PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2017, 44(12): 1118~1124
www.pibb.ac.cn

Phosphorylation Site Prediction Integrating The Position Feature With Sequence Evolution Information*

TAN Si-Qiao[1, 2]**, LI Qian[2, 3], CHEN Yuan[4], PENG Jian[1, 2]

([1] *College of Information Science and Technology, Hunan Agricultural University, Changsha* 410128, China;
[2] *Hunan Engineering Research Center for Information Technology in Agriculture and Rural, Changsha* 410128, China;
[3] *Center of Information Construction and Management, Hunan Agricultural University, Changsha* 410128, China;
[4] *College of Plant Protection, Hunan Agricultural University, Changsha* 410128, China)

Abstract Phosphorylation is the major post-translation modification to proteins, and it can be classified as kinase-specific and non-kinase-specific. This paper focuses on the prediction methods of non-kinase-specificity and using Dou's dataset of phosphorylation sites as the template, this paper develops a position-based chi-square table feature, χ^2-pos, and then integrates this feature with the pseudo position-specific scoring matrix (PsePSSM). A Support Vector Machine (SVM) classifier with balanced positive and negative samples was created, and the *S*, *T*, *Y* independent testing results for the Matthew correlation coefficient, the inferior surface integral of the ROC curve and the precision were (0.59, 0.87, 79.74%), (0.55, 0.85, 77.68%) and (0.50, 0.81, 75.22%), respectively, which are significantly superior to the results reported previously. The integration of the χ^2-pos and the PsePSSM offers a promising method to predict phosphorylation sites more accurately in proteins.

Key words phosphorylation, prediction, chi-square table feature, PsePSSM, SVM
DOI: 10.16476/j.pibb.2016.0351

Protein phosphorylation involves the transfer of phosphate groups of ATP or GTP to specific sites on proteins, namely serine, threonine and tyrosine residues, by protein kinases. The process transforms external stimuli into intracellular signals and represents an important post-translation modification, with >30% of proteins phosphorylated[1-2]. In addition, dysfunctional protein phosphorylation is the hallmark of a number of diseases, such as rheumatic arthritis[3], Alzheimer's disease[4] and diabetes[5]. The identification of phosphorylation sites in proteins depends mainly on experimental high through-put methods [6]. However, these experimental methods are time-consuming, and thus the feasibility of using these tests to examine a large number of protein sequences is very low. This highlights the extreme importance of applying machine learning ways to predict phosphorylation sites.

Currently, there are primarily two methods for predicting protein phosphorylation: kinase-specificity and non-kinase-specificity. On the one hand, kinase-specificity methods [7-10] only use proteins and kinases as inputs, and the resulting phosphorylation sites identified are catalyzed by the kinases used as input. On the other hand, non-kinase-specificity methods[11-14] only use proteins as input and the output provides all possible phosphorylation sites. With the development of sequencing technologies, many non-model organisms have been sequenced and most

*This work was supported by a grant from Natural Science Foundation of Hunan province (14JJ2082).

**Corresponding author.
Tel: 86-13974893577, E-mail: 545534721@qq.com

Received: July 9, 2017 Accepted: August 28, 2017

kinases have been identified. Nonetheless, substrate information required for developing kinase-specificity algorithms is insufficient, and thus, this paper focuses on the study of non-kinase-specificity prediction methods.

Research on protein phosphorylation includes examining the sequence encoding scheme covering the amino acid frequency[11, 15], evolution information[12, 15], predicted secondary structure[13], predicted disordered zone[11, 13], physical and chemical properties[11], and the K-nearest neighboring feature [11, 13]. The classifier methods are primarily neural network [16–18] and SVM[11–12, 19–22] based. In 2010, Swaminathan *et al.*[22] reported precision of the non-kinase-specificity methods up to 80%. In 2014, on the basis of eight kinds of features including the Shannon entropy, Dou *et al.*[13] adopted a SVM classifier and different window lengths of S, T and Y, and the Area Under the Curve, also briefed as AUC, values after ten intersections were 0.8405, 0.8183 and 0.7383, whereas the AUC values obtained by independent tests were 0.7761, 0.6652 and 0.5958.

This paper presents the results of a position-based chi-square table (χ^2-pos) using the dataset created by Dou *et al.*[13] while considering its integration with the pseudo amino acid sequence evolution information expressed by PsePSSM, and then a SVM classifier is used to obtain ideal independent predicted results on Dou's dataset. In order to ensure efficient prediction of the positive and negative samples, the MCC parameter is treated as the primary measure standard. The following sections report our results in detail.

1 Data and methods

1.1 Datasets

Table 1 shows the number of sample sequences used from the animal protein sequences reported by Dou *et al.*[13] through taking $S/T/Y$ as the center residue and forty one residues as the window length. For positive samples, our method randomly selects 70% of the sequences for training and the other 30% for independent testing. For negative samples, the training set and the testing set were randomly selected from sequences where the positive samples and the negative samples account for 50%, respectively.

Table 1 Numbers of known phosphorylation sites for P.ELM datasets (window size is 41)

Residue	Sequences	Positive	Negative	Training	Test
S	6 635	18 902	18 902	26 462	11 342
T	3 227	5 183	5 183	7 256	3 110
Y	1 392	1 925	1 925	2 696	1 154

1.2 Position-based chi-square table feature (χ^2-pos)

Statistics of the frequencies at the ith (i = 1, 2, ⋯, 41) position of the positive samples and the negative samples of the twenty kinds of amino acid residues in the training set is outlined in the following 2 × 20 table (Table 2).

Table 2 Frequency distribution of amino acid residues between positives and negatives for the ith position

Sample	Amino acid residue						Total
	1(A)	2(R)	⋯	j	⋯	20(V)	
True	$f^+_{i,1}$	$f^+_{i,2}$	⋯	$f^+_{i,j}$	⋯	$f^+_{i,20}$	f^+_i
False	$f^-_{i,1}$	$f^-_{i,2}$	⋯	$f^-_{i,j}$	⋯	$f^-_{i,20}$	f^-_i
Total	$f_{i,1}$	$f_{i,2}$	⋯	$f_{i,j}$	⋯	$f_{i,20}$	N

In the above table, $f^+_{i,j}$ represents the frequency at the ith position of the positive samples of the jth kind of residue, $f^-_{i,j}$ represents the frequency at the ith position of the negative samples of the jth kind of residue, and $f_{i,j}$ represents the frequency at the ith position of both samples of the jth kind of residue. The chi-square value is calculated by the following expression:

$$\chi^2=\frac{N^2}{f_i^{+}\times f_i^{-}}\left[\sum_{i=1}^{20}\frac{f_{i,j}^{+2}}{f_{i,j}}-\frac{f_i^{+2}}{N}\right] \tag{1}$$

If a new training sample is added and the ith position of it is the jth kind of amino acid: (i) assuming the new sample is a positive one then substitute $f_{i,j}^{+}$ for $f_{i,j}^{+}$ +1, and calculate a chi-square value $\chi_{i,j}^{+}$ according to the expression (1); and (ii) assuming the new sample is a negative one then substitute $f_{i,j}^{-}$ for $f_{i,j}^{-}$ +1, and calculate a chi-square value $\chi_{i,j}^{-}$ according to the expression (1). Thus, the score for the chi-square table with the jth kind of residue at the ith position is $\Delta\chi_{i,j}=\chi_{i,j}^{+}-\chi_{i,j}^{-}$. Taking the window length of 41 as the example, the following 20×41 chi-square table(Table 3) can be obtained.

Table 3 Chi-square difference values for 20 amino acid residues and 41 positions

Amino acid residue	Protein(P)			
	Position(−20)	···	Position(i)	Position(+20)
1(A)	$\Delta\chi_{-20,1}$	···	$\Delta\chi_{i,1}$	$\Delta\chi_{20,1}$
···	···	···	···	···
j	$\Delta\chi_{-20,j}$	···	$\Delta\chi_{i,j}$	$\Delta\chi_{20,j}$
20(V)	$\Delta\chi_{-20,20}$	···	$\Delta\chi_{i,20}$	$\Delta\chi_{20,20}$

As a result, for each sequence in the training set and the testing set, if the jth amino acid appears at the ith position, then the value $\Delta\chi_{i,j}$ can be assigned.

1.3 PsePSSM

The sequence evolution information is expressed by the Position-Specific Scoring Matrix (PSSM) and obtained through three iterative searches using the Swiss-Prot database executed locally by PSI-BLAST[23] with the E-value set to 0.001. As an extension of BLAST, the PSI-BLAST program allows an iterative search and is good at finding distant relationships between different sequences[23]. The normalized PSSM for a sequence of a length L is:

$$P_{\text{PSSM}}=P_{i\to j}(i=1,2,\cdots,L;j=1,2,\cdots,20) \tag{2}$$

In the above expression, $P_{i\to j}$ represents the normalized score when the ith residue mutates into the jth kind of natural amino acid.

Shen *et al* [24] proposed the use of the pseudo amino acid sequence evolution information expressed by the PsePSSM to solve the problem of inconsistent sample characterization dimensions resulting from unequal sequences.

$$\begin{cases}P_{\text{psePSSM}}^{m}=\left[G_1^{m},\ G_2^{m},\ \cdots,\ G_j^{m},\ \cdots,\ G_{20}^{m}\right]\\ G_j^{m}=\dfrac{1}{L-m}\sum\limits_{i=1}^{L-m}\left[P_{i\to j}-P_{(i+m)\to j}\right]^2\end{cases}$$

$$(i=1,2,\cdots,L;j=1,2,\cdots,20;m=0,1,\cdots,\lambda) \tag{3}$$

In expression (3), G_j^{m} represents the correlation factor when the jth kind of amino acid is at a spacing of m. Thus, each sequence can obtain the PsePSSM feature of 20 × (λ + 1) dimensions.

1.4 Classifier and model

This paper uses Libsvm 3.1[25] as the classifier and fixes the radial basis kernel as the kernel function. The parameters c and g are automatically obtained through the 5-fold cross-test search on the training set executed by the grid.py. Considering the SVM's immunity to the feature dimensions and common feature selection methods, such as the MRMR, recursive feature extraction (RFE) have negligvble or even adverse effects on this dataset, so this paper does not carry out the feature selection.

Indexes such as sensitivity (Sn), specificity (Sp), accuracy (Ac) and the Matthew correlation coefficient (MCC) are used to evaluate the performance of this model:

$$Sn=\frac{TP}{TP+FN}\times100\% \tag{4}$$

$$Sp=\frac{TN}{TN+FP}\times100\% \tag{5}$$

$$Ac=\frac{TP+TN}{TP+TN+FP+FN}\times100\% \tag{6}$$

$$MCC=\frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \tag{7}$$

In the above expressions, the TP (true positive), the TN (true negative), the FN (false negative) and the FP (false positive) represent true judgments over the positive samples, true judgments over the negative samples, false judgments over the positive samples and false judgments over the negative samples, respectively.

The receive operating characteristic (ROC) curve is also extensively used to measure the performance of the prediction models[26-27], which takes $1-Sp$ as the x-coordinate and Sn as the y-coordinate to show the curves mapped under all possible thresholds. The area under the ROC curve (AUC) ranges from 0 to 1, with an AUC value closer to 1 when the prediction performance is better.

2 Results

2.1 Optimization of the PsePSSM parameters

The PsePSSM has three parameters that are determined: the left length (ll), the right length (rl) and the maximum spacing λ. This paper executes optimizations in the left window (ll) by taking five residues for the shortest ll, nineteen residues for the longest ll, two residues for the step, and the number of levels is set to eight. The above settings are also applied to the right window. For λ, the minimum, the maximum and the step are taken as five residues, nineteen residues and two residues, respectively, with the number of levels set to eight. The whole combination needs to execute 152 processes. For each process, the PsePSSM characterization, the five-fold crossing of the training set and the testing precision are collectively used as the measuring criteria for the optimization of each sequence. The optimized results are shown in Table 4.

Table 4　Optimum parameters computed by the 5-fold crossing of the number of PsePSSM features

Site	ll	rl	λ	Feature number
S	19	19	13	280
T	15	15	19	400
Y	15	13	19	400

2.2 Feature integration

The performances for the independent separate integration prediction and the independent two-feature integration prediction based on the χ^2-pos and the PsePSSM are shown in Table 5. The results obtained by the two-feature integration prediction are the optimum for the three sites.

Table 5　Prediction accuracy of phosphorylation sites based on different encoding schemes

Site	Encoding schemes	AUC	MCC	Ac	Sn	Sp
S	χ^2-pos	0.83	0.51	75.52	75.43	75.61
	PsePSSM	0.86	0.56	77.92	78.34	77.58
	χ^2-pos + PsePSSM	0.87	0.59	79.74	78.57	80.90
T	χ^2-pos	0.81	0.45	72.69	72.72	72.67
	PsePSSM	0.85	0.55	77.68	76.66	78.71
	χ^2-pos + PsePSSM	0.86	0.56	78.00	76.32	79.68
Y	χ^2-pos	0.73	0.35	67.59	67.59	67.59
	PsePSSM	0.76	0.41	70.45	63.77	77.12
	χ^2-pos + PsePSSM	0.81	0.50	75.22	77.64	72.79

2.3 Comparison with the PhosphoSVM method

The PhosphoSVM method[13] extracts eight kinds of features from the datasets with balanced positive samples and negative samples, and executes ten crossings through the SVM. Our method extracted the position-based chi-square table features (χ^2-pos) and the pseudo amino acid evolution information from the same datasets, and uses the SVM for the independent tests. The comparison between PhosphoSVM[13] and the results of our method are presented in Table 6, showing that the method proposed herein is significantly superior.

Table 6 Comparison of predictive accuracy for three sites with other method

Site	Method	*AUC*	*MCC*	*ACC*	*Sn*	*Sp*
S	This paper	0.87	0.59	79.74	78.57	80.90
	PhosphoSVM	0.84	0.30	69.24	44.43	94.04
T	This paper	0.86	0.56	78.00	76.32	79.68
	PhosphoSVM	0.82	0.25	66.15	37.31	94.99
Y	This paper	0.81	0.50	75.22	77.64	72.79
	PhosphoSVM	0.74	0.21	64.63	41.92	87.34

3 Discussion

3.1 Merits of the χ^2-pos feature

Common position-based sequence characterizations include the 0/1 code for the residue and the code for the physicochemical property of the amino acid. If the 0/1 code for residues is adopted, then the expression of each position needs twenty dimensions of 0/1 features; the obtained feature matrix is very sparse and does not reflect the degree of difference among residues in a physicochemical property. For example, there are three kinds of residues, *S*, *N* and *W*, at a site. Assuming the dependent variables, phosphorylation or non-phosphorylation, are mainly related to the hydrophobic nature of the residue at the site, their hydrophobic indexes are 0.05, 0.06 and 2.65, respectively. The *S*-*N* hydrophobic nature gap is extremely small, whereas the *N*-*W* hydrophobic nature gap is larger; however, both the *S*-*N* distance and the *N*-*W* distance are equal to 1 when the 0/1 code is adopted. The AAindex database(see http://www.genome.jp/aaindex/) includes 531 kinds of physicochemical properties from the twenty kinds of natural amino acids. When the code for physicochemical properties of amino acids is used, then the expression of each position requires 531 kinds of physicochemical properties, because it is unknown which physicochemical properties are related to the dependent variables. Consequently, the existence of a large number of irrelevant features and redundant features among them increases the difficulty of feature selection.

When there is high similarity between the positive sample sequences and the negative sample sequences, such as two segments from the protein A0AVK6 with the true phosphorylation site at position 357, the use of the component-based or the association-based methods will yield feature vectors that are very similar and the classifiers will fail to distinguish the two sequences efficiently. In contrast, much larger differences among the feature vectors extracted by the position-based methods like χ^2-pos are determined, and this ensures that the classifiers can efficiently distinguish the two sequences. Positive sample P1: (342) AFKWTGPEIS-PNTSG**S**SPVIHFTPSDLEVRR(372); Negative sample N1: (343)FKWTGPEISPNTSGS**S**PVIHFTPSDLEVR-RS (373).

The data-driven χ^2-pos method is able to reflect the scoring differences of different residues at the same position or the same residue at different positions, and differentiate highly similar positive sample sequences and negative sample sequences efficiently. In addition, this method also has a number of valuable merits, such as less feature dimensions, low redundancy and a non-sparse feature matrix. Thus, its application prospect in molecular sequence characterization is expected to be very extensive.

3.2 Necessity of combining the improved PSSM and the χ^2-pos

The position-based PSSM method is able to differentiate between highly similar positive sample sequences and negative sample sequences efficiently but is not very robustness when there are insertions and deletions in the amino acid sequence. In contrast, the PsePSSM method solves the unequal sequence problem and makes the sequence evolution information features more robust. The position-based χ^2-pos method can differentiate between highly similar positive sample sequences and negative sample sequences, whereas the PsePSSM method can reflect the sequence evolution information or even distant relationships, and the PsePSSM method is fault-tolerant. These two kinds of features can characterize a sequence from different aspects and complement each other. Both are necessary, and the prediction performance obtained by integrating χ^2-pos and PsePSSM is optimum (Table 5).

References

[1] Ficarro S B, McCleland M L, Stukenberg P T, *et al*. Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*. Nature Biotechnology, 2002, **20**(3): 301-305.

[2] Krupa A, Preethi G, Srinivasan N. Structural modes of stabilization of permissive phosphorylation sites in protein kinases: distinct strategies in Ser/Thr and Tyr kinases. Journal of Molecular Biology, 2004, **339**(5): 1025-1039

[3] Grabiec A M, Korchynskyi O, Tak P P, *et al*. Histone deacetylase inhibitors suppress rheumatoid arthritis fibroblast-like synoviocyte and macrophage IL-6 production by accelerating mRNA decay. Annals of the Rheumatic Diseases, 2012, **71**(3): 424-431

[4] Avila J. Tau phosphorylation and aggregation in Alzheimer's disease pathology. FEBS Letters, 2006, **580**(12): 2922-2927

[5] Cohen P. The role of protein phosphorylation in human health and disease. European Journal of Biochemistry, 2001, **268**(19): 5001-5010

[6] Trost B, Kusalik A. Computational prediction of eukaryotic phosphorylation sites. Bioinformatics, 2011, **27**(21): 2927-2935

[7] Blom N, Sicheritz-Pontén T, Gupta R, *et al*. Prediction of post-translational glycosylation and phosphorylation of proteins from the amino acid sequence. Proteomics, 2004, **4**(6): 1633-1649

[8] Neuberger G, Schneider G, Eisenhaber F. pkaPS: prediction of protein kinase A phosphorylation sites with the simplified kinase-substrate binding model. Biol Direct, 2007, **2**(1): 1-23

[9] Yoo P D, Ho Y, Zhou B, *et al*. SiteSeek: post-translational modification analysis using adaptive locality-effective kernel methods and new profiles. BMC Bioinformatics, 2008, **9**(1): 272-289

[10] Sobolev B, Filimonov D, Lagunin A, *et al*. Functional classification of proteins based on projection of amino acid sequences: application for prediction of protein kinase substrates. BMC Bioinformatics, 2010, **11**(1): 313-331

[11] Gao J, Agrawal G K, Thelen J J, *et al*. A new machine learning method for protein phosphorylation site prediction in plants// Bioinformatics and Computational Biology. Springer Berlin Heidelberg, 2009: 18-29

[12] Biswas A K, Noman N, Sikder A R. Machine learning method to predict protein phosphorylation sites by incorporating evolutionary information. BMC Bioinformatics, 2010, **11**(1): 273-290

[13] Dou Y, Yao B, Zhang C. PhosphoSVM: prediction of phosphorylation sites by integrating various protein sequence attributes with a support vector machine. Amino Acids, 2014, **46**(6): 1459-1469

[14] Huang S Y, Shi S P, Qiu J D, *et al*. Using support vector machines to identify protein phosphorylation sites in viruses. Journal of Molecular Graphics and Modelling, 2015, **56**(3): 84-90

[15] Iakoucheva L M, Radivojac P, Brown C J, *et al*. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Research, 2004, **32**(3): 1037-1049

[16] Blom N, Gammeltoft S, Brunak S. Sequence and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of Molecular Biology, 1999, **294**(5): 1351-1362

[17] Hjerrild M, Stensballe A, Rasmussen T E, *et al*. Identification of phosphorylation sites in protein kinase A substrates using artificial neural networks and mass spectrometry. Journal of Proteome Research, 2004, **3**(3): 426-433

[18] Ingrell C R, Miller M L, Jensen O N, *et al*. NetPhosYeast: prediction of protein phosphorylation sites in yeast. Bioinformatics, 2007, **23**(7): 895-897

[19] Plewczyń ski D, Tkacz A, Godzik A, *et al*. A support vector machine method to the identification of phosphorylation sites. Cell Mol Biol Lett, 2005, **10**(1): 73-89

[20] Wong Y H, Lee T Y, Liang H K, *et al*. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. Nucleic Acids Research, 2007, **35**(suppl 2): W588-W594

[21] Heazlewood J L, Durek P, Hummel J, *et al*. PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor. Nucleic Acids Research, 2008, **36**(suppl 1): D1015-D1021

[22] Swaminathan K, Adamczak R, Porollo A, *et al*. Enhanced prediction of conformational flexibility and phosphorylation in proteins [M]//Advances in Computational Biology. New York Springer, 2010: 307-319

[23] Schöffer A A, Aravind L, Madden T L, *et al*. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Research, 2001, **29**(14): 2994-3005

[24] Shen H B, Chou K C. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Engineering Design and Selection, 2007, **20**(11): 561-567

[25] Chang C C, Lin C J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2011, **2**(3): 1-27

[26] Bewick V, Cheek L, Ball J. Statistics review 13: receiver operating characteristic curves. Critical Care, 2004, **8**(6): 508-512

[27] Centor R M. Signal detectability the use of ROC curves and their analyses. Medical Decision Making, 1991, **11**(2): 102-106

融合位置特征与序列进化信息的磷酸化位点预测 *

谭泗桥 [1,2]** 李　钎 [2,3] 陈　渊 [4] 彭　剑 [1,2]

([1] 湖南农业大学信息科学技术学院，长沙 410128；[2] 湖南省农村农业信息化工程技术研究中心，长沙 410128；
[3] 湖南农业大学信息化建设与管理中心，长沙 410128；[4] 湖南农业大学植物保护学院，长沙 410128)

摘要　磷酸化是蛋白质翻译后的主要修饰，可分为激酶特异性和非激酶特异性两种类型．以非激酶特异性磷酸化位点 Dou 数据集为基础，本文发展了一种基于位置的卡方差表特征χ^2-pos，融合伪氨基酸序列进化信息 PsePSSM 表征序列，构建正负样本均衡的支持向量机分类器，*S, T, Y* 独立测试 Matthew 相关系数、ROC 曲线下面积分及准确率分别达到了(0.59、0.87、79.74%)，(0.55、0.85、77.68%)和(0.50、0.81、75.22%)，明显优于文献报道结果．χ^2-pos、PsePSSM 两种特征的融合在蛋白质磷酸化位点预测中有广泛应用前景．

关键词　磷酸化，预测，卡方差表特征，伪氨基酸序列进化信息，支持向量机

学科分类号　Q51，Q61　　**DOI**: 10.16476/j.pibb.2016.0351

* 湖南省自然科学基金资助项目(14JJ2082).

** 通讯联系人.

Tel: 013974893577, E-mail: 545534721@qq.com

收稿日期：2017-07-09，接受日期：2017-08-28