W226–W229 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki471

KinasePhos: a web tool for identifying protein kinase-specific phosphorylation sites

Hsien-Da Huang*, Tzong-Yi Lee, Shih-Wei Tzeng[1] and Jorng-Tzong Horng[1,2]

Department of Biological Science and Technology, Institute of Bioinformatics, National Chiao Tung University, Hsin-Chu 300, Taiwan, [1]Department of Life Science and [2]Department of Computer Science and Information Engineering, National Central University, Chung-Li 320, Taiwan

Received February 13, 2005; Revised and Accepted April 15, 2005

ABSTRACT

KinasePhos is a novel web server for computationally identifying catalytic kinase-specific phosphorylation sites. The known phosphorylation sites from public domain data sources are categorized by their annotated protein kinases. Based on the profile hidden Markov model, computational models are learned from the kinase-specific groups of the phosphorylation sites. After evaluating the learned models, the model with highest accuracy was selected from each kinase-specific group, for use in a web-based prediction tool for identifying protein phosphorylation sites. Therefore, this work developed a kinase-specific phosphorylation site prediction tool with both high sensitivity and specificity. The prediction tool is freely available at http://KinasePhos.mbc.nctu.edu.tw/.

INTRODUCTION

Protein phosphorylation, performed by a group of enzymes known as kinases and phosphotransferases, is a post-translational modification essential to correct functioning within cells (1). The post-translational modification of proteins by phosphorylation is the most abundant form of cellular regulation, affecting many cellular signal pathways, including metabolism, growth, differentiation and membrane transport (2). The enzymes must be specific and act only on a defined subset of cellular targets to ensure signal fidelity.

Because of owing to its importance in cellular control, a computational scheme to quickly and efficiently identify phosphorylation sites in protein sequences and the catalytic kinases involved in the phosphorylation is desirable. Such a tool would improve the efficiency of characterization of new protein sequences. Therefore, in this work, a prediction method was designed and implemented to facilitate the identification of the phosphorylation sites and the related catalytic kinases.

NetPhos (2), DIPHOS (3) and Berry *et al*. (1) presented several prediction methods for identifying the phosphorylation site prediction concentrating on only the substrate specificity. NetPhosK (4) is an artificial neural network algorithm to identify protein kinase A (PKA) phosphorylation sites with 100% sensitivity and 40% specificity in experiments. Scansite 2.0 (5) identified short protein motifs that are recognized by phosphorylation protein serine/threonine or tyrosine kinases. Each motif used in the Scansite was constructed from a set of experimentally validated phosphorylation sites and was represented as a position-specific scoring matrix. Rather than search a protein motif of phosphorylation substrate against the target sequences based on the homolog to the motifs, the KinasePhos web server developed here was based on the concept of machine learning, the same as NetPhos and DIPHOS. Computer models were trained for the detection of phosphorylation sites. By comparison of the prediction accuracy between the predictive computer model methods and the motif search tools, the predictive computer models contribute more specificity for the detection of phosphorylation sites.

The proposed scheme considers the catalytic kinases of protein phosphorylation. The known phosphorylation sites from data sources in public domain were categorized by their annotated protein kinases. Based on the profile hidden Markov model (HMM), computational models were determined from the kinase-specific groups of the phosphorylation sites. A web-based prediction application was implemented to facilitate the identification of protein kinase-specific phosphorylation sites.

MATERIALS AND METHODS

The PhosphoBase (6) consists of 1883 experimentally verified phosphorylation sites within 597 protein entries. The number of serine, threonine and tyrosine sites is 984, 246 and 653, respectively. Swiss-Prot (7) (release 45 of October 2004) maintains 163 500 protein entries, of which 3614 have phosphorylation annotation. Among these entries, the number of serine, threonine and tyrosine sites was 1005, 281 and 321,

*To whom correspondence should be addressed. Tel: +886 3 5712121, ext. 56952; Fax: +886 3 5729288; Email: bryan@mail.nctu.edu.tw
Correspondence may also be addressed to Jorng-Tzong Horng. Tel: +886 3 4227151, ext. 35307; Fax: +886 3 4222681; Email: horng@db.csie.ncu.edu.tw

respectively. Generally, the serine, threonine and tyrosine, which are not annotated as phosphorylation residues, within the experimentally validated phosphorylated proteins, are selected as negative sets, i.e. the non-phosphorylated sites. Therefore, two negative (non-phosphorylated) datasets were obtained from the PhosphoBase and Swiss-Prot based on the phosphorylation annotation. Because of the absence of good negative dataset exists for non-phosphorylated sites, the residues that had not been previously annotated as phosphorylated in phosphorylation annotated proteins were chosen as a reflection of more general non-phosphorylated sites. Supplementary Table S1 summarizes the statistics of kinase-specific phosphorylated sites used for learning models in the proposed application. This work confirms the existence of two major protein kinases phosphorylating either at serine/threonine residues or at tyrosine residues.

Figure 1 depicts a flowchart of the proposed method. Phosphorylated sites were first extracted as positive sets; non-phosphorylated sites were extracted as negative sets, and the catalytic kinase annotations were obtained from PhosphoBase and Swiss-Prot. The positive sets were then categorized by catalytic kinases. Alternatively, in larger positive groups, the sequences of the phosphorylated sites can be clustered into subgroups by maximal dependence decomposition (MDD) (8). The MDD was first applied in nucleotides and is a recursive process to divide a sequence set into tree-like subgroups based on the positional dependency of the sequences. Here, we applied the MDD to group protein phosphorylation substrates into subgroups. As the example given in Figure 1, 232 phosphorylation serine substrates are grouped into subgroups. When applying MDD to cluster the sequences of a positive set, a parameter, i.e. the minimum-cluster-size, should be set. If the size of a subgroup is less than the minimum-cluster-size, the subgroup is terminated to be divided. The MDD process terminates until all the subgroup sizes are less than the minimum-cluster-size.

Thereupon, the concept of the profile HMM was adopted to learn computational models from positive sets of phosphorylation sites. To evaluate the learned models, k-fold cross-validation and leave-one-out cross-validation were performed on them. After evaluating the models, the model with highest accuracy in each dataset was chosen.

For each kinase-specific positive set of the phosphorylated sites, the best performed model is selected and used to identify the phosphorylation sites within the input protein sequences by HMMsearch (9). To search the hits of a model, HMMER returns both a HMMER bit score and an expectation value (*E*-value). The HMMER bit score is used as the criterion to define a HMM match. We select the HMMER score as the criterion to define a HMM match. A search of a model with the HMMER score greater than the threshold t is defined as a positive prediction, i.e. a HMM recognizes a phosphorylation site. The threshold t of each model is decided by maximizing the accuracy measure during a variety of cross-validations with the HMM bit score value range from 0 to −10. For example, Supplementary Figure S1 depicts the optimization of the threshold of the HMM bit scores in the S_PKA model. The threshold of the S_PKA model is set to −4.5 to maximize the accuracy measure of the model.

When considering a MDD-clustered dataset, for example, MDD-clustered PKA catalytic serine (S_PKA), the HMMs are trained separately from the subgroups of the phosphorylated sites resulted by MDD. Each model is used to search in the given protein sequences for the phosphorylated sites. A positive prediction of a model group is defined by at least one of the models that makes a positive prediction, whereas

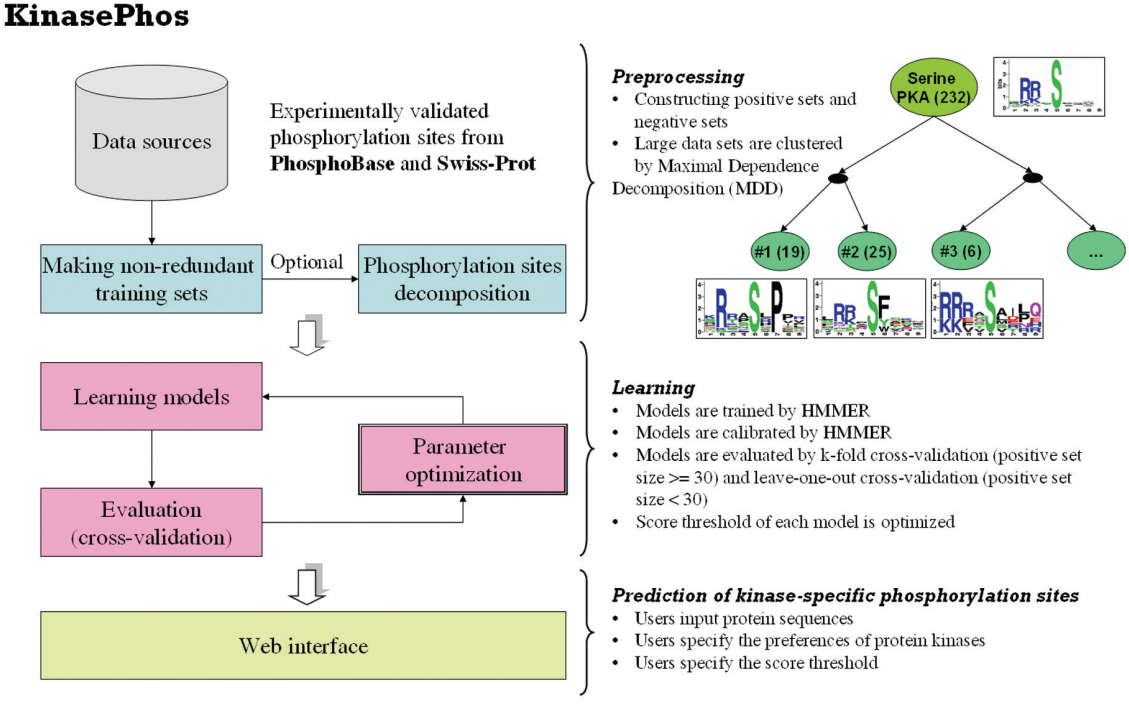

Figure 1. The flow of the proposed scheme.

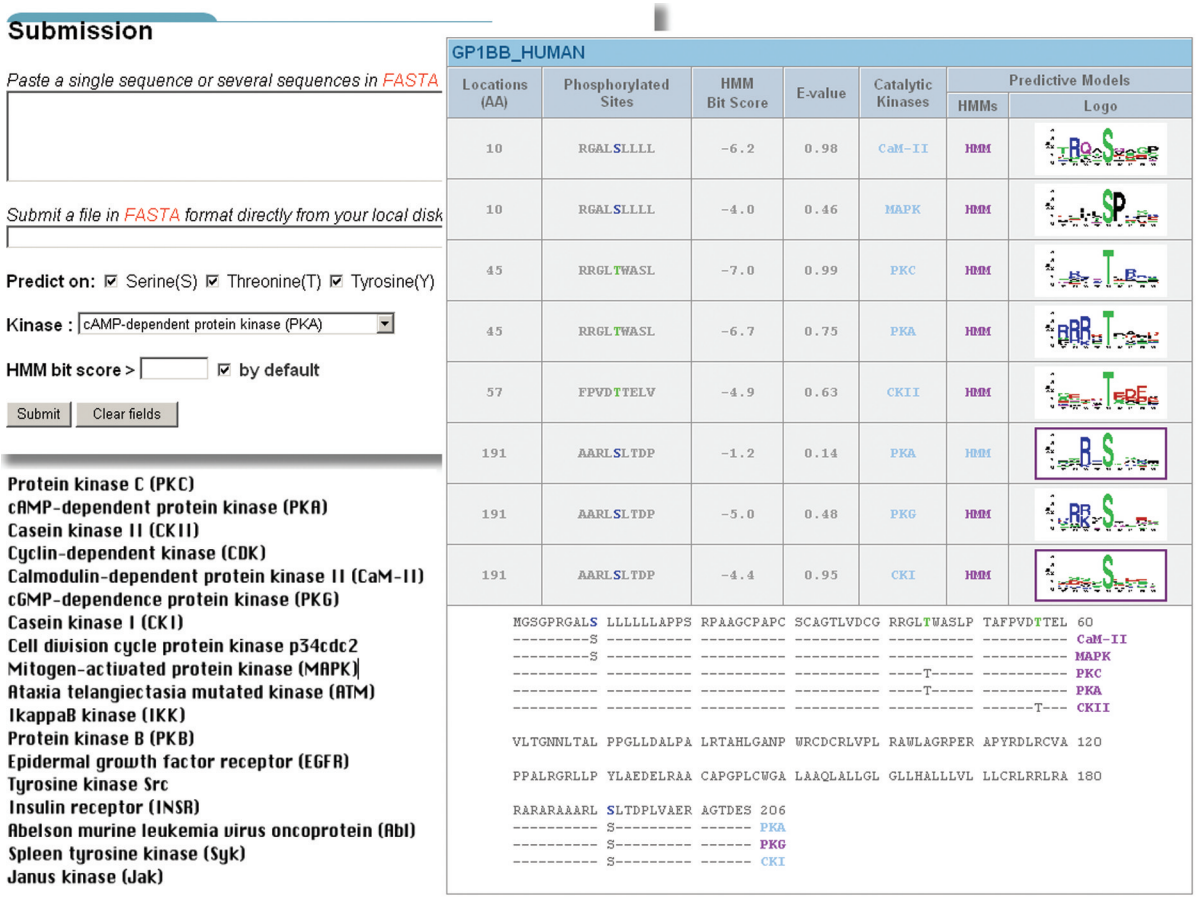

Figure 2. The KinasePhos web interface.

a negative prediction is defined as all the models that make negative predictions.

USAGE

The predictive models for protein phosphorylation sites were learned and evaluated using experimentally validating datasets from Swiss-Prot (7) and PhosphoBase (6). The chosen models were collected and provided for the web-based predicting services. The web server can accurately and efficiently predict the kinase-specific phosphorylation sites in the input protein sequences. As depicted in Figure 2, users can submit their uncharacterized protein sequences to the query interface and choose the kinase-specific models to predict phosphorylated serine, threonine or tyrosine. The web server locates the predictive phosphorylation sites and the involved catalytic protein kinases. In order to reveal the characteristics of the phosphorylated sites, including the phosphorylated residues and surrounding sequences, the training phosphorylation sites and constructed sequence logos corresponding to each protein kinase are also provided on the application interfaces on the web.

PREDICTION ACCURACY

The following measures of the predictive performance of the models are then calculated: precision = TP/(TP+FP), sensitivity = TP/(TP+FN), specificity = TN/(TN+FP) and the accuracy = (sensitivity + specificity)/2, where TP, TN, FP and FN are true positive, true negative, false positive and false negative predictions, respectively. Especially, we make the equal sizes of the positive samples and the negative samples during the cross-validation processes. For each kinase-specific group, the most accurate model was chosen and applied in the web application, as presented in Table 1. For example, the precision, sensitivity, specificity and accuracy of the model learned from the MDD-clustered S_PKA dataset, which was constructed by the combined PhosphoBase and Swiss-Prot data sources, were 0.85, 0.91, 0.84 and 0.88, respectively. The average precision, sensitivity, specificity and accuracy of all the kinase-specific serine models were 0.88, 0.84, 0.88 and 0.86, respectively.

The prediction accuracy of the proposed web application was compared with several previously developed phosphorylation prediction tools, such as NetPhos (2), DISPHOS (3) and rBPNN (1). Previous tools did not consider the catalytic kinase annotations. In particular, this work constructed the kinase-specific models for phosphorylation sites. Only the average accuracy from the best model was chosen in each kinase-specific model. As presented in Table 2, the average accuracy of the models learned from the serine, threonine and tyrosine sets was 0.86, 0.91 and 0.84, respectively. The average accuracy for KinasePhos was 0.87. All the accuracies of the serine, threonine and tyrosine models were higher than those from NetPhos and DISPHOS.

Table 1. The selected models learned and used in the web server

Residues	Protein kinases	Score threshold	Precision	Sensitivity	Specificity	Accuracy
Serine	S_PKA[a] (232)	−4.5	0.85	0.91	0.84	0.88
	S_PKC[a] (176)	−4.5	0.87	0.77	0.88	0.82
	S_PKG (27)	−9.5	0.94	0.96	0.93	0.95
	S_PKB (37)	−6.5	0.88	0.76	0.89	0.82
	S_CaM-II (37)	−8.0	0.84	0.76	0.86	0.81
	S_CKI (30)	−7.0	0.82	0.65	0.86	0.76
	S_CKII[a] (85)	−3.5	0.95	0.79	0.96	0.87
	S_cdc2 (43)	−10	0.94	0.94	0.94	0.94
	S_MAPK (27)	−6.0	0.97	0.77	0.97	0.87
	S_CDK[a] (71)	−6.5	0.83	0.87	0.82	0.85
	S_ATM (38)	−8.0	0.92	0.87	0.92	0.90
	S_IKK (32)	−8.0	0.75	0.75	0.75	0.75
	Average		0.88	0.84	0.88	0.86
Threonine	T_PKA (19)	−7.0	0.97	0.94	0.97	0.95
	T_PKC (37)	−8.5	0.85	0.83	0.85	0.84
	T_CKII (17)	−9.0	0.79	0.98	0.75	0.86
	T_cdc2 (23)	−9.5	1.00	0.95	1.00	0.97
	T_MAPK (15)	−9.5	1.00	1.00	1.00	1.00
	T_CDK (35)	−6.5	0.94	0.86	0.94	0.90
	Average		0.91	0.92	0.91	0.91
Tyrosine	Y_EGFR (30)	−5.5	0.89	0.83	0.89	0.86
	Y_INSR (16)	−9.5	0.82	0.78	0.83	0.80
	Y_Src (28)	−5.0	0.86	0.81	0.87	0.84
	Y_Abl (27)	−2.0	0.93	0.48	0.96	0.72
	Y_Syk (22)	−8.5	0.83	0.91	0.82	0.86
	Y_Jak[a] (42)	−3.5	0.91	0.66	0.93	0.80
	Average		0.86	0.81	0.87	0.84

[a]The dataset is clustered by MDD.

Table 2. The prediction accuracy comparison between NetPhos, DISPHOS, rBPNN and KinasePhos

Residue types	NetPhos	DISPHOS	rBPNN	KinasePhos
Serine	0.69	0.75	No data	0.86
Threonine	0.72	0.80	No data	0.91
Tyrosine	0.61	0.82	No data	0.84
Total or average	0.67	0.79	0.87	0.87

FUTURE DEVELOPMENT

Prospective works to enhance the accuracy of predictive schemes are addressed as follows. First, the species-specific phosphorylation sites can be considered to evaluate protein phosphorylation-related mechanisms in different organisms, possibly improving the accuracy of the models learned from a species-specific dataset. Second, protein structural properties, such as solvent accessibility, of the phosphorylated sites can be considered to reduce the number of false positive predictions of phosphorylated sites located in buried regions. For proteins with known structures, the solvent accessibility of a phosphorylated site can be calculated trivially. However, as to the proteins without structures, the solvent accessibility of a residue should be computationally determined.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

The authors would like to thank the National Science Council of the Republic of China for financially supporting this research under Contract No. NSC 93-2213-E-008-024 and Contract No. NSC-93-2213-E-009-075. Funding to pay the Open Access publication charges for this article was provided by National Science Council of the Republic of China.

Conflict of interest statement. None declared.

REFERENCES

1. Berry,E.A., Dalby,A.R. and Yang,Z.R. (2004) Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms. *Comput. Biol. Chem.*, **28**, 75–85.
2. Blom,N., Gammeltoft,S. and Brunak,S. (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. *J. Mol. Biol.*, **294**, 1351–1362.
3. Iakoucheva,L.M., Radivojac,P., Brown,C.J., O'Connor,T.R., Sikes,J.G., Obradovic,Z. and Dunker,A.K. (2004) The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.*, **32**, 1037–1049.
4. Hjerrild,M., Stensballe,A., Rasmussen,T.E., Kofoed,C.B., Blom,N., Sicheritz-Ponten,T., Larsen,M.R., Brunak,S., Jensen,O.N. and Gammeltoft,S. (2004) Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. *J. Proteome Res.*, **3**, 426–433.
5. 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.
6. Blom,N., Kreegipuu,A. and Brunak,S. (1998) PhosphoBase: a database of phosphorylation sites. *Nucleic Acids Res.*, **26**, 382–386.
7. Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I. *et al.* (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.
8. Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.
9. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.