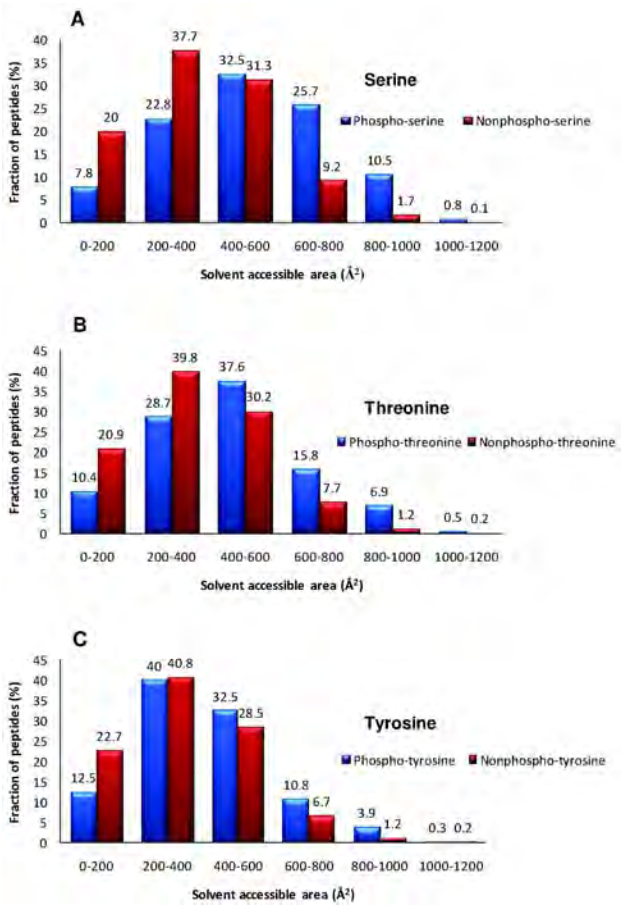

Fig. 4: Distribution of solvent accessible surface areas of phosphorylation site containing peptides as compared to nonphospho-site containing peptides, when phosphorylation site is serine (A), threonine (B) and tyrosine (C)

than the other potential sites. This suggests that observation of known phosphorylation sites at low accessibility values is not an artefact arising from structural homologs included in our analysis.

A large number of phosphorylation sites included in our dataset had been identified by high throughput mass spectroscopic experiments. The protein kinases that phosphorylate these sites are not known and there is a possibility that some of these sites are false positives arising because of the uncertainty in correctly assigning the sequence of the phosphorylation site from the mass spectrometry data. For these reasons, we reanalyzed the average solvent accessible surface area and the distribution of the solvent accessible surface area after removing the phosphorylation sites identified by high throughput studies from the dataset. After removing such site, we were left with 223, 90 and 119 instances containing Ser, Thr and Tyr residues respectively. Fig. 5 shows the average relative and average absolute solvent accessible surface area of phosphorylation site residues and the 7mer peptide harboring phosphorylation site as compared to their nonphosporylated Ser, Thr and Tyr containing

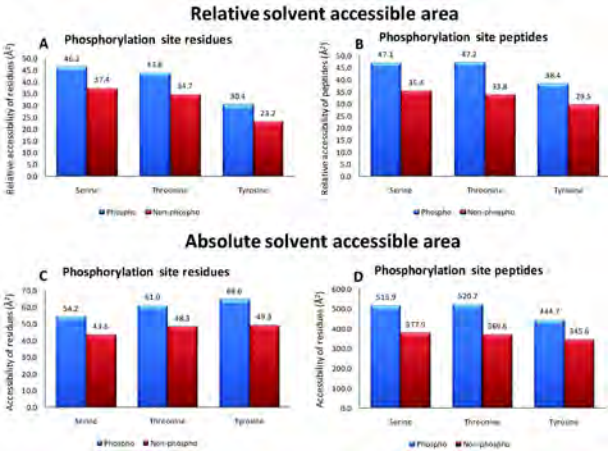

Fig. 5: The average solvent accessible surface area values of known phosphorylation sites, without including sites from high throughput mass spectrometry experiments, as compared to non-phosphorylation sites. (A) Relative solvent accessible surface area of Ser/Thr/Tyr residues. (B) Relative solvent accessible area of heptapeptides containing Ser /Thr /Tyr as the central residue. (C) Absolute solvent accessible surface area of Ser/Thr/Tyr residues. (D) Absolute solvent accessible area of heptapeptides containing Ser/Thr/Tyr as the central residues

counterparts when instances identified by mass spectrometry were not considered. It can be seen that the average surface area values in every case remain more or less same as when such site were considered (Fig. 3). Fig. 6 shows the comparison of distribution of absolute solvent accessible area values of phosphorylation sites containing peptides with their counterparts which are not phosphorylated, when the phosphorylation instances identified by mass spectrometry were not included. It can be seen clearly from Fig. 6, at low accessibility value (<400Å^2) the percentage of phosphorylated peptides has reduced as compared to the values when sites by mass spectrometry were included in the analysis (Fig. 4). It may be noted that, earlier study (Jimenez *et al.*, 2007) on structural analysis of protein phosphoryaltion sites had also reported human and mouse phosphorylation sites at low accessibility.

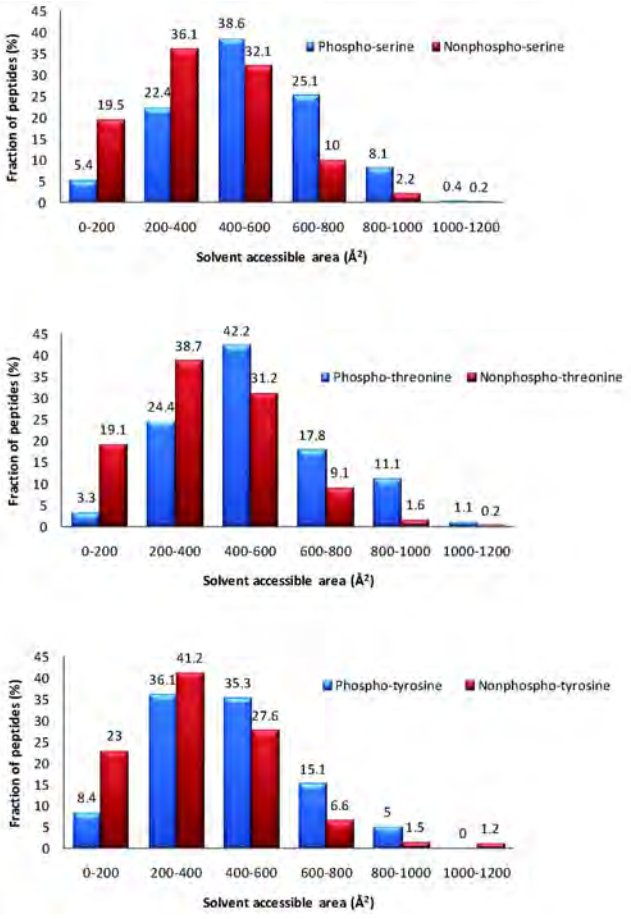

Fig. 6: Distribution of solvent accessible surface area of phosphorylation site containing peptides, without including the sites identified from high throughput mass spectrometric experiments, as compared to Ser/Thr/Tyr containing peptides which are not phosphorylated

We analyzed in detail the substrate proteins having known phosphorylation sites at low accessibility. Table 2 lists substrate proteins with phosphorylation sites whose relative surface areas are less that 10 Å^2. Many of these sites are buried in the respective proteins and phosphorylation of such sites might require conformational shift in the protein, before they can be accessed and phosphorylated by the protein kinases. For example, phosphorylation site containing peptide RYL**S**EVA in signalling protein, human 14-3-3 beta, has been found to be buried as judged from its very low solvent accessible area (0.8Å^2) from crystal structure 2C23 (Fig. 7). It would be interesting to analyze the phosphorylation mechanism of such proteins. It would also be interesting to see if the binding of such substrates to protein kinases lead to the unfolding and binding of phosphorylation site, or interaction with other cellular factors unfolds the proteins and makes the phosphorylation site accessible to protein kinases.

As different protein kinases are known to be differentially regulated with each one having a unique mechanism of regulation, we investigated if certain specific kinase families are responsible for phosphorylation of these sites, which are buried largely. Out of 1088 phosphorylation sites mapped to crystal structures, 628 sites did not have any information about the kinase responsible for their phosphorylation. The remaining 460 sites are phosphorylated by 134 kinase types as reported in phospho.ELM database. Among these mapped phosphorylation sites in the crystal structure templates, we calculated the average accessible surface area for various kinase types. However, sufficient number of phosphorylation instances was not mapped for most of the protein kinases. Therefore, we grouped together the phosphorylation instances assigned to similar kinases belonging to a specific group (Fig. S4). We selected those groups, which contained more than five phosphorylation instances. For this study, we considered only Ser/Thr kinases. 237 phosphorylation instances could be grouped into

Table 2: The substrate protein of protein kinases with relative surface accessible area of phosphorylation site less than 10Å^{2*}

A	B	C	D	E	F	G	H
P08575	1216	feqYqfl	1YGU	603	93	0.0	Csk
P10275	791	rhlSqef	2PNU	250	98	0.0	PKB group
P16452	247	wigSvdi	2Q3Z	693	49	0.0	PKA group
P39748	187	tfgSpvl	1UL1	355	98	0.0	CDK1
P51617	209	knvTnnf	2OID	326	48	0.0	IRAK1
Q63270	138	radSlqk	2B3Y	887	97	0.0	PKC group
Q9UEW8	373	plhTsrv	2OZA	360	42	0.0	Wnk1
P18031	153	ktyYtvr	1L8G	297	100	0.1	InsR
P68369	432	ekdYeev	1SA1	438	97	0.1	Syk
P17655	369	rgsTagg	1KFX	699	91	0.2	PKA group
P18031	66	dndYina	1L8G	297	100	0.3	EGFR, InsR
P31152	196	wyrSprl	2I6L	296	80	0.4	MAPK4
Q9P286	573	ksdSill	2F57	297	98	0.4	PAK5
P22001	135	irfYelg	2R9R	389	83	0.5	Src
P04167	128	rrfSlat	2Q6N	464	89	0.6	PKA group
P35270	213	retSvdp	1Z6Z	257	97	0.6	CCDPK
P18206	1100	nlqSvke	1TR2	1129	85	0.8	PKC alpha
P31946	132	rylSeva	2C23	230	97	0.8	PKC zeta
P43403	474	nrhYaki	2OZO	605	88	0.8	Lck
P11799	1748	griSnys	1KOA	430	57	1.0	CaM-KII group
Q9UQM7	286	rqeTvec	2V7O	297	94	1.2	CaM-KII alpha
P29322	839	erpYwem	2QOK	278	84	1.4	EphA8
Q63270	711	sygSrrg	2B3Y	887	97	1.9	PKC group
P61763	158	sfySphk	1DN1	589	94	3.1	CDK group
O43293	225	kqeTltn	1YRP	276	99	3.5	DAPK3
P05532	934	revSfyy	1P4O	380	45	3.8	Fyn
Q00169	164	kfkSikt	1UW5	264	97	3.8	PKC alpha
Q29502	197	nveSlld	2H9V	326	47	3.8	PAK2
P18031	50	rdvSpfd	1L8G	297	100	3.9	PKB group CLK1
Q9UM73	1278	rdiYetd	1P4O	295	63	3.9	ALK
P68104	432	mrqTvav	2B7C	439	89	4.0	PKC group

Contd ...

Table 2 contd ...

O96017	516	lhtSrvl	2OZA	319	51	4.1	CHK2
P13569	660	rrnSilt	1XMI	283	93	4.6	PKA group PKG/cGK group
P37173	424	trrYmap	2QLU	299	65	4.8	TGFbR2
P18031	152	iktYytv	1L8G	297	100	5.0	InsR
P22681	371	yelYcem	1FBV	388	100	6.2	EGFR, InsR
O00571	322	lvaTpgr	2I4I	413	98	6.3	CDK1
P55211	153	dlaYils	1JXQ	269	79	6.6	Abl
P04040	230	eavYckf	1QQW	499	100	6.8	Abl, Abl2
Q14934	676	rkrSqpq	2AS5	287	78	7.1	MAPK1, MAPK3, RSK-2
Q07497	601	lveYlkl	2OZO	306	53	7.6	EphB5
P35241	564	kykTlre	2I1K	581	73	7.7	ROCK group
P35241	573	kgnTkrr	2I1K	581	73	7.9	ROCK group
Q13882	342	dneYtar	2H8H	445	61	8.0	Brk
P17655	368	rrgStag	1KFX	699	91	8.4	PKA group
P61763	574	hilTpqk	1DN1	589	94	8.8	CDK group
P48025	130	vrdYvrq	2OZO	252	76	9.0	Lyn, Syk
P53355	289	rreSvvn	1WMK	311	90	9.4	RSK-2, RSK group
P22607	724	nelYmmm	2PSQ	299	91	9.8	FGFR3
P41743	271	driYamk	1ZRZ	340	90	9.8	Src

*All these substrates show a significant alignment over 200 amino acids with the sequence of a match from the PDB database. (A) Swissprot/TrEmbl code for substrate protein containing the phophorylation site for protein kinase in column H. (B) Residue number of phosphorylation site Ser/Thr/Tyr in the protein sequence of substrate. (C) Heptapeptide sequence containing the phosphorylation site (Upper case). (D) PDB id of structural match of substrate protein. (E) Alignment length of the substrate with the structural match (F) The percentage identity of substrate protein with structural match over the alignment length. (G) Relative solvent accessible area of phosphorylation site Ser/Thr/Tyr residue (shown in Upper case in column C) as calculated by NACCESS program. (H) Protein kinase responsible for phosphorylation of the substrate at the given site

13 Ser/Thr kinase groups containing 56 individual kinases. Fig. 8 shows the average accessibilities of phosphorylation sites for these kinase groups. CDK1, DAPK and PKB are three groups whose substrates have lowest average surface accessible areas among all kinases in our study. Based on this, it is tempting to speculate that, these kinase groups might employ a mechanism involving slight unfolding/ conformational shift in their substrate proteins, thus favouring induced fit mechanism for phosphorylation. On the other hand, phosphorylation sites in substrates of some kinases like CK1, GSK and PLK show a very high average accessible area, suggesting that the substrates of these kinases can be easily accessed by the protein kinase catalytic site.

We also explored the possibility that the substrates wherein the phosphorylation sites are largely buried inside the protein core, could get exposed to solvent and serve as targets of phosphorylation during thermal fluctuations to which cellular proteins are constantly subjected to. For this

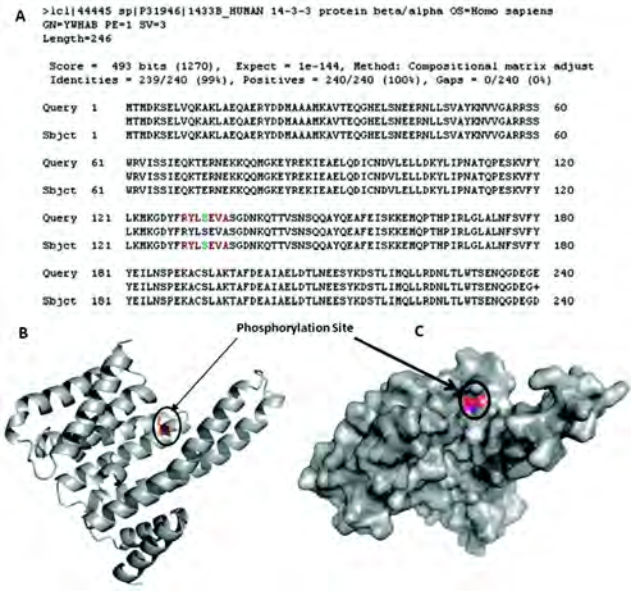

Fig. 7: An example of substrate protein human 14-3-3 beta containing a phosphorylation site with very low solvent accessible area (0.8Å^2). The crystal structure of human 14-3-3 beta has been identified from the PDB by blast alignments. (A) Alignment of human 14-3-3 beta (P31946) with the sequence of PDB entry 2C23. (B) Cartoon representation of 2C23. The Phosphorylation site has been encircled. (C) Surface representation of 2C23 showing the phosphorylation site encircled

purpose, we carried out explicit solvent molecular dynamics simulations on 2C23 (human 14-3-3 beta protein) for 15ns using GROMACS suite. Fig. 9 indicates that, the buried phosphorylation site mentioned above (RYL**S**EVA) in fact gets exposed to solvent intermittently over a period of dynamics. This could, in general, be true for many such buried phosphorylation sites, which could only be studied by subjecting each one of the substrates to thermal fluctuations.

Discussion

Most phosphorylation site prediction methods take into consideration only the sequence or structure of a short peptide stretch flanking the Ser/Thr/Tyr amino acid. However, in reality the protein kinases phosphorylate a folded and structurally intact substrate protein in its functional form. Therefore, solvent accessibility of the phosphorylation site and its flanking residues is a major factor, which is expected to affect the phosphorylation of a substrate protein. In this work, we have analyzed the solvent accessible areas of known phosphorylation sites in various substrate proteins of different kinases. Calculations of solvent accessible surface areas of phosphorylation sites have been carried out in the crystal structures of the substrate protein or their structural homologues. The results of our analysis indicate that, the phosphorylation site residues are significantly more exposed to the solvent as compared to the other sites containing Ser/Thr/Tyr residues. Our results are in agreement with earlier studies which have also suggested that phosphorylation sites are mostly found in regions of proteins that are likely to adopt loop or coil secondary structural states and prefer to be exposed to the solvent (Gnad *et al*., 2007). Our current results based on the actual accessibility values from crystal structures reaffirm the same.

Although, in general, the phosphorylation sites are more exposed than their counterparts which are not phosphorylated, our analysis has revealed few interesting examples of substrate proteins on which the phosphorylation sites have a very low solvent accessible surface area indicating that they are buried inside. Such sites cannot be easily phosphorylated by the protein kinases in the absence of a considerable conformational change in the structure of the substrate proteins. Molecular dynamics simulations on one of the substrate proteins harbouring a buried phosphorylation site indicated that, such buried phosphorylation sites can be exposed intermittently during thermal fluctuations. Thus they might get phosphorylated by their cognate protein kinases. It will be interesting to analyze more number of such proteins individually to find out mechanistic details about how they are phosphorylated. Previous studies (Jimenez *et al*., 2007) have also reported human and mouse phosphosites at low accessibilities based on analysis of both high throughput as well as low throughput phosphorylation data and have individually analyzed several proteins containing such buried phosphosites. On the contrary, we have systematically analysed low throughput phosphorylation site data in Phospho.ELM. This extensive analysis indicates that the actual number of phosphorylation instances at such buried sites is lower, when high throughput data is excluded.

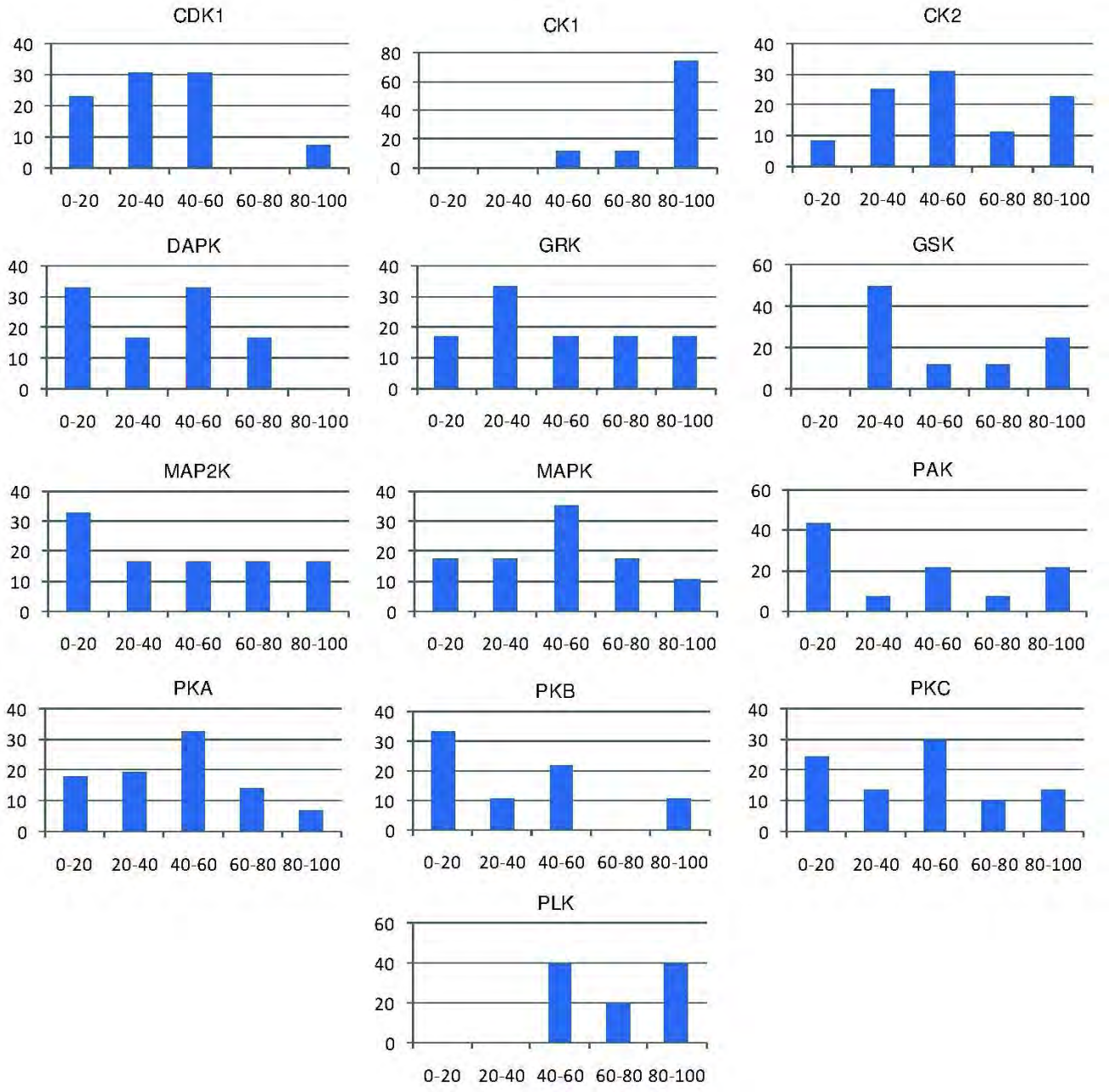

Fig. S4: Distribution of relative solvent accessible surface areas of phosphorylation sites of various protein kinases as calculated by NACCESS program from the crystal structures of homologous proteins. On each panel, x-axis has relative accessible area, and y-axis has the number of substrates for the protein kinase

Additionally we have also investigated if specific kinase families preferentially phosphorylate these largely buried sites.

The results from analysis of accessibilities of phosphorylation sites suggest that, the inclusion of the solvent accessibility terms in the phosphorylation site prediction programs might help in improving their prediction accuracy. However, our observation of several phosphorylation sites at very low accessibility values suggests that it would be difficult to fix a deterministic criteria based on accessibility value for identifying potential phosphorylation sites. Accessibility and structural flexibility of the potential phosphorylation sites can probably be combined with interactions between phosphorylation site and kinase

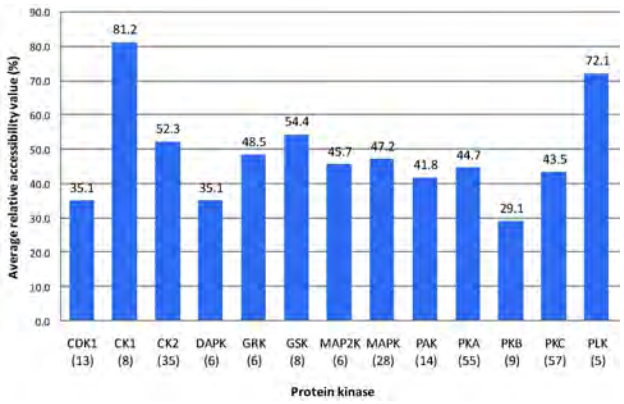

Fig. 8: Average relative solvent accessible surface accessible areas for the various protein kinase groups. Only those kinases groups have been shown which had more than five (values in brackets below the kinase group name) phosphorylation instances mapped to structural templates

to develop more powerful structure based methods for prediction of substrates for kinases.

Acknowledgements

Authors thank Director, NII for encouragement and support. NPD thanks Department of Biotechnology (DBT), India for award of BINC senior research fellowship. The work has been supported by grants to NII from DBT, India. DM also acknowledges financial support from DBT, India under BTIS project and National Bioscience Career Development award.

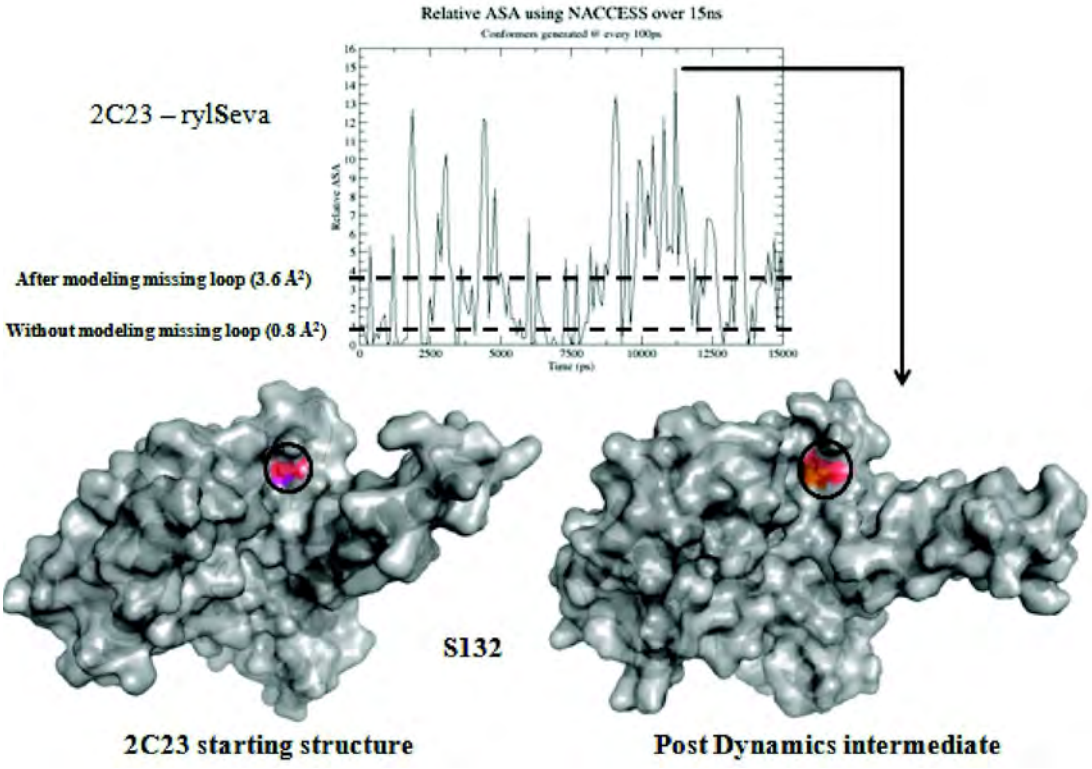

Fig. 9: Intermittent exposure over 15ns of phosphosite S132 in the heptameric strech rylSeva on human 14-3-3 beta protein shown. Lower panels indicate surface representation of the same structure 2C23 before and after dynamics (at the arrow-indicated time point) with phosphosite encircled

References

Berendsen H J C, Grigera J R and Straatsma T P (1987) The missing term in effective pair potentials *J Phys Chem* **91** 6269-6271

Berman H M *et al.* (2000) The Protein Data Bank, *Nucleic Acids Res* **28** 235-242

Biondi R M and Nebreda A R (2003) Signalling specificity of Ser/Thr protein kinases through docking-site-mediated interactions *Biochem J* **372** 1-13

Blom N *et al.* (2004) Prediction of post-translational glycosylation

and phosphorylation of proteins from the amino acid sequence *Proteomics* **4** 1633-1649

Caspari T (2000) How to activate p53, *Curr Biol* **10** R315-317

Cho Y *et al.* (1994) Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations *Science* **265** 346-355

Darden T, York D and Pedersen L (1993) Particle mesh Ewald: An N-log(N) method for Ewald sums in large systems *J Chem Phys* **98** 10089-10092

Diella F *et al.* (2004) Phospho.ELM: a database of experimentally verified phosphorylation sites in eukaryotic proteins *BMC bioinformatics* **5** 79

Durek P *et al.* (2009) Detection and characterization of 3D-signature phosphorylation site motifs and their contribution towards improved phosphorylation site prediction in proteins *BMC bioinformatics* **10** 117

Eisenhaber B and Eisenhaber F (2007) Posttranslational modifications and subcellular localization signals: indicators of sequence regions without inherent 3D structure? *Current protein & peptide science* **8** 197-203

Ellis J J and Kobe B (2011) Predicting protein kinase specificity: Predikin update and performance in the DREAM4 challenge *PloS one* **6** e21169

Essmann U, Perera L, Berkowitz M L, Darden T, Lee H and Pedersen L G (1995) A smooth particle mesh ewald potential *J Chem Phys* **103** 8577-8592

Gnad F *et al.* (2007) PHOSIDA (phosphorylation site database): management, structural and evolutionary investigation, and prediction of phosphosites *Genome Biol* **8** R250

Hess B (2008) GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation *J Chem Theory Comput* **4** 435-447

Hess B, Bekker H, Berendsen H J C and Fraaije J G E M (1997) LINCS: A linear constraint solver for molecular simulations *J Comp Chem* **18** 1463-1472

Hubbard S and Thorntan J M (1993) NACCESS Computer Program

Iakoucheva L M *et al.* (2004) The importance of intrinsic disorder for protein phosphorylation *Nucleic Acids Res* **32** 1037-1049

Jimenez J L *et al.* (2007) A systematic comparative and structural analysis of protein phosphorylation sites based on the mtcPTM database *Genome Biol* **8** R90

Kumar N and Mohanty D (2010) Identification of substrates for Ser/Thr kinases using residue-based statistical pair potentials *Bioinformatics* **26** 189-197

Linding R *et al.* (2007) Systematic discovery of in vivo phosphorylation networks *Cell* **129** 1415-1426

Neuberger G, Schneider G and Eisenhaber F (2007) pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model *Biology direct* **2** 1

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

Olsen J V *et al.* (2006) Global, in vivo, and site-specific phosphorylation dynamics in signaling networks *Cell* **127** 635-648

Parrinello M and Rahman A (1981) Polymorphic transitions in single crystals: A new molecular dynamics method *J Appl Phys* **52** 7182-7190

Song C *et al.* (2012) Systematic analysis of protein phosphorylation networks from phosphoproteomic data *Molecular & cellular proteomics : MCP* **11** 1070-1083

Songyang Z *et al.* (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases *Curr Biol* **4** 973-982

Szklarczyk D *et al.* (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored *Nucleic Acids Res* **39** D561-568

Trost B and Kusalik A (2011) Computational prediction of eukaryotic phosphorylation sites *Bioinformatics* **27** 2927-2935

Ubersax J A and Ferrell J E Jr (2007) Mechanisms of specificity in protein phosphorylation *Nat Rev Mol Cell Biol* **8** 530-541

van Gunsteren W F, Billeter S R, Eising A A, Hünenberger P H, Krüger P, Mark A E, Scott W R P and Tironi I G (1996) Biomolecular Simulation: The GROMOS96 manual and user guide

Xue Y *et al.* (2008) GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy *Molecular & cellular proteomics : MCP* **7** 1598-1608

Yang X *et al.* (2006) Structural basis for protein-protein interactions in the 14-3-3 protein family *Proceedings of the National Academy of Sciences of the United States of America* **103** 17237-17242

Zanzoni A *et al.* (2011) Phospho3D 2.0: an enhanced database of three-dimensional structures of phosphorylation sites, *Nucleic Acids Res* **39** D268-271.