Bioinformatics and Biosystems 2007, Vol. 2, No. 1, pp. 28-32

Prediction of Protein Kinase Specific Phosphorylation Sites with Multiple SVMs

Wonchul Lee[1] and Dongsup Kim[1]

[1] Department of Biosystems., Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, Republic of Korea 305-701

Abstract

The protein phosphorylation is one of the important processes in the cell signaling pathway. A variety of protein kinase families are involved in this process, and each kinase family phosphorylates different kinds of substrate proteins. Many methods to predict the kinase-specific phosphoryrated sites or different types of phosphorylated residues (Serine/Threonine or Tyrosin) have been developed. We employed Supprot Vector Machine (SVM) to attempt the prediction of protein kinase specific phosphorylation sites. 10 different kinds of protein kinase families (PKA, PKC, CK2, CDK, CaM-KII, PKB, MAPK, EGFR) were considered in this study. We defined 9 residues around a phosphorylated residue as a deterministic instance from which protein kinases determine whether they act on. The subsets of PSI-BALST profile was converted to the numerical vectors to represent postive or negative instances. When SVM training, We took advantage of multiple SVMs because of the unbalanced training sets. Representative negative instances were drawn multiple times, and generated new traing sets with the same positive instances in the original traing set. When testing, the final decisions were made by the votes of those multiple SVMs. Generally, RBF kernel was used for the SVMs, and several parameters such as gamma and cost factor were tested. Our approach achieved more than 90% specificity throughout the protein kinase families, while the sensitivities recorded 60% on average.

Introduction

Proteins are usually phosphorylated on their specific residues such as Serine, Threonine, and Tyrosin after their synthesis. The phosphorylation plays crucial roles in a variety of biological cellular processes, including transcription, translation, cell cycle and signal transduction.

If potential phosphorylated sites and involved protein kinases could be revealed, it would greatly help extend our knowledge on the biological cellular processes. Already known phosphorylation sites can be divided into the sites with known protein kinases acting on them and the sites with no such information available. Many researches on the prediction of phosphorylation sites have been done. Some of them focused on the specific substrate residues (prediction for Serine/Threonine or Tyrosin), while others approached in terms of protein kinase family or group (prediction for the sites catalyzed by PKA or CDK).

Corresponding author: Dongsup Kim (Tel: +82-42-869-4317, Fax:+82-42-869-4310, Email: kds@kaist.ac.kr)
This work is supported by CHUNG Moon Soul center for BioInformaion and BioElectronics (CMSC).

Generally, local sequence patterns (consensus sequences or motifs) and profiles were used. Sequence patterns are derived by aligning the local regions of proteins that contain phosphorylated residues. In the profile method, pre-compiled profile (or weight matrix) is compared with a target protein sequence, and similarity score is driven. The profile is constructed by aligning only phosphorylated sequences. Scansite (Yaffe et al., 2001) utilized this profile approach, and correctly predicted ~70% of known phosphorylation sites in PhosphoBase.

On the other hand, machine learning techniques also have been implemented. NetPhos(Blom N et al., 1999) is implemented in the artificial neural networks (ANNs) with the consensus-motif-based approaches. The improved version, NetPhosK can perform PK-specific predictions as well. Support vector machine (SVM)-based method was also developed and implemented in PredPhospho (J.H.Kim et al., 2004). PredPhospho can predict the phosphorylation sites by 8 kinds of different protein kinase families and groups, and performs well both in specificity and sensitivity. It attempted to optimize the system by identifying SVM parameters such as gamma and penalty parameter, kernel type, and window size that maximize the performance.

Here, we also attempted the PK-specific phosphorylation

sites prediction with the SVM. We used a subset of Psi-Blast profile as features to include the evolutionary information. In addition, decisions were made by the votes of multiple SVMs that were trained with different negative instances. This guarantees relatively higher specificity than the sensitivity, because the system can experience many different negative instances by constructing multiple SVMs when training.

Materials and Methods

Dataset

SwissProt protein sequences with annotated phosphorylation sites were obtained from Phospho.ELM database v5.0 (Francesca Diella et al., 2004). This database is a collection of experimentally verified serine, threonine, and tyrosin sites in eukaryotic proteins. It contains 7,071 phosphorylation sites in total, and more than 60 kinds of protein kinases annotated.

Dariusz Plewczynski et al.(2005) suggested that sequence specificity determinants in the phosphorylation process are not that strict, they are located within a 9-amino acid segment around a phosphorylated site. We therefore set the window size to 9, which has 4-upstream and 4-downstream residues of the phosphorylated sites.

On the other hand, non-phosphorylated serine/threonine or tyrosin are needed as negative controls in order to use machine-learning techniques such as SVM and the neural network. We defined those serine/threonine or tyrosin that are not annotated in the sequences as non-phosphorylated sites. For example, CDK (cyclin-dependent kinase) phosphorylates serine or threonine residues in its substrate. Thereby, all the non-annotated serine or threonine residues in the substrate can be regarded as non-phosphorylated sites.

Table 1. Data Statistics of of the database. The ratios between positive and negative instances are large. Among more than 60 kinds of kinases in the database, only 10 has more than 50 verified phosphorylation sites.

PK Family	Num. of Positives	Num. of Negatives	Ratio
PKA_group	254	14622	1:58
PKC_group	249	10584	1:43
CK2	237	8071	1:34
CDK	106	3874	1:37
Src	96	1341	1:14
CDK1	85	4353	1:51
CaM-KII_group	57	4541	1:80
PKB_group	55	5422	1:99
MAPK_group	52	2827	1:54
EGFR	50	383	1:8

Dataset Statistics and Protein Kinase Selection

We excluded all the phosphorylated serine/threonine or tyrosin that do not have 8 neighboring residues because they are located around the sequence terminals. Table 1 shows the number of positive and negative instances of protein kinase families examined in our study.

In the database, there are more than 60 kinds of annotated protein kinases. However, we only chose those 10 proteins (PKA, PKC CK2, CDK, Src, CDK1, CaM-KII, PKB, MAPK, EGFR), because other kinases does not have enough known phosphorylation sites. In order to make reliable SVMs, we limit the minimum size of positive instances for each protein kinase to 50, but other kinases can be included if more positive instances are available in the future. Table 1 also shows ratios between positive and negative instances of each kinase. The number of negative instances outweighs the number of positive instances in all cases. Those unbalances make it difficult to train SVMs, because the SVMs trained in favor of negative instances would predict almost all the positive instances incorrectly. To solve this problem, we used the SVM parameter (cost factor) and multiple-SVM system. The explanation about this will be followed.

Profile Feature Extraction

In order to make the SVM input format, positive and negative instances should be represented as numerical vectors. For each positive or negative instance, we ran PSI-BLAST (Altschul et al., 1997) with the sequence, from which the instance is derived, against NR90 database (More than 90% homology-reduced non-redundant protein database). A resultant profile was obtained after 3 iterations, and a subset of values corresponding to the instance was retrieved. As a result, 180-dim-vectors were generated to represent 9-long-instances.

Multiple-SVM Training Procedure

We used SVM-Light (T. Joachims, 1999) with RBF kernel as a default. There were two parameters that we could take into consideration: gamma for the RBF kernel and cost factor ratio (Morik et al., 1999) by which training errors on positive instances outweight errors on negative instances. Previous works on phosphorylation sites prediction suffers from relatively lower sensitivity than specificity. We included this parameter and set it to larger than the default (1) to improve

sensitivity of our SVMs.

For the SVM parameters selection, we performed 3-fold cross validation and multiple-SVM training method. First, positive and negative instances were randomly divided into training and test sets in a 2:1 ratio. Then, for each pair of training and test set, negative instances in the training set were grouped into 1.5~2 times the number of positive instances in the training set. The grouping procedure was carried out using BLOSUM62 scoring matrix between a pair of negative instances, and it is needed to balance the number of positive and negative instances when training. In our previous experiment, training with a balanced set resulted in a better performance. Once the negative instances are clustered, randomly chosen representative negative instances from each group become a training set with the positive instances from the original training set. This procedure is repeated multiple times to construct multiple SVMs. Then multiple (usually odd number) SVMs perform predictions on the instances in the test set. If more than half of the SVMs produce non-negative raw scores for an instance, the instance is regarded as a positive one. For 3-fold cross validation, this process is applied to 3 pairs of training and test set. Finally, median MCC (Matthew's correlation coefficient) value is recorded. MCC value ranges from -1 to 1, and the values near to 1 mean a good result in terms of specificity and sensitivity.

$$MCC = \frac{(TP*TN)-(FN*FP)}{\sqrt{(TP+FN)*(TN+FP)*(TP+FP)*(TN+FN)}}*100\%$$

To search for the optimal SVM parameters, we repeated the 3-fold cross validations while changing the parameters: gamma (0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1, 2) and cost factor (1~3).

A pair of parameters that produced a maximum median MCC value were considered as the optimal parameters, and assigned to each protein kinase family.

Assessment Scheme

In order to measure the performance of our approach, we used three indicators: accuracy (Acc), specificity (SP), and sensitivity (SN).

$$Acc = \frac{TP+TN}{TP+FP+TN+FN}*100\%,$$

$$SP = \frac{TN}{TN+FP}*100\%,$$

$$SN = \frac{TP}{TP+FN}*100\%,$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives and FN is the number of false negatives. The accuracy Acc provides an overall correctness of the prediction, while the specificity SPmeasures a power to identify negatives, and the SN indicates how well positives are predicted as positives.

RESULTS and DISCUSSION

Prediction of phosphorylation sites recognized by 10 protein kinase families

Phosphorylation sites recognized by 10 protein kinase fami-

Table 2. The optimal parameters and the performance with those parameters of 10 protein kinase famlilies. Each kinase family required different *r* (gamma) parameters, while the cost factor ratios did not change much. The accuracy, specificity, and sensitivity were measured. The MCC values are dependent on the number of positive and negative instances in the test set. Relatively low MCC values were obtained because there were much more negative instances than positive instances in our test sets.

PK Family	Optimal *r*	Optimal Cost Factor Ratio	Acc (%)	SP (%)	SN (%)	MCC
PKA_group	2.0	1.0	95.86	96.44	61.90	0.3618
PKC_group	1.0	1.0	92.70	93.63	53.01	0.2667
CK2	1.0	1.0	91.65	92.48	63.29	0.3228
CDK	0.5	1.0	89.20	89.44	80.55	0.3446
Src	0.5	1.0	89.53	92.82	43.75	0.3099
CDK1	1.0	1.0	92.96	93.59	60.71	0.2820
CaM-KII_group	1.0	1.0	97.91	98.74	31.57	0.2648
PKB_group	0.5	1.5	93.47	93.68	72.22	0.2559
MAPK_group	0.01	1.0	93.31	94.04	52.94	0.2466
EGFR	0.01	2.0	66.43	63.77	87.50	0.3275

lies were predicted by SVMs with the RBF kernel. The optimal SVM parameters (gamma and cost factor), accuracy, specificity, and sensitivity were measured to evaluate the performance. For the system of multiple SVMs, we implemented 5 SVMs so that 3 votes are needed to claim that an instance is positive. Table 2 shows the result.

The predictions for the 8 protein kinase families out of 10 recorded more than 90% specificity, while sensitivities range from 31.57% to 87.50%. The optimal gamma parameters were different for each protein kinase family, while the cost factors did not show big differences throughout the kinase families. As stated, our system has advantages in that it pursues the enhanced performance on specificity by utilizing multiple SVMs and it simplified the prediction procedure by applying the common options such as SVM kernel function and window size to all the protein kinase families examined in our study.

Performance Comparison with the another work

PredPhospho (J.H.Kim et al., 2004) also employed SVMs to predict PK-specific phosphorylation sites. Thus it can be useful to compare our results with the performance of PredPhospho. However, PredPhospho only used 4 kinds of PK families and it optimized the system by considering all the possible variables: SVM kernel function, gamma, penalty parameter, and even window size. Hence, it is difficult to directly compare the two methods. Table 3 below shows the comparison of the two approaches for the 4 protein kinase families (found in both methods).

In the case of CDK, PredPhospho reached the excellent performance with the window size 18 and Sigmoid kernel function. Our method for CDK recorded relatively lower performance than PredPhospho in sensitivity. The window size 9 might not be enough to distinguish the positive and negative instances for CDK. On the other hand, our method for the PKC resulted in the similar or slightly better performance than PredPhospho.

Src and EGFR examined in our study are sub-families of Tyrosin Kinase (TK). The result of the kinase families was compared with the performance of TK group done by PredPhospho, and listed in table 4.

Extraction of Negative Instances

When extracting negative instances from the database, we assumed that all the non-annotated serine/threonine or tyrosin are non-phosphorylated residues. However, those residues can be proved to be positive sites in future experiments. In addition, the protein kinases that phosphorylate serine or threonine do not act on those two kinds of residues at the same rate. For example, protein kinase A (PKA) phosphorylates 241 serines and 22 threonine out of 263 in total. The ratio between serine and threonine is 11:1, but we considered all the non-annotated threonines as negative sites. If we take this prior-probability into account, we can reasonably reduce the number of negative instances, which leads to imporve the sensitivity.

Table 3. Performance Comparison with PredPhspho.

	PK family	Kernel Function	Feature Method	Window Size	Acc(%)	SP(%)	SN(%)
Our approach	CDK	RBF	Profile	9	89.20	89.44	80.55
PredPhospho		Sigmoid	BIN	18	95.09	95.10	95.02
Our Approach	CK2	RBF	Profile	9	91.65	92.48	63.29
PredPhospho		RBF	BIN	10	91.47	96.43	83.90
Our Approach	PKA	RBF	Profile	9	95.86	96.44	61.90
PredPhospho		RBF	BIN	5	89.98	91.11	88.32
Our Approach	PKC	RBF	Profile	9	92.70 (85.49)	93.63 (85.58)	53.01 (81.92)
PredPhospho		Sigmoid	BIN	11	82.9	85.79	78.71

Table 4. Src and EGFR versus TK in PredPhospho.

	PK family	Kernel Function	Feature Method	Window Size	Acc(%)	SP(%)	SN(%)
Our approach	CDK	RBF	Profile	9	89.53	92.82	43.75
PredPhospho					66.43	63.77	87.50
PredPhospho	TK	RBF	BIN	19	75.89	88.69	56.70

Structural Features

We only referred to the sequence information (profile) in making feature vectors. However, to improve the prediction performance and to apply this approach to the protein kinase families in which only small number of phosphorylated sites are known, structural features should be incorporated in the system. Phosphorylated residues are susceptible to be located on the surface of the proteins. In addition, secondary strucures can influence the phosphorylation process. Recently, Kunpeng Zhang et al, 2006 used the solvent accessibility and the secondary structure prediction result as features for the training of neural network. More detailed structural constraints, from which the phosphorylation is determined, also can be identified if the proteins with known structure like PDB can be analyzed.

Conclusion

In our work, we performed the protein kinase specific phosphorylation sites prediction for 10 kinds of kinase families with SVMs. While sensitivities for several kinase families were lower than those from the similar approach with SVM, we reached the evenly good performance throughout the kinase families in terms of specificity. In order to solve the unbalance problem between the number of positive and negative instances, representative negative instances were extracted several times to make multiple negative sets, and the final decisions were made by the voting system. For better phosphorylation sites prediction system, more careful extraction of negative instances and other structural features will be needed.

REFERENCES

[1] Altschul et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

[2] Blom N, Gammeltoft S, Brunak S, (1999) Sequence and structure-based prediction of eukaryotic protein phosphorylation sites, J Mol Biol, 294(5):1351-1362.

[3] Dariusz Plewczynski et al. (2005) A support vector machine approach to the identification of phosphorylation sites, Cellular & Molecular biology letters, 10:73-89.

[4] Francesca Diella et al. (2004) Phospho.ELM: A database of experimentally verified phosphorylation sites in eukaryotic proteins. BMC Bioinformatics, 5:79.

[5] Jong Hun Kim, Juyoung Lee, Bermseok Oh, Kuchan Kimm and InSong Koh, (2004) Prediction of phosphorylation sites using SVMs, Bioinformatics, 20(17):3179-3184.

[6] K. Morik, P. Brockhausen, and T. Joachims, (1999) Combining statistical learning with a knowledge-based approach - A case study in intensive care monitoring. Proc. 16th Int'l Conf. on Machine Learning (ICML-99).

[7] Kunpeng Zhang et al., (2006) Using a Neural Netwoking Method to Predict the Protein Phosphorylation Sites with Specific Kinase, LNCS 3973, pp. 682-689.

[8] T. Joachims, (1999) Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Schlkopf and C. Burges and A. Smola (ed.), MIT-Press.

[9] Yaffe et al. (2001) A motif-based profile scanning approach for genome-wide prediction of signaling pathways. Nature Biotechnology., 19:348-353.