다중 커널 학습을 이용한 단백질의 인산화 부위 예측

김종경, 최승진

포항공과대학교 컴퓨터공학과

{blkimjk, Seungjin}@postech.ac.kr

Prediction of phosphorylation sites using multiple kernel learning

Jong Kyoung Kim, Seungjin Choi

Dept. of Computer Science, POSTECH

Abstract

Phosphorylation is one of the most important post translational modifications which regulate the activity of proteins. The problem of predicting phosphorylation sites is the first step of understanding various biological processes that initiate the actual function of proteins in each signaling pathway. Although many prediction methods using single or multiple features extracted from protein sequences have been proposed, systematic data integration approach has not been applied in order to improve the accuracy of predicting general phosphorylation sites. In this paper, we propose an optimal way of integrating multiple features in the framework of multiple kernel learning. We optimally combine seven kernels extracted from sequence, physico-chemical properties, pairwise alignment, and structural information. Using the data set of Phospho.ELM, the accuracy evaluated by 5-fold cross-validation reaches 85% for serine, 85% for threonine, and 81% for tyrosine. Our computational experiments show significant improvement in the performance of prediction relative to a single feature, or to the combined feature with equal weights. Moreover, our systematic integration method significantly improves the prediction preformance compared with the previous well-known methods.

1. Introduction

Post translational modifications (PTMs) are the permanent or reversible processes which make proteins function correctly. Most proteins are subject to modifications including proteolytic cleavage, protein folding, and covalent modification. Phosphorylation is one of the most significant PTMs because it regulates the activity of proteins, as well as occurs in over 30% of all proteins in the eukaryotic cell [1]. The specific animo acids of serine (S), threonine (T), and tyrosine (Y) in the primary peptide are the major targets for phosphorylation. Accurate prediction of phosphorylation sites is an important issue in systems biology since cell signaling network is regulated by intricate phosphorylation relay. Also it requires a lot of work to identify phosphoproteins using high-throughput techniques such as mass spectrometry [2]. In the problem of predicting phosphorylation sites, one of the most critical factors in deciding prediction accuracy is a way of extracting significant features from the protein sequences because the sequence motifs around phosphorylation sites are highly variable. The extracted features should be represented as either a numerical vector or generalized similarity relationship between two protein sequences such as kernel functions, in order to be used as an input in the machine learning algorithms for classification. NetPhos [3], a widely used web server for predicting general phosphorylation sites, is a predictor based on neural networks. NetPhos uses a sequence window around a candidate site as an input to the neural network classifier. Another web-based predictor, Scansite [4], predicts the phosphorylation sites using the position-specific scoring matrices for each motif constructed from peptide library. Unlike the two methods using only the sequence features, DISPHOS [5] uses heterogeneous features extracted from sequence, physico-chemical properties, and structure. DISPHOS collects the multiple features in a single feature vector, and then performs prediction using logistic regression after the preprocessing steps of feature selection and dimensionality reduction. Although DISPHOS does not use systematic data integration approach, it shows the significantly improved performance over the other two methods using only the sequence information.

In our previous work [8], we proposed an optimal way of integrating multiple features in the framework of multiple kernel learning in order to predict general phosphorylation

sites, without specifying the kind of protein kinases. In our multiple kernel learning approach, each feature was represented via a kernel function implying generalized similarity, and then the heterogeneous kernels are optimally combined using a convex optimization [6]. Here, to develop a more discriminative prediction method, we present four different kinds of kernels which are extracted from sequence, physico-chemical properties, pairwise alignment, and structural information. Using the data set of Phospho.ELM [7], our method achieves the accuracy of 85% for S, of 85% for T, and of 81% for Y. Our computational experiments show that the optimal data integration method based on kernels significantly improves the prediction performance compared with the previous methods.

2. Kernels for prediction of phosphorylation sites

In the problem of classifying a protein sequence into its functional classes, the bottleneck in prediction accuracy is a way of extracting useful features from the protein sequence since the raw sequence cannot be used as an input data in classifiers. The extracted features should be represented as either a numerical vector or a generalized similarity relationship between two protein sequences. Since it is impossible to find a feature extraction method minimizing information loss, one should carefully choose one of available methods or combine them. Combining multiple features or integrating multiple data sources is a central issue in bioinformatics. In recent years, numerous studies have attempted to develop systematic methodology for biological data integration, including Bayesian and kernel methods. However, those principled frameworks have not been widely used in a variety of biological applications because of their difficulties with implementation. Instead, heuristic methods have been more frequently applied in order to combine multiple features, including majority voting and single concatenated feature vector. Although the heuristic methods are relatively easier to implement, they are lack of optimality in combination. In order to integrate multiple features extracted from a protein sequence in the framework of multiple kernel learning described in Section 3, we introduce seven kernel functions designed for predicting phosphorylation sites, which can be classified into the four types of kernels: sequence, physico-chemical properties, pairwise alignment, and structural kernels.

2.1. Kernels for protein sequences

The main assumption on predicting phosphorylation sites is that the amino acid sequence near the candidate site of S, T, or Y determines its phosphorylation. Based on it, we propose the following kinds of feature vectors constructed exclusively from the sequence information: (1) amino acid composition; (2) dipeptide composition; (3) orthogonal binary representation. In the prediction of phosphorylation sites, we assume that the distribution of amino acids of positive sites is different to that of negative sites. The composition of the $i$ith amino acid $f^{ac}(i)$ is given by

$$f^{ac}(i) = N^{ac}(i),$$

where $N^{ac}(i)$ is the total count of the $i$ith amino acid in the sequence window centered at the candidate site. Then, the feature vector $\mathbf{x}_t$ for the tth site is given by

$$\mathbf{x}_t = [f_t^{ac}(1) f_t^{ac}(2) \ldots f_t^{ac}(20)]^T,$$

where $f_t^{ac}(i)$ is the composition of the ith amino acid in the sequence window centered at the tth site.

The dipeptide composition is an extension of amino acid composition, where we add the information on the local order of amino acids. The dipeptide means two consecutive amino acids in a protein sequence. Twenty different amino acids lead to 400 combinations of dipeptide. In practice, it is proved that the dipeptide composition has superior predictive power, compared to the amino acid composition [9]. The composition of the ith dipeptide $f^{dc}(i)$ is given by

$$f^{dc}(i) = N^{dc}(i),$$

where $N^{dc}(i)$ is the total count of the ith dipeptide in the sequence window centered at the candidate site. Then, the feature vector $\mathbf{x}_t$ for the tth site is given by

$$\mathbf{x}_t = [f_t^{dc}(1) f_t^{dc}(2) \ldots f_t^{dc}(400)]^T,$$

where $f_t^{dc}(i)$ is the composition of the ith dipeptide in the sequence window centered at the tth site.

The orthogonal binary representation is a standard feature extraction method in prediction of phosphorylation sites since it considers position specific information of amino acids within the sequence window, as well as is easy to implement. For each position of a sequence window, the amino acid of the position is represented as a 20-dimensional vector of zeros with a single one for the residue observed at the position. We have applied it to a sequence window with a length of 25 residues, resulting a 24X20=480 dimensional binary vector by excluding the central residue of S, T, or Y.

2.2. Kernels for physico-chemical properties

It is well known that the specificity of protein kinases is influenced by acidic, basic, or hydrophobic residues adjacent to the phosphorylated sites [1]. To incorporate this biological knowledge, we consider 121 physico-chemical properties in order to represent a sequence window by a 121-dimensional feature vector based on amino acid composition. We use the AAindex database [10] to get the values of physico-chemical properties for all 20 amino acids, which are thought to be related to protein functions. The average value of the i th physico-chemical property is defined by

$$\varphi(i) = \sum_{j=1}^{20} A_i(j) f^{ac}(j),$$

where $A_i(j)$ is the value of the j th amino acid of the i th physico-chemical property and $f^{ac}(j)$ is the composition of the j th amino acid. The feature vector $\mathbf{x}_t$ for the t th site is given by

$$\mathbf{x}_t = [\varphi_t(1)\varphi_t(2)\ldots\varphi_t(121)]^T,$$

where $\varphi_t(i)$ is the average value of the i th physico-chemical property in the t th site.

2.3. Kernels for pairwise alignment

The fundamental assumption that makes it possible to predict phosphorylation sites from protein sequences is the existence of multiple sequence motifs for kinases near the phosphorylation sites. However, it is difficult to directly identify such motifs because of the sequence divergence. We have proposed several kernel functions which extract the underlying motifs indirectly by utilizing the scores of global pairwise alignments [9]. For the problem of predicting phosphorylation sites, we first perform hierarchical clustering on the training data set of phosphorylation sites to make several clusters, assuming that all sequences belonging to the same cluster share common functional motifs. For clustering, we make a hierarchical tree on sequence windows of positive sites using ward linkage algorithm. For pairwise distances, we calculate the Jukes-Cantor distance between each pair of windows after aligning them with the Needleman-Wunsch algorithm [11]. Then, we make clusters from the constructed hierarchical tree by specifying the number of clusters. The number of clusters is chosen as 200. Next, we select randomly a representative sequence from each cluster and convert a target sequence into the corresponding numerical feature vector by computing the scores of global pairwise sequence alignment between the target sequence and the chosen representative sequences. In the step of performing global pairwise sequence alignment, we assume that the true functional motifs of the target sequence will be properly aligned with the corresponding motifs of at least one of representative sequences if the target sequence contains a true phosphorylation site. To increase discriminative power, we perform the same procedures to the training data set of negative sites. A d-dimensional feature vector $\mathbf{x}_t$ for the t th site has the form

$$\mathbf{x}_t = [x_t(1)x_t(2)\ldots x_t(d)]^T,$$

where $x_t(i)$ is the score of the Needleman-Wunsch algorithm between the t th sequence and the i th representative sequence. Note that d is equal to the total number of selected representative sequences of the training data set of positive and negative sites, and therefore 400. The gap penalty is set to be -3 and the BLOSUM50 matrix is used as a substitution matrix. The assumption on using global pairwise alignment may be incorrect if the functional motifs are localized. In this case, the Smith-Waterman algorithm can be used to compute the scores of local pairwise sequence alignment [12]. The general procedure is equal to that of the above method using global alignment except for using the Smith-Waterman algorithm in the step of constructing feature vectors.

2.4. Kernels for protein structure

Protein kinases are required to be in close contact with the target sites for phosphorylation. If the target site is within inaccessible regions such as the core hydrophobic domain, steric hindrance would prevent the binding of kinases on the site. To incorporate this structural basis with the problem of predicting phosphorylation sites, we consider three concepts of inaccessible regions: (1)transmembrane, (2)protein secondary structure (helix, sheet, and loop), (3)intrinsic disorder. We construct a feature vector by exploiting the outputs from three different kinds of predictors of transmembrane helices, secondary structure, and intrinsic disorder regions. The output is represented in a 1-of-K coding scheme where the output vector has the length of K such that if the output value is k, then all elements of the vector are zero except the kth element taking the value 1. For prediction of transmembrane helices in proteins, three prediction methods of TMHMM [13] , Sosui [14] , and HMMTOP [15] are employed. Next, we also use three predictors of NNPREDICT [16], SOPMA [17], and HNN [18] for the

prediction of secondary structure. Finally, two more predictors of GLOBPLOT2 [19], DISEMBL [20] are used to predict intrinsic disorder. Since DISEMBL gives three prediction results according to their different definitions of intrinsic disorder, the total number of predictors is equal to four. The number of output values, K, in a 1-of-K coding scheme is two except the outputs of predictors for secondary structure, which take the value 3. The feature vector combining the prediction results of three inaccessible regions is constructed by simply concatenating all output vectors.

3. Multiple kernel learning

Multiple kernel learning provides an optimal way of integrating information extracted from heterogeneous data mainly in the supervised learning problems. Integrating information derived from different types of data can be reduced to the problem of combining kernel functions linearly in an optimal way. The optimal kernel combination can be formulated as a convex optimization problem in the framework of support vector machines (SVM) [6]. In multiple kernel learning, the problem of finding optimal kernel weights μ_j can be formulated into the following semi-infinite linear program (SILP) [21] which can be solved efficiently by using standard linear program (LP) solver and standard SVM implementations.

$$\begin{aligned} &\text{maximize}_{\theta,\boldsymbol{\mu}} \quad \theta \\ &\text{subject to} \quad \boldsymbol{\mu} \geq \mathbf{0}, \sum_{j=1}^{m} \mu_j = 1 \\ &\qquad \sum_{j=1}^{m} \mu_j S_j(\boldsymbol{\alpha}) \geq \theta \text{ for all } \boldsymbol{\alpha} \in \mathbb{R}^n \\ &\qquad \text{with } 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}, \boldsymbol{\alpha}^\top \boldsymbol{y} = 0 \end{aligned}$$

For simplicity,we let

$$\begin{aligned} S_j(\boldsymbol{\alpha}) &= \frac{1}{2} \sum_{i,l=1}^{n} \alpha_i \alpha_l y_i y_l k_j(\boldsymbol{x}_i, \boldsymbol{x}_l) - \sum_{i=1}^{n} \alpha_i \\ \mathcal{C} &= \{\boldsymbol{\alpha} | 0 \leq \boldsymbol{\alpha} \leq C\mathbf{1}, \boldsymbol{\alpha}^\top \boldsymbol{y} = 0\}. \end{aligned}$$

To solve the SILP, we use an *exchange method* which is one of the well-known numerical algorithms for SILP. The more detailed description on multiple kernel learning is available in [8]. A sequence window centered at S, T, or Y residue enters a phase of kernel construction fully described in Section 2. The seven feature vectors extracted from the information of sequence, physico-chemical properties, pairwise alignment, and structure are converted into the corresponding kernel function by using the RBF kernel. Then, the multiple kernel functions are linearly combined with the optimal kernel weights learned from the exchange method [8]. The optimally combined kernel function is applied into the standard SVM solver as a single kernel. The predicted class label on phosphorylation site is obtained from the trained SVM classifier with the combined kernel function.

4. Results

4.1. Datasets

A key problem for collecting positive and negative phosphorylation sites is to define a reliable set of training and test data. Although there exist high-throughput methods experimentally confirming positive phosphorylation sites, it is extremely difficult to verify negative sites. We considered all sites of S, T, and Y without annotations of phosphorylation as negative sites. In order to reduce false negative, we removed all negative sites with high sequence identity with positive sites. We obtained the positive and negative sites by extracting sequence windows with a length of 25 residues centered at S, T, or Y from Phospho.ELM (ver 4.0, May. 2006) [7]. We excluded all sequence windows of less than 25 residues. To remove redundancy in the positive sites, we eliminated all positive sites with over 50% sequence identity (30% for S) in the pairwise alignments without gaps. The negative sites may contain false negative, which means un-annotated positive sites. We discard all negative sites with over 30% sequence identity with non-redundant positive sites. In the classification of phosphorylated sites, the numbers of data in two classes are unbalanced because there are much more negative sites than positive sites. This unbalance might cause a poor performance in prediction. To reduce the number of negative sites, we grouped all negative sites into clusters whose number is equal to about 1.5 times the number of non-redundant positive sites. We selected a single representative sequence window for each cluster. The summary of the resulting data sets is shown in Table 1. The performance measures was described in [8].

Table 1
The number of positive and negative phosphorylation sites.

	Number of positive sites	Number of negative sites
S	1219	1829
T	592	890
Y	1029	1544

4.2. Experimental results

The user-provided parameters of our prediction system were chosen by applying 5-fold cross-validation. The RBF

kernel widths γ for the seven feature vectors were selected in such a way that the SVM classifier trained with a single RBF kernel maximizes the accuracy of test data sets in 5-fold cross-validation. During the training and testing, we fix the regularization parameter C to 1. We first performed computational experiments which study the performance of the optimally combined kernel compared to the single kernel constructed from the seven feature vectors independently. Then, we compared the performance of the optimal kernel combination to the kernel combination with equal weights. We finally compared the performance of our method to previous works for predicting phosphorylation sites. The optimally combined kernel outperforms the best result using only one kernel function. The improvement takes place in both accuracy and MCC. The accuracy increases by 1.86% from 82.95% to 84.81% for S, by 1.17% from 83.34% to 84.51% for T, and by 2% from 78.81% to 80.81% for Y. The improvement in MCC corresponds to a change by 0.0395 from 0.6421 to 0.6816, by 0.0241 from 0.6533 to 0.6774, and by 0.0424 from 0.5536 to 0.5960. To confirm the significance of the learned kernel weights, we gave the equal kernel weight of 1/7 to the all single kernel function. The performance of the optimal kernel weights is better than of the equal kernel weights. The accuracy and MCC of the equal kernel weights were 84.41% and 0.6731 for S, 83.69% and 0.6592 for T, and 80.10% and 0.5806 for Y. The kernels of global and local alignment constructed from the scores of pairwise alignment yielded the best individual performance. The result supports the underlying assumptions of alignment kernels that the functional motifs with large sequence variation can be captured via the score of pairwise alignment, and the representative sequences randomly chosen from the clusters can identify the highly diverse motifs. Among the sequence kernels, the kernel constructed from the feature vector of orthogonal binary representation gave the highest performance, reflecting the fact that it fully uses the positional information of amino acids. The kernels based on physico-chemical properties and protein structure showed the marginal performance relatively. However, when the two kernels were removed from the list of kernels for multiple kernel learning, the performance decreased significantly compared to all seven kernels. The performance of our method based on multiple kernel learning is compared with the two other previous methods, using the data set of Table 1. First, the accuracy of NetPhos reaches 68.45% for S, 70.60% for T, and 66.88% for Y. The MCC is 0.3755 for S, 0.3807 for T, and 0.3169 for Y. Next, the accuracy of DISPHOS reaches 84.71% for S, 80.81% for T, and 78.38% for Y. The MCC is 0.6800 for S, 0.6043 for T, and 0.5557 for Y. This result shows that the discriminative method optimally combining multiple features outperforms the method using single feature (NetPhos) or equally combining multiple features (DISPHOS).

5. Discussion

We described a discriminative method for predicting phosphorylation sites of proteins which attempts to find a decision boundary in a feature space constructed from multiple heterogeneous kernel functions. The algorithm based on SVM and SILP finds the optimal kernel weights for the linear kernel combination in order to map the heterogeneous input spaces implicitly into the optimally combined feature space. The computational experiments show significant improvement in the performance of prediction relative to a SVM classifier with the single kernel, or relative to the kernel combination with the equally fixed kernel weights. Our method has three main features that distinguish it from the previous works on phosphorylation prediction. First, kernels provide a highly flexible framework which incorporates heterogeneous features naturally existing in kernels. The four different types of features including sequence, pairwise alignment, physico-chemical properties, and protein structure are represented within the same mathematical object of a kernel function. The identical representation gives an opportunity to integrate features efficiently. Second, the multiple kernels constructed from the heterogeneous features are combined optimally in the framework of SVM and SILP. The optimal criterion is given by maximizing the 1-norm soft margin in SVM. Finally, the algorithm automatically performs feature selection, finding optimal kernel weights. The nonsupport kernels with zero weights are eliminated when the single kernels are linearly combined. We also do not need to consider dimensionality reduction for a high dimensional feature vector because kernels implicitly assume very high dimensionality. Although the method based on multiple kernel learning offered the state-of-the-art result for predicting phosphorylation sites, there remain two basic limitations in need of further research. The first difficulty in learning SVM classifiers is the unreliable negative training set. We considered the un-annotated sites of S, T, and Y as negative sets, where the sites which have over 30% sequence identity with the annotated positive sites were removed. The second limitation is the fact that the overall performance of the optimally combined kernel strongly depends on the single kernel showing the best

performance. Therefore, we need to develop more discriminative kernels reflecting the characteristics of sequence divergence and diversity in phosphorylation sites.

6. Acknowledgments

This work was supported by National Core Research Center for Systems Bio-Dynamics and Basic Research Fund in POSTECH.

7. References

[1] Blom, N., Sicheritz-Ponten, T., Gupta, R., Gammeltoft, S., and Brunak, S. (2004). Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. *Proteomics*, 4, 1633-1649.

[2] Mann, M., Ong, S., Gronborg, M., Steen, H., Jensen, O. N., and Pandey, A. (2002). Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome. *Trends in Biotechnology*, 20, 261-268.

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

[4] Yaffe, M. B., Leparc, G. G., Lai, J., Obata, T., Volinia, S., and Cantley, L. C. (2001). A motif-based pro_le scanning approach for genome-wide prediction of signaling pathways. *Nat. Biotechnol.*, 19, 348-353.

[5] 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.

[6] Lanckriet, G. R. G., Cristianini, N., Bartlett, P., Ghaoui, L. E., and Jordan, M. I. (2004). Learning the kernel matrix with semide_nite programming. *Journal of Machine Learning Research*, 5, 27-72.

[7] Diella, F., Cameron, S., Gem ünd, C., Linding, R., Via, A., Kuster, B., Sicheritz-Ponten, T., Blom, N., and Gibson, T. J. (2004). Phospho.elm: A database of experimentally veri_ed phosphorylation sites in eukaryotic proteins. *BMC Bioinformatics*, 5, 79-83.

[8] Kim J., Kim S., and Choi S. (2006). Prediction of post-translational phosphorylation of proteins using multiple kernel learning. IEIC Technical Report (Institute of Electronics, Information and Communication Engineers) 106(376) 79-84

[9] Kim, J. K., Bang, S. Y., and Choi, S. (2006). Sequence driven features for prediction of subcellular localization of proteins. *Pattern Recognition*, 39, 2301.2311.

[10] Kawashima, S. and Kanehisa, M. (2000). AAindex: amino acid index database. *Nucelic Acids Res.*, 28, 374.374.

[11] Needleman, S. B. and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol. Biol.*, 48, 443.453.

[12] Smith, T. F. and Waterman, M. S. (1981). Identi_cation of common molecular subsequences. *J Mol. Biol.*, 147, 195.197.

[13] Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes. *J Mol. Biol.*, 305, 567.580.

[14] Hirokawa, T., Boon-Chieng, S., and Mitaku, S. (1998). Sosui: classi_cation and secondary structure prediction system for membrane proteins. *Bioinformatics*, 14, 378.379.

[15] Tusnady, G. and Simon, I. (2001). The hmmtop transmembrane topology prediction server. *Bioinformatics*, 17, 849.850.

[16] Kneller, D. G., Cohen, F. E., and Langridge, R. (1990). Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol. Biol.*, 214, 171.182.

[17] Geourjon, C. and Deleage, G. (1995). Sopma: signi_cant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Bioinformatics*, 11, 681.684.

[18] Guermeur, Y. (1997). *Combinaison de classi_eurs statistiques ,application a la prediction de la structure secondaire des proteines*. Ph.D. thesis, Universite Paris 6.

[19] Linding, R., Russell, R. B., Neduva, V., and Gibson, T. J. (2003). Globplot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.*, 31, 3701.3708.

[20] Linding, R., Jensen, L. J., Diella, F., Bork, P., Gibson, T. J., and Russell, R. B. (2003). Protein disorder prediction: implications for structural proteomics. *Structure*, 11, 1453.1459.

[21] Sonnenburg, S., R ätsch, G., Sch äfer, C., and Sch ölkopf, B. (2006). Large scale multiple kernel learning. *Journal of Machine Learning Research*, **7**, 1531.1565.