



Prediction of phosphorylation sites using SVMs

Jong Hun Kim^{1,2}, Juyoung Lee¹, Bermseok Oh¹, Kuchan Kimm¹
and InSong Koh^{1,3,*}

¹National Genome Research Institute, 5 Nokbun-Dong, Eunpyung-Gu, Seoul, 122-701 Korea, ²Samsung Medical Center, 50 Ilwon-dong, Kangnam-ku, Seoul, 135-710 Korea and ³Brain Korea 21 Program, College of Medicine, Korea University, 5 Anam-Dong, Sungbuk-Ku, Seoul, 136-701 Korea

Received on November 26, 2003; revised on June 19, 2004; accepted on June 22, 2004
Advance Access publication July 1, 2004

ABSTRACT

Motivation: Phosphorylation is involved in diverse signal transduction pathways. By predicting phosphorylation sites and their kinases from primary protein sequences, we can obtain much valuable information that can form the basis for further research. Using support vector machines, we attempted to predict phosphorylation sites and the type of kinase that acts at each site.

Results: Our prediction system was limited to phosphorylation sites catalyzed by four protein kinase families and four protein kinase groups. The accuracy of the predictions ranged from 83 to 95% at the kinase family level, and 76–91% at the kinase group level. The prediction system used—PredPhospho—can be applied to the functional study of proteins, and can help predict the changes in phosphorylation sites caused by amino acid variations at intra- and interspecies levels.

Availability: PredPhospho is available at <http://www.ngri.re.kr/proteo/PredPhospho.htm>.

Contact: jh7521@ngri.re.kr

Supplementary information: <http://www.ngri.re.kr/proteo/supplementary.doc>

INTRODUCTION

Protein phosphorylation is involved in signal transduction in various processes, including the cell cycle, proliferation and apoptosis. Between 30 and 50% of eukaryotic proteins undergo phosphorylation (Pinna and Ruzzene, 1996). Accordingly, accurate predictions of the phosphorylation sites of eukaryotic proteins will help in understanding the overall intracellular events.

Consensus sequences have generally been used to predict the sites that are recognized by a given kinase. The consensus sequence of a kinase can provide a general picture of the primary structures of any sites that can be phosphorylated efficiently by that kinase. Deriving consensus sequences is very

time-consuming. Moreover, consensus sequences cannot successfully reflect kinase recognition sequences because kinases recognize the residues surrounding the target, and the amino acids bordering phosphorylation sites are in turn dependent on other nearby residues (Pinna and Ruzzene, 1996).

Blom *et al.* (1999) used neural networks to predict phosphorylation sites, which resulted in better outcomes than predictions based on consensus sequences. However, their prediction system—NetPhos—cannot provide information on the kinases involved. Yaffe *et al.* (2001) used a profile method and developed the web program Scansite. They derived the profiles of 63 kinases from experimental data. The profile approach requires only the phosphorylated sequences to construct the profiles and can avoid the hazards of using non-phosphorylated sequence data, which may contain false negative sites. Scansite correctly predicted ~70% of known phosphorylation sites in PhosphoBase at the low stringency—the top 5% cutoff level (Yaffe *et al.*, 2001).

Manning *et al.* (2002) found 518 human protein kinase genes in the human genome sequence with the hidden Markov model (HMM) profile and confirmed the identities of more than 90% of the identified kinase genes using cDNA cloning. They also classified the protein kinase superfamily into 9 broad groups and subdivided the groups into 134 families and 204 subfamilies, using sequence comparisons of kinase catalytic domains (Manning *et al.*, 2002).

Based on this classification, we have attempted to predict the phosphorylation sites in kinase group-specific and family-specific ways using the support vector machines (SVMs) derived from the statistical learning theory proposed by Vapnik and Chervonenkis in 1995 (Kecman, 2001; Vapnik, 1995, 1998). In contrast to consensus sequences SVMs, which are classes of neural networks, can consider positional correlations of amino acids (Blom *et al.*, 1999).

The data on phosphorylation sites in the public databases mainly concern four kinase families: cyclin-dependent kinase (CDK), casein kinase 2 (CK2), protein kinase A (PKA) and protein kinase C (PKC). Furthermore, most available data

*To whom correspondence should be addressed.

Table 1. The number of phosphorylation sites recognized by the protein kinase families and groups used in the study

AGC group		CAMK group		TK group		CMGC group		Other group	
Family	Number	Family	Number	Family	Number	Family	Number	Family	Number
GRK	27–36	CAMK1	4–6	Abl	7–9	CDK	72–91	CK2	58–81
PKA	133–187	CAMK2	21–34	CTK	2–3	MAPK	22–30		
PKB	24–28	MAPKAPK	2–3	EGFR	20–26	GSK	28–31		
PKC	118–174	MLCK	2–4	Fak	1–2	Total ^a	118–141		
PKG	25–29	CAMKL	15–16	Fer	4				
RSK	2–7	PHK	9–13	InsR	20–25				
SGK	1	PKD	1	JakA	5				
Total ^a	285–383	Others ^b	5	Met	1				
		Total ^a	56–71	PDGFR	12–13				
				Src	25–30				
				Others ^c	134–154				
				Total ^a	219–254				

^aBecause of redundancies in the data, the total is less than the arithmetic sum of the sites of each family.

^bPhosphorylation sites in the CAMK group that cannot be sorted according to Manning's classification.

^cTyrosine phosphorylation sites lacking kinase information.

Abbreviations for kinases: Abl, cellular homolog of Abelson murine leukemia virus oncogene; AGC, protein kinase AGC; AKT, v-akt murine thymoma viral oncogene homolog; CAMK, calcium/calmodulin-dependent kinase; CAMKL, CAMK-like protein kinase; CDK, cyclin-dependent kinase; CMGC, kinases including CDK, MAPK, GSK3 and cyclin-dependent kinase-like kinase; CK2, casein kinase 2; CTK, protein tyrosine kinase related to Csk; EGFR, epidermal-growth-factor receptor; Fak, focal adhesion kinase; Fer, Fps (Fujinami poultry sarcoma)/Fes (feline sarcoma)-related protein; GRK, G-protein-coupled receptor kinase; GSK, glycogen synthase kinase; InsR, insulin receptor; JakA, Janus kinase A; MAPK, mitogen-activated protein kinase; MAPKAP, MAPK-activated protein kinase; MET, Met proto-oncogene; MLCK, myosin light chain I kinase; PDGFR, platelet-derived-growth-factor receptor; PHK, phosphorylase kinase; PKA, protein kinase A; PKB, protein kinase B; PKC, protein kinase C; PKD, polycystic kidney disease; PKG, protein kinase G; RAD53, protein kinase Chk2; RSK, ribosomal protein S6 protein kinase; SGK, serum- and glucocorticoid-inducible kinase; Src, cellular homolog of Rous sarcoma virus oncogene; and TK, tyrosine kinase.

on kinases relate to the four main kinase groups: protein kinase AGC (AGC), calcium/calmodulin-dependent kinase (CAMK), the group of kinases including CDK, mitogen-activated protein kinase (MAPK), glycogen synthase kinase 3 (GSK3), cyclin-dependent kinase-like kinase (CMGC) and tyrosine kinase (TK) (Boeckmann *et al.*, 2003; Kreegipuu *et al.*, 1999; Manning *et al.*, 2002). The prediction systems were thus inevitably limited to these four families and four groups. By considering kinase classification in predicting phosphorylation sites, we could enhance the accuracy of the predictions and obtain kinase information for each possible phosphorylation site. We have implemented the prediction system in the WWW program, PredPhospho (<http://www.ngri.re.kr/proteo/PredPhospho.htm>).

SYSTEMS AND METHODS

Homology-reduced (+) datasets

Phosphorylation site sequences were obtained from PhosphoBase and the feature table of SWISS-PROT (release 42.10) (Kreegipuu *et al.*, 1999). The considered window size was 3–25 symmetrical residues. For example, a window size of three residues means seven amino acids centering on serine, threonine or tyrosine residues. We herein designate phosphorylated site sequences as (+) sites and non-phosphorylated site sequences as (–) sites. Phosphorylation sites not annotated by experimental results were excluded from the data obtained from the SWISS-PROT.

To test the generalization ability of the SVMs more accurately, we discarded over 70% of identical sequences in the (+) dataset before cross-validation testing. If the test set data are highly homologous to the training set data, the accuracy of a prediction can be overestimated. Homology reduction allows us to avoid such bias.

As significant amounts of data were needed to train and test the SVMs, only those families and groups of kinases with more than 50 data entries in the databases were used. The CDK family has 72–91 (+) sites, the CK2 family 58–81, the PKA family 133–187 and the PKC family 118–174. The AGC, CAMK, CMGC and TK groups have 285–383, 56–71, 118–141 and 219–254 (+) sites, respectively. Since the amount of data varies according to window size, the data have been expressed as ranges. That is, the larger the window size, the more likely is the phosphorylation sites near the N- or C-terminals of proteins will be excluded. A small window, such as three residues, has too small a resolution to represent its unique sequence. Generally, a window of 25 residues contains the smallest number of data, and a window of 4–5 residues has the largest. The protein kinase families and groups used in the study are shown in Table 1. To compare the accuracy of the kinase-specific and non-specific predictions, we divided the phosphorylation sites in the public databases into serine and threonine sites. For the (+) sites, there were 667–855 serine sites and 163–216 threonine sites.

(-) Datasets

Although, the SWISS-PROT provides much information about phosphorylation sites, it contains many more false (-) sites than PhosphoBase, in which the data are filtered manually by reference to many publications (Kreegipuu *et al.*, 1999). Therefore, we used the non-annotated sites of the PhosphoBase as (-) sites.

The (-) sites of the PhosphoBase can be proved to be (+) sites in future experiments. However, (-) sites in the database were used in the study for the following reasons. First, only a few serine, threonine and tyrosine residues are phosphorylated (Yaffe *et al.*, 2001). Second, the composition of the amino acids flanking phosphorylated serine, threonine and tyrosine residues is skewed away from that of the amino acids flanking non-phosphorylated serine, threonine and tyrosine residues (Yaffe *et al.*, 2001). Therefore, SVMs, which in practice allow some training errors, would regard false (-) sites as errors. Third, the family- and group-wise predictions of phosphorylation sites will remove many false (-) sites. For example, if there are some false (-) sites that are actually phosphorylated by the PKA kinase family, they are false (-) sites in the context of the PKA kinase family. However, they are true (-) sites when we are predicting the phosphorylation sites of other kinases.

Selection of (-) sites and cross-validation

As there are many more (-) sites than (+) sites, the SVMs trained with all the (-) sites will predict all sites as (-) sites. To avoid such overweighting by (-) sites, a similar number of (-) sites and (+) sites should be selected for use in the study. The (-) sites selected randomly cannot reflect the distribution of (-) sites and homologous sequences can be chosen by random selection. Therefore, we grouped (-) sites and selected representative sequences in each group to reduce the number of (-) sites to 1.5 times the number of (+) sites. In that ratio, the accuracy (Ac) of the preliminary test was maximum (data not shown).

Because of limitations in memory and other computational resources, existing multiple alignment software could not be used to group over several thousands of (-) sites. Instead, the grouping method described in the supplementary material was used.

Ten sets of (-) sites for each kinase family and kinase group were constructed. For each set of (-) sites, a 7-fold cross-validation test was performed by randomly dividing (+) sites and (-) sites into training and test sets in a 6:1 ratio. The 7-fold cross-validation test was performed 10 times for each kinase family and kinase group. The optimal parameter set was determined by selecting the highest median Ac value for ten 7-fold cross-validation tests.

Prediction system assessment

Accuracy (Ac), sensitivity (Sn) and specificity (Sp) are often used to evaluate prediction systems. However, when the

accuracies of the positive predictions and those of the negative predictions must be considered simultaneously, Sn and Sp values are both inadequate. Furthermore, if the numbers in both classes are different, the Ac value—the measurement that considers only the number of correct predictions—is not useful either. For example, ~70% of natural secondary protein structures are non-helical, whereas only 30% are helical. A constant prediction of ‘non-helix’ is bound to be accurate 70% of the time, although it is useless (Baldi, 2001). In our study, we were able to use the Ac value as a criterion for selecting an optimal parameter set because the number of (-) sites was adjusted to 1.5 times the number of (+) sites. In addition to the Ac, Sn and Sp values, a correlation coefficient (CC) can be used to assess a prediction system. The CC has a value ranging from -1 to +1. The closer the CC value is to +1, the better the prediction system. If Sn, Sp, Ac and CC are each expressed in terms of true positive (TP), false negative (FN), true negative (TN) and false positive (FP) predictions, each measurement is given as follows:

$$Sn = \frac{TP}{TP + FN}, \quad Sp = \frac{TN}{TN + FP},$$

$$Ac = \frac{TP + TN}{TP + FP + TN + FN},$$

and

$$CC = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP + FN) \times (TN + FP) \times (TP + FP) \times (TN + FN)}}.$$

Support vector machines

The algorithm and parameters used for the SVMs are detailed in the supplementary material.

WWW program

The PredPhospho was implemented using the PERL (version 5.8) programming language. Predictions using the SVMs can find possible phosphorylation sites recognized by four kinase groups (AGC, CAMK, CMGC and TK) and four kinase families (CDK, CK2, PKA and PKC). Predictions using consensus sequences can also identify the potential phosphorylation sites of the 29 kinds of kinases. References for the consensus sequences of the 29 kinases are given on the webpage. By selecting the kinds of kinases that the user is interested in and by inputting the substrate sequence, the user can obtain predictions on phosphorylation sites in the query sequence and the kinds of kinases that may act at these sites.

RESULTS

Prediction of phosphorylation sites recognized by kinase families (CDK, CK2, PKA and PKC) and groups (AGC, CAMK, CMGC and TK)

Phosphorylation sites recognized by the four kinase families were predicted by the SVMs, and the accuracies of the

Table 2. The results of phosphorylation site predictions for the CDK, CK2, PKA and PKC families

Kinase family	Method	Optimal kernel	Optimal C	Optimal γ	Optimal window size	Ac (%)	Sn (%)	Sp (%)	CC
CDK	SVM	Sigmoid	500	0.004	18	95.09	95.02	95.10	0.90
	Consensus					77.70	45.45	99.16	0.56
CK2	SVM	RBF	1	0.04	10	91.47	83.90	96.43	0.82
	Consensus					83.95	76.51	88.78	0.66
PKA	SVM	RBF	1	0.04	5	89.98	88.32	91.11	0.79
	Consensus					81.61	56.09	98.70	0.64
PKC	SVM	Sigmoid	1	0.04	11	82.90	78.71	85.79	0.65
	Consensus					72.59	47.54	89.78	0.42

The consensus sequences of the CDK, CK2, PKA and PKC families are [S/T]-P-x-[K/R], [S/T]-x(2)-[D/E] (Prosite ID; PDOC0006), [R/K]-[R/K]-x-[S/T] (Prosite ID; PDOC0004) and [S/T]-x-[R/K] (Prosite ID; PDOC0006), respectively; [S/T] being phosphorylation sites.

RBF (radial basis function) and sigmoid are types of kernel functions of SVMs. C and γ are parameters of SVMs. These kernel functions and parameters of SVMs are described in the supplementary material.

Table 3. The results of phosphorylation site predictions for the AGC, CAMK, CMGC and TK groups of kinases

Kinase group	Optimal kernel	Optimal C	Optimal γ	Optimal window size	Ac (%)	Sn (%)	Sp (%)	CC
AGC	RBF	1	0.004	17	83.31	67.81	93.63	0.65
CAMK	RBF	1	0.04	11	80.40	61.59	92.90	0.60
CMGC	RBF	500	0.04	14	91.13	83.96	95.94	0.82
TK	RBF	1	0.004	19	75.89	56.70	88.69	0.49

predictions was compared with that of the predictions made from consensus sequences (Table 2). The selection of an optimal parameter set for the SVMs and an optimal window size is described in the supplementary material. A consensus sequence for the CDK family was formulated based on the consensus sequences of CDC2, CDK5 and cdc2-like protein kinases (Beaudette *et al.*, 1993; Songyang *et al.*, 1994, 1996). The consensus sequences of other protein kinase families (CK2, PKA and PKC) were taken from PROSITE (Table 2) (Falquet *et al.*, 2002). For the familywise predictions made by the SVMs, the Ac value ranged from 83 to 95% and the CC value from 0.65 to 0.90 (Table 2). SVMs could predict the phosphorylation sites of kinase families with Ac values 8–17% higher than those achieved with consensus sequences.

If the catalytic domains of two protein kinases are homologous, they might phosphorylate similar amino acid sequences. For example, the PKA family belonging to the AGC group phosphorylates many sites recognized by the PKC family members that are also in the AGC group. For that reason, we thought that the SVMs trained with phosphorylation sites of a particular kinase group would be able to recognize the general pattern of phosphorylation sites in that group. For the groupwise predictions made by the SVMs, the Ac values were 76–91% and the CC values 0.49–0.82 (Table 3).

Comparison of the performances of kinase-specific predictions and kinase-non-specific predictions

We wanted to compare our system with another system—NetPhos. However, the accuracy of the NetPhos can be overestimated because NetPhos uses the data of the PhosphoBase and its training data are identical to some of our test data. Therefore, instead of comparing the PredPhospho with the NetPhos, we compared it with the kinase-non-specific method of the NetPhos using SVMs.

As described by Blom *et al.* (1999), all phosphorylation sites, except tyrosine residues, in PhosphoBase and SWISS-PROT are divided into serine and threonine sites. We used the sorted sites in training the SVMs and in evaluating their performance. The Ac value of the predictions was 81% for the serine residues and 77% for the threonine residues. The CC value was 0.59 for the serine residues and 0.52 for the threonine residues (Table 4). In contrast, the Ac values for the groupwise predictions of serine/threonine phosphorylation sites made by the SVMs were 80–91% and the CC values were 0.60–0.82 (Table 3). As the kinase-specific predictions can simultaneously predict the serine and threonine phosphorylation sites, all our systems including the prediction system for CAMK group provide more accurate predictions than the kinase-non-specific predictions do. These results are to be expected, because protein kinases are so diverse that

Table 4. The results of phosphorylation site predictions according to residues (serine or threonine)

Residue	Optimal kernel	Optimal C	Optimal γ	Optimal window size	Ac (%)	Sn (%)	Sp (%)	CC
Ser	Sigmoid	1	0.004	24	80.50	72.46	85.86	0.59
Thr	RBF	1	0.04	13	77.19	57.14	90.51	0.52

human protein kinase genes comprise 1.7% of known human genes (Manning *et al.*, 2002). Therefore, it may be difficult for the kinase-non-specific systems to accurately identify common patterns that the classifiers recognize as phosphorylation sites of all the kinases. In other words, the sequences of the substrates of one kinase group might act as noise in determining the substrate sequences of another kinase group. Kinase-specific predictions could provide the criteria for classifying the substrate sequences of diverse kinases and thus enhance the accuracy of the predictions.

Comparisons of PredPhospho and Scansite

We attempted to compare the PredPhospho and the Scansite. However, the Scansite provides phosphorylation site predictions for 63 kinases that are classified differently from the classification we used. Therefore, we cannot compare Scansite with our system across all the kinases analyzed with Scansite. For example, according to Menning's classification, which we adopted, the CDK kinase family consists of 24 kinase subfamilies (Manning *et al.*, 2002). However, Scansite provides predictions for three kinase subfamilies of the CDK family: CDC2 (cell-division-control protein 2), CDK5 and CDK2. For this reason, comparisons between the PredPhospho and the Scansite were restricted to two kinase families only, the CK2 and PKA families.

The results of predictions made with the Scansite are shown in Table 5. Our system gave Ac values $\sim 2\%$ higher for the CK2 family and $\sim 5\%$ higher for the PKA family than the values produced by the Scansite at low stringency (Tables 2 and 5).

Since these are the results for only two kinase families, we cannot say that our system is superior to the Scansite, but we suggest that the two systems can be used complementarily. The Scansite has been constructed from datasets that have been experimentally verified by that system's authors and provides predictions for diverse specific kinases. On the other hand, not only can PredPhospho predict the phosphorylation sites of the CK2 and PKA families with higher accuracy than Scansite can, but it can also be used to predict the phosphorylation sites of other kinases that cannot be analyzed with the Scansite. For example, Scansite can predict the phosphorylation sites for only three kinase subfamilies of the CDK family—CDC2, CDK5 and CDK2—whereas the PredPhospho can be used to predict the phosphorylation sites of all members of the CDK family.

Table 5. The results of phosphorylation site predictions for the CK2 and PKA families using Scansite

Kinase family	Stringency	Ac (%)	Sn (%)	Sp (%)	CC
CK2	High	63.29	7.92	100.00	0.17
	Medium	73.39	33.24	100.00	0.48
	Low	89.37	74.68	99.11	0.79
PKA	High	67.64	18.99	100.00	0.35
	Medium	78.35	46.99	99.25	0.57
	Low	84.85	67.63	96.30	0.69

Scansite has three levels of stringency: high, medium and low. High stringency involves low sensitivity and high specificity, whereas low stringency involves high sensitivity and low specificity.

Web program

This prediction system—PredPhospho—is available at <http://www.ngri.re.kr/proteo/PredPhospho.htm>. The program enables users to predict phosphorylation sites in protein sequences using SVMs or consensus sequences. The group-specific predictions made by the SVMs give more-or-less non-specific information on the kinds of kinases involved in phosphorylation. Predictions made using consensus sequences may help users guess the family of kinases predicted by SVMs.

DISCUSSION

Believing that kinases with similar kinase-activity domains recognize similar phosphorylation sites, we made predictions at both the family and group levels. PredPhospho predicts not only phosphorylation sites in proteins, but also the possible kinases that are involved in phosphorylation at those sites. Our prediction systems performed better than did kinase-non-specific predictions, and can be used compatibly with the Scansite, a Web program that is widely used to predict phosphorylation sites (Yaffe *et al.*, 2001).

This study had several shortcomings. The most critical limitation of our study is that all non-annotated serine, threonine and tyrosine residues are considered to be (–) sites. Although phosphorylated residues make up a minority of all serine, threonine and tyrosine residues in proteins and false (–) sites are reduced by the family-specific and group-specific predictions, some of the (–) sites in the study might be determined

to be (+) sites in the future. These false (−) sites act as noise in training the SVMs and add bias to the assessment of prediction accuracy. This problem may be addressed by experiments that accumulate large numbers of unique sequences that are not phosphorylated by protein kinases, although these experiments are labor-intensive and expensive. Second, we thought that SVMs trained with the phosphorylation sites of a group of kinases would find common patterns in the phosphorylation sites of that group. However, the numbers of data entries in a group are unequally distributed among the kinase families. In one case, data from the PKA and PKC families made up over 90% of the AGC group data. SVMs trained with such data can predict the sites of the PKA and PKC families better than those of other families in the AGC group. Third, we regarded serine, threonine and tyrosine residues as only two classes: (+) and (−) sites. However, (+) sites should be subdivided by individual criteria on the basis of efficiency of phosphorylation.

To use the system in proteome analyses and to make it more useful, the system should be able to ensure greater accuracy and predict the phosphorylation sites of other groups. Biological knowledge related to phosphorylation, improvements in the SVM algorithm, and the collection of more data with which to train the SVMs will increase the performance of the prediction system. For instance, with information about the cellular localization of a specific kinase, proteins in other cellular compartments can be ruled out in phosphorylation site predictions (Blom *et al.*, 1999). The surface accessibility calculated by protein structures can help eliminate false (+) sites that occur in the internal regions of proteins (Yaffe *et al.*, 2001). Other SVM kernel functions that are more suitable for peptide analysis of kinase recognition sites also need to be developed to ensure greater accuracy of the prediction system. Furthermore, for the purpose of predicting phosphorylation sites of other kinase groups, more data sets on phosphorylation sites with their respective kinase information are necessary. Collecting these data can be accomplished by intensively reviewing papers or by developing a natural language processor.

Our system can be used in the analysis of proteins for various purposes. For example, the system can be applied to the analysis of protein functions and the effects of non-synonymous single nucleotide polymorphisms (SNPs) on phosphorylation. Information about kinds of kinases at phosphorylation sites can be used to predict whether a protein might be involved in some signal transduction pathways and can give clues as to the possible functions of the proteins. Moreover, changes in amino acids near phosphorylation sites can alter kinase kinetics and may result in the substitution of kinases that recognize those sites with other kinases. Substitutions of amino acids can also add or remove phosphorylation sites.

These variations within or between species that are related to phosphorylation sites could cause phenotypic variations or disease. The PredPhospho can thus help us predict such effects of protein variations.

ACKNOWLEDGEMENTS

We thank Dr Ki-Bong Kim for his thoughtful advice.

REFERENCES

- Baldi,P. (2001) *Bioinformatics: The Machine Learning Approach*, 2nd edn. The MIT Press, Cambridge.
- Beaudette,K.N., Lew,J. and Wang,J.H. (1993) Substrate specificity characterization of a cdc2-like protein kinase purified from bovine brain. *J. Biol. Chem.*, **268**, 20825–20830.
- Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.
- Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
- Falquet,L., Pagni,M., Bucher,P., Hulo,N., Sigrist,C.J.A., Hofmann,K. and Bairoch,A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.*, **30**, 235–238.
- Kecman,V. (2001) *Learning and Soft Computing*. The MIT Press, Cambridge.
- Kreegipuu,A., Blom,N. and Brunak,S. (1999) PhosphoBase, a database of phosphorylation sites: release 2.0. *Nucleic Acids Res.*, **27**, 237–239.
- Manning,G., Whyte,D.B., Martinez,R., Hunter,T. and Sudarsanam,S. (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.
- Pinna,L.A. and Ruzzene,M. (1996) How do protein kinases recognize their substrates? *Biochim. Biophys. Acta*, **1314**, 191–225.
- Songyang,Z., Blechner,S., Hoagland,N., Hoekstra,M.F., Piwnicka-Worms,H. and Cantley,L.C. (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.*, **4**, 973–982.
- Songyang,Z., Lu,K.P., Kwon,Y.T., Tsai,L.H., Filhol,O., Cochet,C., Brickey,D.A., Soderling,T.R., Bartleson,C., Graves,D.J. *et al.* (1996) A structural basis for substrate specificities of protein Ser/Thr kinases: primary sequence preference of casein kinases I and II, NIMA, phosphorylase kinase, calmodulin-dependent kinase II, CDK5, and Erk1. *Mol. Cell. Biol.*, **16**, 6486–6493.
- Vapnik,V. (1995) *The Nature of Statistical Learning Theory*. Springer, New York.
- Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.
- Yaffe,M.B., Leparo,G.G., Lai,J., Obata,T., Volinia,S. and Cantley,L.C. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.*, **19**, 348–353.