BIOINFORMATICS ORIGINAL PAPER

Vol. 26 no. 2 2010, pages 189–197
doi:10.1093/bioinformatics/btp633

Structural bioinformatics

Identification of substrates for Ser/Thr kinases using residue-based statistical pair potentials

Narendra Kumar and Debasisa Mohanty*

National Institute of Immunology, Aruna Asaf Ali Marg, New Delhi 110067, India

Received on May 6, 2009; revised on November 5, 2009; accepted on November 7, 2009

Advance Access publication November 12, 2009

Associate Editor: Anna Tramontano

ABSTRACT

Motivation: *In silico* methods are being widely used for identifying substrates for various kinases and deciphering cell signaling networks. However, most of the available phosphorylation site prediction methods use motifs or profiles derived from a known data set of kinase substrates and hence, their applicability is limited to only those kinase families for which experimental substrate data is available. This prompted us to develop a novel multi-scale structure-based approach which does not require training using experimental substrate data.

Results: In this work, for the first time, we have used residue-based statistical pair potentials for scoring the binding energy of various substrate peptides in complex with kinases. Extensive benchmarking on Phospho.ELM data set indicate that our method outperforms other structure-based methods and has a prediction accuracy comparable to available sequence-based methods. We also demonstrate that the rank of the true substrate can be further improved, if the high-scoring candidate substrates that are short-listed based on pair potential score, are modeled using all atom forcefield and MM/PBSA approach.

Contact: deb@nii.res.in

Supplementary information: Supplementary data are available at *Bioinformatics* Online.

1 INTRODUCTION

Post-translational modification of proteins by phosphorylation is a major regulatory mechanism for a variety of cellular processes (Cohen, 2000; Ubersax and Ferrell, 2007). Therefore, the correct identification of the substrate protein that a kinase is likely to phosphorylate, is crucial for understanding the molecular details of various cellular and disease processes. With the availability of the sequences of complete genomes, bioinformatics analyses (Caenepeel *et al*., 2004; Krupa and Srinivasan, 2002; Manning *et al*., 2002; Plowman *et al*., 1999) have classified the large number of kinases in different organisms into various kinase groups and families, thus giving valuable clues about putative signaling pathways in which they could possibly be involved. However, deciphering the specific substrate proteins of these large number of kinases and reconstruction of phosphorylation networks at a genomic scale (Linding *et al*., 2007) still remains a major challenge. Therefore, it is necessary to develop powerful computational methods for predicting the substrate proteins for a given kinase and precisely identifying the sites of phosphorylation.

The various prediction programs available for the identification of substrates of kinases can be broadly classified into two major groups, namely, the sequence-based and the structure-based methods. All sequence-based methods like NetPhosK (Blom *et al*., 2004), SCANSITE (Obenauer *et al*., 2003), GPS (Zhou *et al*., 2004) and PPSP (Xue *et al*., 2006) are trained on sequence information derived from experimentally characterized phosphorylation sites (Amanchy *et al*., 2007; Songyang *et al*., 1994) for various kinases, thus limiting their applicability to only those kinase families for which sufficient amount of substrate information is available from experimental studies. Hence, they often fail to predict substrates for the other protein kinases for which little or no substrate information is available. In contrast to sequence-based methods, structure-based methods attempt to predict the substrate peptides for kinases based on structural modeling of the putative Ser/Thr/Tyr containing peptides in the peptide binding pocket of the kinase and ranking various peptide ligands as per their interaction energy with the receptor kinase. Thus, the structure-based methods do not require information about known substrates for a given kinase, as the preferred substrates are predicted based on physico-chemical interactions between the kinase and peptide. Therefore, they can in principle, be applied for predicting substrates for novel kinase families for which no experimental information is available. However, the survey of available literature indicates that, in contrast to a large number of sequence-based methods for predicting substrates for kinases, as of now, PREDIKIN (Brinkworth *et al*., 2003) is the only structure-based program available for prediction of phosphorylation sites in proteins. Even though PREDIKIN was a major development in demonstrating for the first time, the utility of structural information in successful prediction of kinase substrates, it has not been extensively benchmarked on various kinase families. Secondly, the scoring scheme used in PREDIKIN for estimating the binding energy of the substrate peptides is based on information from peptide library data. In view of its reliance on experimental substrate peptide data, PREDIKIN also share some of the disadvantages of sequence-based methods so far as its applicability to new kinase families is concerned. This prompted us to explore the feasibility of developing alternative structure-based approach which does not use any experimental information about known kinase substrates. In fact, threading type approach involving statistical pair potentials has been successfully used for predicting binder peptides for MHC proteins (Altuvia *et al*., 1995; Schueler-Furman *et al*., 1998, 2000).

*To whom correspondence should be addressed.

Since, the recognition of substrate peptides by protein kinases is analogous to the binding of peptides by MHC proteins, we wanted to investigate whether residue-based statistical potentials can be used for predicting substrate peptides for various kinases.

In this work, we have developed a novel multi-scale structure-based approach, where putative high-scoring substrate peptide candidates are identified by threading of peptides on structural templates of kinase–peptide complexes and scoring them by residue-based statistical pair potentials. High-scorer peptides, short listed by initial screening are modeled in the peptide binding pocket of the kinase using rotamer library approach and ranked as per their MM/PBSA binding free energy values. Benchmarking of our prediction method on the available experimental data and extensive comparison with other kinase substrate prediction programs, indicate that the prediction accuracy of our method is comparable to other sequence-based methods, even though it does not use any kinase specific experimental data.

2 METHODS

2.1 Modeling of peptides in the active site pocket of kinases and scoring the binding energy by statistical pair potentials

The crystal structures of kinase–peptide complexes belonging to Ser/Thr kinase families were used as templates for modeling various query peptides in complex with their respective kinases. The search in PDB (Berman *et al*., 2000) showed a total of 49 kinase–peptide complexes having at least three residues on each side of the phosphorylation site Ser/Thr residue in the bound substrate peptide. They belonged to four families namely PKA, PKB, PHK and CDK2. However, a majority of these complexes were redundant structures which have been solved in complex with identical, or very similar substrate peptides, but with different inhibitor compounds bound in the ATP binding site. Therefore, after removing the redundancy, we selected 1JBP, 1O6K, 1QMZ and 2PHK as representative kinase–peptide complex templates from PKA, PKB, CDK2 and PHK family, respectively. Detailed analysis of these representative structures showed that they share a highly conserved structural fold. For example, PKA and PHK could be superimposed with a RMSD of 1.5 Å in spite of a sequence similarity of only 46%. Therefore, in the absence of a kinase–peptide complex for a given family, a reasonably accurate model can be built based on the crystal structures of these four kinase–peptide complexes. In order to model different query peptides, the backbone was kept fixed in the peptide binding pocket of the kinase, in the bound conformation and the side chains were modeled using backbone dependent rotamer library approach of SCWRL (Canutescu *et al*., 2003). All the modeling tasks were carried out using MODPROPEP (http://www.nii.res.in/modpropep.html) (Kumar and Mohanty, 2007), a software developed in our laboratory for knowledge-based modeling of protein–peptide complexes. The contacting residue pairs between the kinase and the peptide were identified using the criteria of any two atoms of the residue pair being at a distance, ≤4.5 Å. Based on the total number of contacts between the kinase and the peptide, the binding energy was evaluated using Betancourt–Thirumalai (BT) statistical pair potential (Betancourt and Thirumalai, 1999) (Supplementary Data 1). It must be noted that the statistical pair potential used for scoring is not derived from kinase–substrate complexes. Only the peptide structures have been modeled using available kinase–peptide complexes, but the pair potentials used for scoring the complexes are based on the earlier work by Betancourt and Thirumalai (1999). These pair potentials have been derived from the analysis of packing preference of amino acids in a non-redundant set of crystal structures corresponding to different fold families. Thus, the pair potential is not specific to kinase–peptide complexes, rather has general applicability in protein folding and protein ligand binding. All the Ser/Thr containing heptameric peptides in a query protein were ranked as per their binding energy using appropriate interface of MODPROPEP. Since, all computations for our structure-based method were carried out using MODPROPEP, we will refer to our structure-based prediction method as MODPROPEP while comparing with other phosphorylation site identification methods.

2.2 Dataset for benchmarking prediction accuracy of MODPROPEP and comparison with results from GPS, PPSP, SCANSITE, NetPhosK and PREDIKIN

Experimentally identified phosphorylation sites cataloged in Phospho.ELM database were used to compare the prediction accuracy of MODPROPEP with other softwares available for prediction of protein kinase substrates. Phospho.ELM (Diella *et al*., 2004) (version 5.0, May 2006) dataset contains a total of 13 603 phosphorylation instances in 4422 proteins by 263 kinase families. Out of these 263 families, 188 families belonged to Ser/Thr kinases. However, various sequence-based methods have grouped many members of these 188 kinase families to single substrate-specific classes. Based on the classification scheme proposed by GPS (Xue *et al*., 2005), 110 out of these 188 Ser/Thr kinases were grouped into 38 classes, with number of member families in different classes varying from 1 to 12. Since, some of these kinase classes contained too few substrates, we removed the classes with <20 phosphorylation instances. Kinases lacking significant homology with structural templates available in MODPROPEP, were also excluded. Finally, out of the 38 substrate specific classes, 22 classes containing 70 kinase families were selected for benchmarking of various substrate prediction programs (Supplementary Table S1). They contained a total of 2457 phosphorylation instances in 1180 proteins by 70 kinase families. During benchmarking of prediction accuracy on the above mentioned data set, various programs whose performances were compared to our structure-based approach MODPROPEP are GPS, PPSP, SCANSITE, NetPhosK and PREDIKIN. While this work was carried out PREDKIN 1.0 was available, but after completion of our analysis PREDIKIN 2.0 (Saunders and Kobe, 2008; Saunders *et al*., 2008) was released. Even though PREDIKIN 2.0 is a structure-based method similar to PREDIKIN 1.0, it uses a different scoring scheme. Therefore, we included both PREDIKIN 1.0 as well as PREDIKIN 2.0 in our benchmarking analysis. Additional details about these programs and benchmarking process are given in Supplementary Methods.

During benchmarking, we tested whether, for a given kinase the programs could rank the true phosphorylation site(s) from among all possible Ser/Thr containing heptameric peptides present in the corresponding protein, which is listed as a substrate for that particular kinase in Phospho.ELM database. In case of MODPROPEP, all the Ser/Thr containing peptides in a substrate protein are modeled in the binding pocket of the corresponding kinase and ranked as per their binding energy score. It must be noted that these peptides are not modeled in the binding pockets of all other kinases. In fact, other methods like GPS, PPSP, SCANSITE, NetPhosK and PREDIKIN also use a similar prediction strategy and compare all the Ser/Thr containing peptides in a given substrate protein against the motifs/profiles of the known substrates for the corresponding kinase. Supplementary Data 2 shows the typical output by various programs for the microtubule associated protein tau (SWISSPROT Id: P10636) which is phosphorylated by PKA. As can be seen from this file, the prediction outcome for the kinase/substrate protein pair is marked as correct or incorrect depending on the rank of the true phosphorylation site. Similar outputs for all the 1180 substrate proteins in the test set is provided in Supplementary Data 3. We also provide a summary of prediction results by each of the prediction methods by listing SWISSPROT id of the substrate protein, and the rank of the true phosphorylation site in a spread sheet format. Based on the results tabulated in these summary files, the prediction accuracies of MODPROPEP as well as the six other prediction methods were calculated for 22 different kinase families. Percentage accuracies were calculated by considering the number of correct predictions out of the total number of substrate proteins which were tested for kinases belonging to a given family.

2.3 Re-ranking of high-scoring substrate peptides using MM/PBSA

For each substrate protein sequence, the high-ranking 30% Ser/Thr containing peptides scored using BT statistical potential by MODPROPEP, were selected and re-ranked as per their binding energy score using all atom force-field. These peptides were modeled in complex with their respective protein kinases using rotamer library approach (Canutescu *et al.*, 2003). The resulting kinase–peptide complexes were energy minimized. Each complex was decomposed into three molecular species namely complex, receptor (kinase), and ligand (peptide). The binding energy of substrate peptide is calculated as following:

$$\Delta G_{\text{binding}} = G_{\text{complex}} - G_{\text{kinase}} - G_{\text{peptide}}.$$

Apart from *in vacuo* interactions between peptide and the kinase, the $\Delta G_{\text{binding}}$ term should also include solvent effects. Therefore, it was necessary to include binding free energy. However, the calculation of free energy in explicit solvent simulations using Free-Energy Perturbation (FEP) (Rao *et al.*, 1987) approach is an extremely compute intensive process. Therefore, it was impractical to use FEP method for such large number of kinase–peptide complexes. Alternative method for computing 'free energy' at a reasonably modest computational cost is the MM/PBSA approach (Kollman *et al.*, 2000), which uses an implicit solvent model. MM/PBSA involves supplementing the conventional molecular mechanics (MM) energy terms with solvation energy terms. The electrostatic component of solvation free energy is evaluated using continuum model and Poisson–Boltzman calculations, while the non-electrostatic component (i.e. energy contribution from desolvation of nonpolar groups) is evaluated using solvent accessible surface areas of atoms in the molecule. This approach has been used successfully in earlier studies (Basdevant *et al.*, 2006; Stoica *et al.*, 2008; Wang *et al.*, 2001a, b) to evaluate free energy of biomolecular complexes and respective simulation results have shown fairly good agreement with experimental data. Hence, the binding free energy of different substrate peptides in complex with the protein kinase was evaluated using MM/PBSA approach. The MM/PBSA module of AMBER9 molecular dynamics package (Case *et al.*, 2006) was used to calculate the binding free energy of these energy minimized kinase–peptide complexes.

2.4 Receiver operating characteristic curves calculations

The discriminatory power of our prediction method was calculated using receiver operating characteristic curves (ROC) method. For each of the prediction exercise during the benchmarking, the ROC function of R-statistical language environment (http://www.r-project.org) was used for calculation of ROC, specificity and sensitivity values (Supplementary Methods).

3 RESULTS

3.1 Identification of phosphorylation sites by statistical pair potentials

Analysis of the crystal structures of various PKA, PKB, PHK and CDK2 Ser/Thr kinases in complex with substrate peptides (Brown *et al.*, 1999; Lowe *et al.*, 1997; Madhusudan *et al.*, 1994; Yang *et al.*, 2002) revealed that, protein kinases homologous to one of these crystal structures are likely to adopt similar structural folds and conserve their peptide binding pockets. The substrate peptide is bound in the similar extended conformation and relative orientation in most of the crystal structures of kinase–peptide complexes. Figure 1 shows the peptide binding pocket on the kinase fold, highlighting the subsites (Kobe *et al.*, 2005) corresponding to binding pockets for each residue of the substrate peptide. The site of phosphorylation on the substrate peptide is referred as P0, while the three residues flanking the phosphorylation site on the N- and C-terminus are referred as P−3, P−2, P−1 and P+1, P+2 and P+3, respectively. The subsites in the protein kinase which accommodate these seven residues of the substrate peptide are referred as S−3, S−2, S−1, S0, S+1, etc. (Kobe *et al.*, 2005). The specificity of the substrate peptide for a given protein kinase is determined by the complementarities between the peptide residues and the residues lining these subsites. We wanted to investigate the predictive ability of the two widely used pair potential matrices viz. Miyazawa and Jernigan (MJ) (1996), and Betancourt and Thirumalai (BT) (1999) for theoretical estimation of the binding energy between the peptide and the kinase. Earlier studies on application of statistical pair potentials for estimating the binding energies of protein peptide complexes have suggested that MJ matrix is appropriate for the interactions involving primarily hydrophobic interfaces, because it has been derived using the solvent as the reference state for the

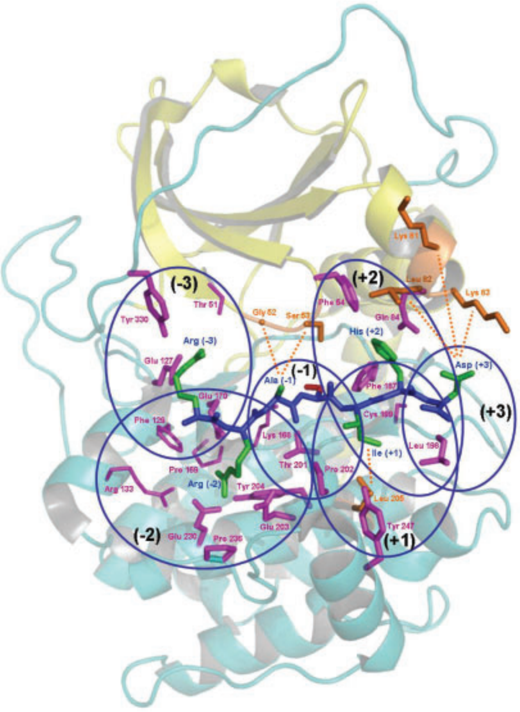

Fig. 1. Crystal structure of PKA (1JBP) in complex with bound substrate peptide. The small N-terminal lobe and large C-terminal lobes are shown in yellow and cyan ribbons, respectively. Binding subsites accommodating the side chain of the peptide residues are shown within the ovals marked −3 to +3. Kinase residues within each subsite which are within a cutoff distance of 4.5 Å from the peptide residues they accommodate are shown in magenta color. Residues which are used in the modified version of the algorithm are shown in the stick representation. Residues represented in orange stick are those which do not come within the distance cut off, but have been reported in the literature to be involved in the determination of substrate specificity. Peptide backbone is shown in blue and the side chains are shown in green.

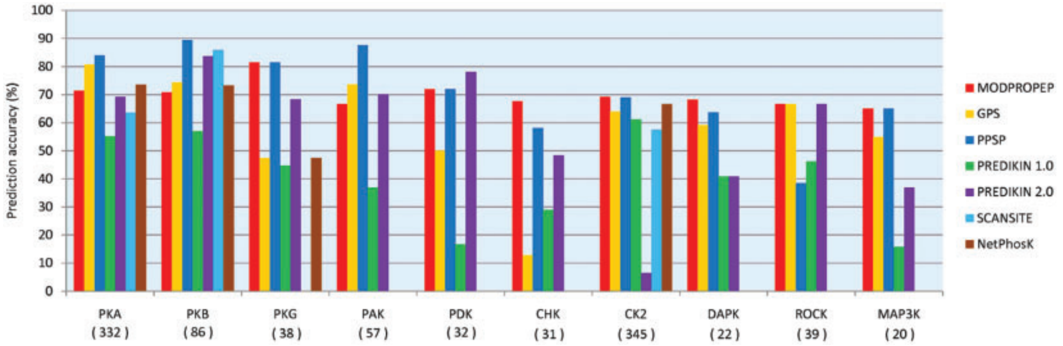

Fig. 2. Comparison of the prediction accuracies of MODPROPEP for 10 different kinase families with other phosphorylation site prediction programs, namely, GPS, PPSP, PREDIKIN, SCANSITE and NetPhosK. MODPROPEP has a prediction accuracy of >60% for these kinases. The total number of known substrate peptides used in prediction is mentioned below the names of each kinase group.

estimation of the favorability of interactions between different amino acid pairs (Schueler-Furman *et al*., 2000). However, it does not score correctly the interactions involving hydrophilic amino acids. On the other hand, BT matrix overcomes this by changing the reference state to a solvent like molecule, threonine. Our analysis on a data set of known kinase substrates also demonstrated that the BT matrix is more suitable for ranking the known kinase substrates with a high score. Since, the substrate peptides for various kinases often contain charged and polar amino acids, in addition to hydrophobic contacts, the BT matrix gives better results compared to MJ pair potential. It may be noted that similar observations have also been made in the context of identification of the MHC binding peptides using statistical pair potential (Altuvia *et al*., 1995; Schueler-Furman *et al*., 1998, 2000). We discuss below the prediction results obtained for various kinase families using BT matrix.

All Ser/Thr kinases showing at least 40% sequence similarity to the four structural templates of kinase–peptide complexes, and having a minimum of 20 different known substrate proteins in Phospho.ELM database, were used to benchmark the predictive power of our structure-based approach. In fact, they corresponded to 22 different kinase families. As discussed before, all Ser/Thr containing heptameric peptides were modeled in the active site pocket of the respective kinases and were ranked as per their binding energy score calculated using the BT matrix. A prediction, for a given substrate protein was considered correct, if the actual experimentally identified phosphorylation site was ranked among the top 30% of all the Ser/Thr containing peptides present in the substrate protein. Since, most prediction methods can only rank the true phosphorylation site among the top few, similar relaxations were also made while evaluating prediction accuracy of other available tools like PREDIKIN, SCANSITE, NetPhosK, GPS and PPSP. Figure 2 shows the results of our structure-based prediction method for 10 different kinase families, namely PKA, PKB, PKG, PAK, PDK, CHK, CK2, DAPK, ROCK and MAP3K. As can be seen from Figure 2, the prediction accuracies of our structure-based method for PKA, PKB and PDK are >70%, while for PKG the prediction accuracy exceeds 80%. For the kinase families, PAK, CHK, CK2, DAPK, ROCK and MAP3K, our structure-based method could also predict with an accuracy >65%. For the purpose of comparison, Figure 2 also shows the results from four other commonly used sequence-based programs GPS, PPSP, SCANSITE, NetPhosK, as well as two different versions of the other structure-based program PREDIKIN. As mentioned earlier, predictions of phosphorylation sites by these programs were carried out using the same dataset which was used for benchmarking the prediction accuracy of MODPROPEP. Our structure-based prediction method outperformed all other sequence-based methods in case of ChK, DAPK and ROCK; while for PGK, PDK, CK2 and MAP3K, the performance of our method was comparable to PPSP. For the protein kinases PKA, PKB and PAK, although MODPROPEP had an accuracy of >65%, PPSP did better than MODPROPEP for PAK, while for PKA and PKB, GPS performed better than MODPROPEP. It must be noted that these sequence-based methods with which we have compared the performance of MODPROPEP need a training data set. Therefore, they might have used a portion of the Phospho.ELM data for training. It was not possible to obtain a common test set which had no overlap with the training set of these methods, because exact information on sequences used as training set by each of these methods was not available. Therefore, this specific choice of test set would result in artificially high prediction-accuracy for only the training-based methods. However, this choice of training set will not lead to artificially high prediction for the structure-based method developed by us, because it uses no training data set for any type of parameter optimization. Figure 2 also shows that the performance of MODPROPEP is superior to that of the other structure-based method PREDIKIN 1.0. However, there has been a significant improvement in the performance of PREDIKIN 2.0 compared to that of PREDIKIN 1.0. Therefore, as can be seen from Figure 2, PREDIKIN 2.0 has higher prediction accuracy than our structure-based method MODPROPEP for kinase families PKB, PAK and PDK, while for other seven kinase groups, performance of MODPROPEP is better or comparable to that of PREDIKIN 2.0. However, we would like to clarify that, unlike our structure-based method, PREDIKIN 2.0 uses a scoring scheme derived from experimental phosphorylation site data and that might have helped in improving its prediction accuracy. Since, PREDIKIN 2.0 uses experimental data for its scoring scheme, it becomes similar in methodology to other profile-based methods. MODPROPEP can also predict substrates for all nine kinase families unlike SCANSITE and NetPhosK, which cannot predict for certain families due to lack of profiles. These results were *per se* extremely encouraging because our approach outperformed other structure-based methods and was

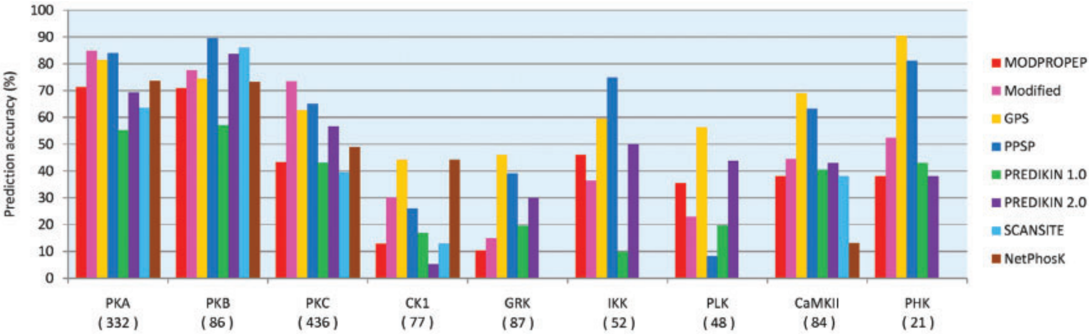

Fig. 3. Comparison of prediction accuracies by modified version of MODPROPEP for those kinase families where prediction accuracy by original MODPROPEP was <60%.

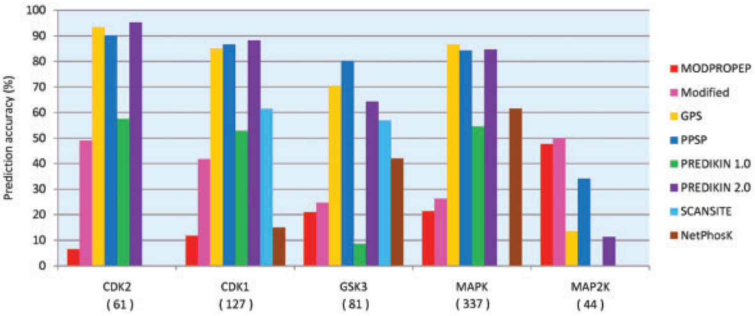

Fig. 4. Comparison of prediction accuracies by modified version of MODPROPEP for the kinases which were modeled using CDK2 as structural template.

comparable in performance to best performing sequence-based methods, even though it does not use any experimental data for training.

Figure 3 shows the prediction results for the protein kinase families PKC, CK1, GRK, IKK, PLK, CaMKII and PHK, while the results for the kinase families CDK2, CDK1, GSK3, MAPK and MAP2K are shown in Figure 4. As can be seen, for PKC, IKK, PLK, CaMKII, PHK and MAP2K the prediction accuracy of MODPROPEP was between 35 and 40%, while for the remaining six kinase families MODPROPEP had a prediction accuracy of 20% or lower. For many of these 12 kinase classes, other sequence-based prediction tools also had a prediction accuracy of <50%. This indicates that they might be genuinely difficult cases for prediction. We also analyzed the prediction accuracy of MODPROPEP for protein kinase families having <20 known phosphorylation sites. As these families contain too few substrates, most other programs which require training using known substrate data, lack the ability to predict substrate for these kinases. Supplementary Figure S1 shows results for these kinases from MODPROPEP predictions alone. As can be seen, prediction accuracy of MODRPOPEP for IPL1 and SGK is >60%, while for CaM-1/IV, LKB1 and RSK-1 prediction accuracy is between 40 and 50%. The prediction accuracy for other families was <40%.

3.2 Improvement of prediction accuracy by alteration of scoring scheme

We proceeded to analyze the possible reasons for the failure of MODPROPEP to rank the known phosphorylation sites with high score in case of kinase families PKC, CK1, GRK, IKK, PLK, CaMKII, PHK, CDK2, CDK1, GSK3, MAPK and MAP2K. As our prediction method involves modeling of protein–peptide complexes, we wanted to investigate whether the accuracy of prediction is dependent on the quality of structural model and hence, on the degree of sequence similarity between the kinase and the structural template. Supplementary Figure S2 shows the prediction accuracy of MODPROPEP as a function of sequence similarity between the kinase and the structural template. As can be seen, prediction accuracy is not directly correlated with the sequence similarity with the template. The prediction accuracy is as high as 69.3% at 42% similarity for CK2, while it is as low as 41.7% at 79% similarity for CDK1. Since, the interactions between peptide and kinase residues considered for calculating binding energy score are restricted to peptide binding site alone, prediction quality is more likely to be dependent on the interaction between the substrate peptide and the binding pocket residues, rather than on the overall accuracy of the structural model of the kinase. Therefore, we analyzed the crystal structures of templates for the composition of each of the subsites which accommodate the peptide residues.

Analysis of the residues lining each of the subsites in PKA indicated that some kinase residues are included in list of subsites even though they are occluded by other residues and do not make direct contact with the residues of the substrate peptide. Similarly, many residues were included as putative substrate binding pocket residues, even though the interactions were mediated primarily by backbone atoms. These anomalies arise primarily because of our simplistic distance-based criteria for identification of binding pocket residues. Such interactions are unlikely to be determinants of specificity of recognition, but their inclusion as binding pocket residues resulted in poor scores for actual substrate peptides. For example, in subsite S−1, kinase residues K168, T201 and P202 are included as pocket residues, even though their side chains do not make contact with the side chain of Ala at P−1 (Figure 1). On the other hand, residues G52 and S53 have been reported to be the specificity determining residues for the amino acid at P−1 position of the peptide, although in the crystal structure of PKA, they do not have direct contact with the Ala at P−1 position in peptide. Therefore, based on careful examination of the crystal structure, we modified the list of binding pocket residues for each of these 12 kinase families by inclusion of additional specificity determining residues (Nishikawa *et al*., 1997; Obata *et al*., 2000) reported in the literature. Predictions were again carried out by MODPROPEP for these 12 kinase families after these modifications. The results from modified version of MODPROPEP for kinase families PKA, PKB, PKC, CK1, GRK, IKK, PLK, CaMKII and PHK are also shown in Figure 3. As can be seen from Figure 3, the inclusion of selected residues resulted in further improvement of prediction accuracy for PKA and PKB. Prediction accuracy of PKA improved from 71.4 to 84.8%, while for PKB the improvement in prediction accuracy was from 70.9 to 77.6%. Most dramatic improvement was observed for PKC with prediction accuracy reaching to 73.4 from 43.3%. The accuracy improved only slightly for CK1 and GRK. However, for these kinases other programs also did not show good prediction accuracy. IKK and PLK showed a decrease in accuracy. CaMKII and PHK whose template was PHK did not show a significant improvement. GSP and PPSP clearly performed better in case of these kinases (Figure 3).

The predictions by MODPROPEP for kinases CDK1, CDK2, GSK3 beta and MAPK had an accuracy of <20%, while GPS, PPSP, PREDIKIN and SCANSITE performed significantly better in these cases (Figure 4). Our analysis of the possible reasons for this poor performance of MODPROPEP indicated that most of the substrates of all these kinases had Pro at P+1 position. Since, these sequence-based methods, as well as PREDIKIN were trained to use Pro as a signature motif, they could predict the substrates of these kinase families with a higher accuracy. These five kinase families were modeled using peptide bound CDK2 as template. We analyzed various modeled peptides in complex with CDK2 based templates to understand, why our method failed to rank known binders with high score. Supplementary Figure S3 shows the interacting residues in various subsites in the crystal structure of CDK2–peptide complex. As can be seen, Arg at P+2 position on the substrate peptide residue is exposed to the solvent and does not make direct contact with any of the kinase residues. Therefore, one would *a priori* expect that substrates having polar or charged residues at P+2 position will be preferred. However, our algorithm does not include any penalty for the hydrophobic residues being exposed to the solvent. Hence, it fails to discriminate peptides having hydrophobic residues at P+2 position. Similarly, the Pro at P+1 position on the substrate peptide has Glu162 and Arg169 as potential interaction partners in S+1 subsite based on distance-based cut off. However, careful examination of the orientations of the side chains in S+1 pocket indicates that Glu162 is oriented away from Pro and Val164 occludes direct contact between Arg169 and Pro. In fact based on analysis of kinase–peptide complexes, Zhu *et al.* (2005) have suggested that in case of PKA the main chain carbonyl of Gly200 which correspond to Val164 of CDK makes hydrogen bond with the backbone amide of the residue at P+1 position. On the other hand, in proline directed kinases like CDK2 the Pro at P+1 lacks the backbone amide group, but the Pro side chain interacts with the conserved hydrophobic residue Val164 in the S+1 pocket. However, our distance-based criteria includes two charged residues as interaction partners for Pro, thus resulting in poor score for peptides containing Pro at P+1. Similarly, the Lys at P+3 is stabilized by interactions with the phospho-Thr160 residue in the kinase (Supplementary Figure S3). However, our current scoring potential does not contain any score for the interaction involving non-standard amino acids. Therefore, while investigating effect of modifications to the scoring scheme and pocket residues, we used the potentials of Asp in place of phospho-Thr for scoring the peptides and excluded the subsites P+1 and P+2 from scoring. As can be seen from Figure 4, this resulted in a dramatic improvement in the prediction accuracy for CDK1 and CDK2. The prediction accuracy for CDK2 changed from 7 to 49%, while for CDK1, the improvement was from 11 to 41%. However, in case of other kinases the improvement was only marginal. This may be because of the involvement of other regulatory mechanisms in determination of the substrate specificity of these kinases. For example, cyclin protein affects the activity and substrate recruitment of CDK2. As seen in the crystal structure of CDK2 (1QMZ), Lysine residue at P+3 position makes contact with E269 in the cyclin protein. However, our current version of MODPROPEP uses only the residues in kinase domain which are in contact with the substrate peptide, for calculation of binding scores. In future versions, use of activated kinase domains with inclusion of contacts from the additional proteins in the complex might help in improving the prediction accuracy. Thus, our detailed analysis of the predictions for these 12 kinase families by MODPROPEP, clearly highlighted the possible reasons for the poor performance of the structure-based method and gave valuable clues for improvement in the prediction accuracy by suitable alteration in the computational protocol.

3.3 ROC analysis

The prediction accuracy of our structure-based method was comparable to, or better than the best available sequence-based methods for 11 kinase families. We decided to further benchmark the robustness of our predictions for these kinase families by a rigorous analysis involving ROC curves. We computed the ROC curve for each of the 10 kinase families, which showed a prediction accuracy of >65% in our analysis by unmodified MODPROPEP. Supplementary Figures S4 shows the ROC curves for the representative case of CK2 and also when all these 10 groups were merged into a single class. As can be seen, the values for area under the curve (AUC) are 0.792 and 0.764, respectively. Supplementary Table S2 gives the values for AUC, sensitivity (Sn) and specificity (Sp) for the predictions by MODPROPEP for each of these 10 kinase families. These results further establish the statistical significance of our predictions. Since, the kinomes of several organism are known to have many members belonging to these families, MODPROPEP can aid in deciphering signaling networks in various genomes.

3.4 Re-ranking of kinase–peptide complexes using MM/PBSA

Thus, our analysis demonstrated that MODPROPEP is a powerful structure-based approach which successfully predicts substrates for 10 different kinase families without utilizing any kinase family

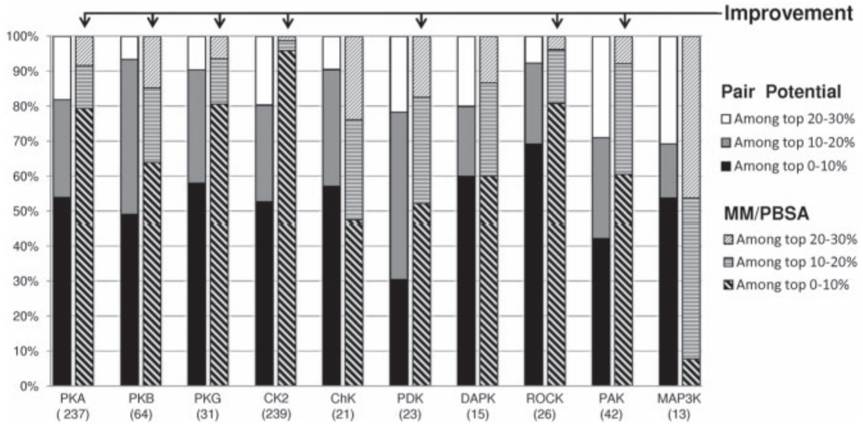

Fig. 5. Comparison of the ranking of true phosphorylation site among top ranked 30% peptides by pair potential and MM/PBSA method. For each of the 10 protein kinase groups for which BT matrix had good prediction accuracy, the distribution of the percentage of cases in which the actual phosphorylation site was among 0–10%, 10–20% and 20–30% of top-scoring peptide is shown. The filled bars represent scoring by pair potential, while bars with stripes represent scoring by MM/PBSA methods.

specific substrate data for training. Moreover, because of the scoring by simple statistical pair potential, it is not compute intensive and can be used for high throughput analysis of sequences in a genomic scale. However, its utility can be further improved if percentile cut off for bracketing the correct phosphorylation site can be further lowered from 30%. As demonstrated earlier, in case of protein-structure prediction problems, multi-scale modeling strategy can potentially help in improving the rank of the correct substrate. Therefore, scoring by pair potentials can be used as a first level of search, while sites ranked within top 30% can be reranked using an all atom forcefield. For the 10 kinase families for which MODPROPEP could successfully rank the correct binding site within top 30%, we carried out detailed all atom modeling of all the Ser/Thr containing peptides in the binding site of the respective kinases. Supplementary Figure S5 shows the C^{α} RMSD values for the kinase and the peptide backbone, from the respective template structures for all the energy minimized kinase peptide complexes analyzed in case of PKA. As can be seen in all cases, modeled complexes remain close to the template structure. Similar trend was also observed for kinase–peptide complexes belonging to other families. For each kinase–peptide complex, interaction energy between the kinase and the peptide was computed using MM-PBSA approach and all the modeled peptides were re-ranked as per their MM-PBSA binding energy values. Figure 5 shows the comparison of the ranking by MM-PBSA and pair potential for all the 10 kinase families. As can be seen for PKA, out of the total of 332 substrate proteins, in case of 237 proteins MODPROPEP could rank the true phosphorylation site within top 30%. As per pair potential ranking, out of these 237 cases, in case of 128 the phosphorylation site was within top 10%, in case of additional 66 proteins the phosphorylation site was ranked within 10–20%. Thus, for a total of 194 cases the phosphorylation site was ranked within top 20% and in case of 43 proteins the phosphorylation site was ranked within 20–30% by MODPROPEP. Interestingly upon re-ranking by MM-PBSA approach, the number of true phosphorylation sites within top 10% increased to 188 and the number of true phosphorylation sites within top 20% increased to 217. Thus, by re-ranking, there is a significant enrichment of true phosphorylation site in top 10 and 20% window. In Figure 5, similar results have been plotted for all 10 kinase families. Since, there were different number of substrate proteins for different families, they have been represented as percentage of the total number of substrate proteins considered for modeling by MM-PBSA. As can be seen, in case of seven out of 10 kinase families, re-ranking has helped in increasing the number of cases where the true phosphorylation site could be bracketed within top 10% window. Figure 6 shows the AUC values obtained from ROC analysis for the ranking using pair potential as well as MM-PBSA. As can be seen, AUC values have increased in all cases except for ChK and MAP3K, thus demonstrating the utility of re-ranking the pair-potential predictions using MM-PBSA approach. Thus, our results demonstrate that the prediction accuracy of MODPROPEP can be further improved if a

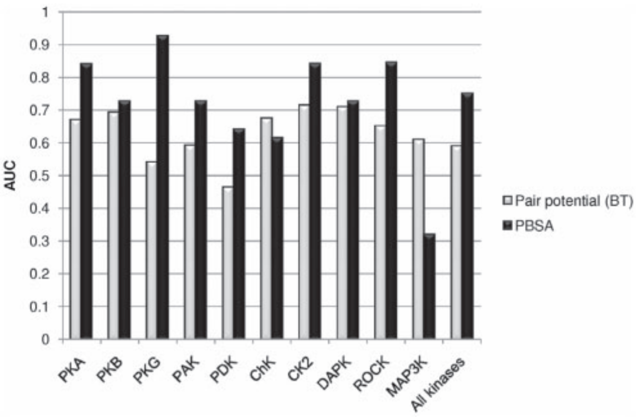

Fig. 6. AUC values in ROC analysis for 30% top ranking peptides scored by BT pair potential matrix. These values are compared with the AUC values obtained when the same set of peptides were re-ranked using the MM/PBSA method.

multi-scale modeling approach involving re-ranking of pair potential predictions by MM-PBSA energy values is implemented.

3.5 DISCUSSIONS AND CONCLUSION

In this work, we have developed a novel structure-based approach for predicting substrates of protein kinases. The putative substrate peptides are modeled in the substrate binding pockets of kinases using the available crystal structures of kinase–peptide complexes as templates. The binding energies of these peptides in complex with the kinase are evaluated using a residue-based statistical pair potential derived by Betancourt and Thirumlai (1999). This pair potential is a generalized scoring function appropriate for protein folding and docking simulations. Thus, our method does not involve any parameter optimization using known examples of substrates. We have carried out a detailed benchmarking of this approach on the experimental data available in the Phospho.ELM database and compared our results with those from a number of other phosphorylation site prediction tools. Our results indicate that the structure-based method developed in this work can predict >60% of the experimentally identified substrates for 11 protein kinases. The prediction accuracies for PKA, PKB, PKG, and PDK were well >70% with PKG having the highest prediction accuracy of 81.5%. The other kinase groups for which our approach showed good prediction accuracy were ChK, CK2, DAPK, ROCK and MAP3K. Our approach also outperformed all other sequence-based prediction tools for PKG, PDK, ChK, CK2, DAPK, ROCK and MAP3K. It must be noted that methods like GPS and PPSP have been trained on the same data which we have used for benchmarking. Thus, they show high-prediction accuracy. We did not divide the data into training and test set as our purpose was to evaluate the prediction accuracy of our structure-based method which does not use any experimental data for training. It is encouraging to note that the prediction accuracy of our method is comparable to other sequence-based methods like GPS, PPSP, SCANSITE and NetPhosK, even though it does not use any experimental phosphorylation site data for training. We also carried out ROC curve analysis for analyzing the robustness of our structure-based prediction approach. The area under curve (AUC) values for these 10 kinases ranged from 0.681 (MAP3K) to 0.838 (PKG). We have also compared our prediction results with PREDIKIN which is the only other structure-based approach, but uses a different scoring scheme. Our results clearly demonstrated the better prediction accuracy of our method compared to PREDIKIN 1.0 (Brinkworth *et al.*, 2003). However, PREDIKIN 2.0 (Saunders and Kobe, 2008; Saunders *et al.*, 2008) uses a scoring scheme, derived from experimental phosphorylation site data. Hence, it performed better than MODPROPEP in 12 out of the 22 kinase families analyzed in our benchmarking exercise. On the other hand, our structure-based method, MODPROPEP does not involve any training and uses the generalized statistical pair potentials for scoring the bound peptides. However, it performs better than PREDIKIN 2.0 for eight kinase families and performance is comparable in case of two other kinase families.

The predictions of MODPROPEP which have been compared with other methods are based on scoring by pair potential alone. We have also demonstrated that the use of a multi scale approach and the re-ranking of the high-scoring peptides identified by pair potential, using all atom MM/PBSA, improves the percentile score of the true phosphorylation site. More detailed studies are necessary to test whether implementation of the complete multi-scale approach can further improve the prediction accuracy of our structure-based approach. The current scoring scheme of MODPROPEP does not include any confidence value for the prediction scores. However, in view of the high variability in the prediction quality, it is desirable to have a confidence score for the predictions. Detailed analysis of the scores for cognate and non-cognate peptides might help in developing a method for assigning confidence value to predictions by MODPROPEP. The feasibility of assigning confidence levels to our prediction scores will be explored in future work.

Many kinases such as GSK3 (Cole *et al.*, 2004), CK2 (Battistutta, 2009) and CDK (Morgan, 1997) etc require prior phosphorylation of one or more residues which modulated the substrate recognition and catalytic activity of kinase. The regulation of the activity of kinases through these phosphorylation events is a fairly complex mechanism. If the phosphorylated residue, as in case of CDK2 (Supplementary Figure S3), makes direct contact with the substrate peptide, it might play a crucial role in substrate recruitment. Currently neither MODPROPEP nor any of the other prediction programs take into consideration the requirement of such phosphorylation priming while making predictions. Hence, these aspects also need to be explored for further improvement of the prediction accuracy.

It must also be noted that MODPROPEP as well as the existing sequence and structure-based methods like GPS, PPSP, NetPhosK, SCANSITE and PREDIKIN etc attempt to predict the substrates for kinases by considering the specificity of the kinase domain alone for the phosphorylation site peptide. In fact this is based on the implicit assumption that the query protein is a known phosphoprotein and is present in the same cellular compartment as the kinase. Therefore, if a user analyzes an extracellular protein or a transmembrane protein with an extracellular domain by these phospho site prediction programs the results may not be biologically relevant. Thus, apart from peptide specificity, the *in vivo* phosphorylation events are also governed by presence of kinase docking motifs, localization and co-expression of kinase and the substrate protein. Therefore, computational methods for deciphering *in vivo* phosphorylation networks should also take into consideration context dependent features like kinase docking motifs, localization and co-expression etc. in addition to specificity for phosphorylation site peptide. Recent work by Linding and coworkers (2007) has demonstrated that inclusion of context dependent features as additional filters over phosphorylation site predictions by NetPhosK and SCANSITE helps in improving the overall prediction accuracy. Similar phosphorylation network identification tools can also be developed by combining predictions of MODPROPEP with other context dependent criteria.

ACKNOWLEDGEMENTS

Authors thank Prof. A. Surolia for his encouragement and support. N.K. thanks CSIR, India for award of senior research fellowship.

Funding: Department of Biotechnology, Government of India grants to National Institute of Immunology; BTIS project of Department of Biotechnology, India grants to D.M.

Conflict of Interest: none declared.

REFERENCES

Altuvia,Y. *et al*. (1995) Ranking potential binding peptides to MHC molecules by a computational threading approach. *J. Mol. Biol*., **249**, 244–250.

Amanchy,R. *et al*. (2007) A curated compendium of phosphorylation motifs. *Nat. Biotechnol.*, **25**, 285–286.

Basdevant,N. *et al*. (2006) Thermodynamic basis for promiscuity and selectivity in protein-protein interactions: PDZ domains, a case study. *J. Am. Chem. Soc*., **128**, 12766–12777.

Battistutta,R. (2009) Protein kinase CK2 in health and disease: Structural bases of protein kinase CK2 inhibition. *Cell Mol. Life Sci*., **66**, 1868–1889.

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

Betancourt,M.R. and Thirumalai,D. (1999) Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci*., **8**, 361–369.

Blom,N. *et al*. (2004) Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, **4**, 1633–1649.

Brinkworth,R.I. *et al*. (2003) Structural basis and prediction of substrate specificity in protein serine/threonine kinases. *Proc. Natl Acad. Sci. USA*, **100**, 74–79.

Brown,N.R. *et al*. (1999) The structural basis for specificity of substrate and recruitment peptides for cyclin-dependent kinases. *Nat. Cell Biol*., **1**, 438–443.

Caenepeel,S. *et al*. (2004) The mouse kinome: discovery and comparative genomics of all mouse protein kinases. *Proc. Natl Acad. Sci. USA*, **101**, 11707–11712.

Canutescu,A.A. *et al*. (2003) A graph-theory algorithm for rapid protein side-chain prediction. *Protein Sci*., **12**, 2001–2014.

Case,D.A. *et al*. (2006) AMBER 9, University of California, San Francisco.

Cohen,P. (2000) The regulation of protein function by multisite phosphorylation—a 25 year update. *Trends Biochem. Sci*., **25**, 596–601.

Cole,A. *et al*. (2004) Further evidence that the tyrosine phosphorylation of glycogen synthase kinase-3 (GSK3) in mammalian cells is an autophosphorylation event. *Biochem. J*., **377**, 249–255.

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

Kobe,B. *et al*. (2005) Substrate specificity of protein kinases and computational prediction of substrates. *Biochim. Biophys. Acta*, **1754**, 200–209.

Kollman,P.A. *et al*. (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res*., **33**, 889–897.

Krupa,A. and Srinivasan,N. (2002) The repertoire of protein kinases encoded in the draft version of the human genome: atypical variations and uncommon domain combinations. *Genome Biol*., **3**, RESEARCH0066.

Kumar,N. and Mohanty,D. (2007) MODPROPEP: a program for knowledge-based modeling of protein-peptide complexes. *Nucleic Acids Res*., **35**, W549–W555.

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

Lowe,E.D. *et al*. (1997) The crystal structure of a phosphorylase kinase peptide substrate complex: kinase substrate recognition. *Embo J*., **16**, 6646–6658.

Madhusudan *et al*. (1994) cAMP-dependent protein kinase: crystallographic insights into substrate recognition and phosphotransfer. *Protein Sci.*, **3**, 176–187.

Manning,G. *et al*. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.

Miyazawa,S. and Jernigan,R.L. (1996) Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**, 623–644.

Morgan,D.O. (1997) Cyclin-dependent kinases: engines, clocks, and microprocessors. *Annu. Rev. Cell Dev. Biol*., **13**, 261–291.

Nishikawa,K. *et al*. (1997) Determination of the specific substrate sequence motifs of protein kinase C isozymes. *J. Biol. Chem*., **272**, 952–960.

Obata,T. *et al*. (2000) Peptide and protein library screening defines optimal substrate motifs for AKT/PKB. *J. Biol. Chem.*, **275**, 36108–36115.

Obenauer,J.C. *et al*. (2003) Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic Acids Res*., **31**, 3635–3641.

Plowman,G.D. *et al*. (1999) The protein kinases of Caenorhabditis elegans: a model for signal transduction in multicellular organisms. *Proc. Natl Acad. Sci. USA*, **96**, 13603–13610.

Rao,S.N. *et al*. (1987) Free energy perturbation calculations on binding and catalysis after mutating Asn 155 in subtilisin. *Nature*, **328**, 551–554.

Saunders,N.F. and Kobe,B. (2008) The Predikin webserver: improved prediction of protein kinase peptide specificity using structural information. *Nucleic Acids Res.*, **36**, W286–W290.

Saunders,N.F. *et al*. (2008) Predikin and PredikinDB: a computational framework for the prediction of protein kinase peptide specificity and an associated database of phosphorylation sites. *BMC Bioinformatics*, **9**, 245.

Schueler-Furman,O. *et al*. (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci*., **9**, 1838–1846.

Schueler-Furman,O. *et al*. (1998) Knowledge-based structure prediction of MHC class I bound peptides: a study of 23 complexes. *Fold Des*., **3**, 549–564.

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

Stoica,I. *et al*. (2008) Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. *J. Am. Chem. Soc.*, **130**, 2639–2648.

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

Wang,J. *et al*. (2001a) Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *J. Am. Chem. Soc*., **123**, 5221–5230.

Wang,W. *et al*. (2001b) An analysis of the interactions between the Sem-5 SH3 domain and its ligands using molecular dynamics, free energy calculations, and sequence analysis. *J. Am. Chem. Soc*., **123**, 3986–3994.

Xue,Y. *et al*. (2005) GPS: a comprehensive www server for phosphorylation sites prediction. *Nucleic Acids Res*., **33**, W184–W187.

Xue,Y. *et al*. (2006) PPSP: prediction of PK-specific phosphorylation site with Bayesian decision theory. *BMC Bioinformatics*, **7**, 163.

Yang,J. *et al*. (2002) Crystal structure of an activated Akt/protein kinase B ternary complex with GSK3-peptide and AMP-PNP. *Nat. Struct. Biol*., **9**, 940–944.

Zhou,F.F. *et al*. (2004) GPS: a novel group-based phosphorylation predicting and scoring method. *Biochem. Biophys. Res. Commun*., **325**, 1443–1448.

Zhu,G. *et al*. (2005) Exceptional disfavor for proline at the P+1 position among AGC and CAMK kinases establishes reciprocal specificity between them and the proline-directed kinases. *J. Biol. Chem*., **280**, 10743–10748.