New encoding schemes for prediction of protein phosphorylation sites

Zimo Yin
College of Science
China Agricultural University
Beijing, China 100083
Email: cau.yinzm@gmail.com

Junyan Tan*
College of Science
China Agricultural University
Beijing, China 100083
Email: tanjunyan0@126.com

***Abstract*—Protein phosphorylation is involved in most cellular functions. Because of the importance of protein phosphorylation, many methods are conducted to identify the phosphorylation sites. Experimental methods for identifying phosphorylation sites are not only costly but also time consuming. Hence, computational methods are highly desired. In this paper, three new encoding methods, BinCTF(Binary-conjoint triad feature), CTF2(new conjoint triad feature) and BinCTF2(Binary-new conjoint triad feature), which are the modification of Binary and CTF encoding, are developed. Then an ensemble support vector machine is applied to predict the phosphorylation sites related to serine (S), threonine (T) and tyrosine (Y) residues. The numerical results indicate that some of the performance of these new methods are better than previous methods.**

Key words: support vector machine; encoding scheme; protein phosphorylation; prediction

I. Introduction

The post-translational modification (PTM) of proteins is a common biological mechanism for regulating protein functions. It makes protein complicate in structure, complete in function, and precise in regulation. Almost all kinds of proteins need some sort of translational modification during its synthetical process or after it. This paper concerns kinase-specific phosphorylation sites prediction. Kinase-specific phosphorylation is the central mechanism of post-translational modification to regulate cellular responses and phenotypes. Signaling defects associated with protein phosphorylation are linked to many diseases, particularly cancers. Characterizing protein kinases and their substrates will enhance our ability of understanding and treating such diseases and broaden our knowledge of signaling networks. So, the ability to predict potential phosphorylation sites has considerable value. Some experimental technologies have been applied to the identification of phosphorylation sites, but these methods have common problems which are time consuming and expensive. Therefore, the computational method is highly desired because of its remarkable merits, quick and convenience.

Some of the currently available computational methods are based on sequence information and machine learning techniques. And the encoding schemes based on sequence information have been successfully applied in the prediction of PTM sites. The currently existed encoding schemes considers both the specific and the characters of amino acids. Kim *et al.* [1] used the standard binary encoding scheme, which considered only the specific amino acid, to predict phosphorylation sites and obtained better outcome than all previous methods. There are also encoding schemes that considered the chemical and physical information of amino acids including the reduce alphabet encoding schemes[2], blosum62 encoding [3], SARAH1 hydrophobicity scale encoding [4] and the Conjoint triad feature (CTF) [5] etc. Another kind of encoding schemes considered the distribution of amino acid pairs, such as the coupling patterns encoding scheme (couple pattern recognition, CPR) [3], position weight matrix (PWM) encoding scheme [16] and the position weight matrices (PWMs) encoding scheme [8]. In this paper, we propose three new encoding schemes which consider the specific amino acid and the chemical characters of amino acids.

After the input vectors are constructed by encoding scheme, another important issue is choosing the classification method. Support vector machine(SVM) has been widely used in the fields of bioinformatics and it also has a good performance in the prediction of PTM sites. In this paper, we use an ensemble of SVM [8], [11] classifiers for classification because of the imbalanced data.

The rest of this paper is organized as follow. In section 2, we introduce the proposed encoding schemes and SVM. The experimental design and results are reported in section 3. In the last section, we conclude this paper.

II. Methods

The ability to precisely predict phosphorylation sites depends strongly on the encoding scheme. We come up with three new encoding scheme based on two pervious encoding schemes to seek for better prediction accuracy.

A. Encoding schemes

Binary-conjoint triad feature (BinCTF) First, we briefly introduce Binary and CTF encoding schemes. Binary encoding is proposed by Kim *et al.* [1]. According to this scheme, each amino acid is mapped to a 21-dimensional vector. The twenty standard amino acids are ordered one to twenty, and the i-th amino acid has the binary codeword of twenty bits with the i-th bit set to 1 and all others to 0, for $i = 1, 2, \cdots, 20$. In

*Corresponding author: Junyan, Tan

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

order to ensure the identical of window size of the sequence, an extra amino acid is added and we call it "the dummy amino acid". It is set to 0 for the first twenty units and 1 for last unit. A sequence fragment of length $2n+1$ is encoded in $21*(2n+1)$ bits with the binary codewords of amino acids concatenated based on their order in the fragment. The traditional CTF encoding method is firstly proposed for the prediction of protein-protein interaction by Shen *et al.* [5] in 2007. According to the dipoles and volumes of the side chains, 20 amino acids were classified into seven classes: {A, G, V }, {I, L, F, P}, {Y, M, T, S}, {H, N, Q, W}, {R, K}, {D, E}, {C}. The conjoint triad considered the properties of one amino acid and its vicinal amino acids, the conjoint triad regarded any three continuous amino acids belonging to the same classes, such as 'ART' and 'VKS', could be treated identically. For the amino acids, there are $7*7*7=343$ different types of conjoint triad. Then a binary space (V, F) is used to represent a protein sequence fragment. Here V is the vector space of the sequence features, and each feature v_i, $i=1,2,\cdots,343$ represents a sort of triad type; F is the frequency vector corresponding to V, and the value of the i-th dimension of F, f_i, $i=1,2,\cdots,343$, is the frequency of type v_i appearing in the protein sequence. Clearly, each sequence fragment has a corresponding F vector. [10] predicted phosphorylation sites by using CTF and Binary encoding scheme, the results are shown in I.

TABLE I
THE AUC OF CTF ENCODING SCHEME (WINDOW SIZE=13)

	CDK	CDK1	CDK2	CK2	MAPK3	GRK	PKA	PKC	SRC
CTF	0.795	0.795	0.735	0.864	0.774	0.792	0.884	0.849	0.576
Binary	0.939	0.962	0.748	0.903	0.938	0.884	0.934	0.878	0.730

From Table I, we can see that CTF encoding scheme is not good enough for predicting phosphorylation sites. It has a lower mean of AUC. The highest AUC is just 0.884 (for predicting PKA) and the lowest is 0.576 (for predicting SRC). Since, CTF encoding scheme has a very good performance in many other protein prediction issue [5], there must be some reasons for its bad performance on predicting phosphorylation sites. The reason might be that the CTF encoding scheme doesn't distinguish the amino acids within the same group, that is, overlooking the difference between those amino acids in the same class. While most of the phosphorylation sites closely relate to not only the physical and chemical property of the amino acids but also the specific amino acid. Based on the above consideration, we propose a new encoding scheme called BinCTF which combines the CTF encoding scheme and Binary encoding scheme. The BinCTF considers both specific amino acid and the chemical and physical character of the amino acid. We hope the BinCTF could perform better than Binary and CTF. This encoding scheme translates an amino acid sequence fragment into a vector by both Binary encoding scheme and CTF encoding scheme. And then, combine the one with the other as its encoding vector, which means concatenate the two vector together and make it into a longer vector. Since Binary encoding gives each amino acid a 21 dimensional vector and CTF encoding make the sequence into a 343 dimensional vector, the output vector by this method should be a $21*n+343$ dimensional vector, where n is the length of the sequence.

New conjoint triad feature (CTF2) The dummy amino acid is often used to ensure the identical of the window size of the amino acid sequence, but the CTF scheme ignores the dummy amino acid. Just from the perspective of encoding, the dummy amino acid should be considered when we construct the input feature vectors. Therefore, we make the dummy amino acid an extra class, which is noted as O. The whole 21 amino acids were classified into eight classes: {A, G, V }, {I, L, F, P}, {Y, M, T, S}, {H, N, Q, W}, {R, K}, {D, E}, {C}, {O}. Then rest of the encoding method is remain the same as the CTF encoding. The output vector should be a 512 $(8\times8\times8)$-dimensional vector. The new scheme is named as CTF2 in the rest of this paper.

Binary+CTF2 (BinCTF2) To consider both the specific amino acid and the physical and chemical information, we also combine Binary and CTF2 encoding scheme. This encoding scheme may solve its two problems: the ignored dummy amino acid and the difference between amino acids which are classified in the same class. We hope that this encoding scheme can achieve better performance on predicting phosphorylation sites. For a sequence of length n, the output vector by this BinCTF2 should be a $21*n+512$ dimensional vector. We abbreviate this encoding scheme as BinCTF2 in the rest of this paper.

Notice that, it would lead to a very high dimensional vector by these new encoding schemes. It is well-known that feature selection can be helpful in selecting important features of vectors which will decrease the dimension of vectors significantly. In this paper, we use t-test[7] as the feature selection method to optimize our three encoding scheme. For the BinCTF, CTF2 and BinCTF2 encoding schemes, we select 80 features of the translated vector as their encoding vector. So, we can get three new different encoding scheme, they are named as t-BinCTF, t-CTF2, t-BinCTF2, separately.

Next, we briefly introduce the classification methods used in this paper.

Support vector machine(SVM)[17]

Given the training set:

$$T=\{(x_1,y_1),\cdots,(x_l,y_l)\}\in(\mathcal{X}\times\mathcal{Y})^l, \tag{1}$$

where $x_i=([x_i]_1,[x_i]_2,\cdots,[x_i]_n)\in\mathcal{X}\subseteq R^n$ is the input and its n components are called 'features'. For the microarray data, the n features are n gene expression levels. $y_i\in\mathcal{Y}=\{-1,1\}$ is the output, it means 'normal' or 'cancerous' for microarray data. The training set T is given by (1), the SVM is to find a hyperplane that separates the two classes of data points by the maximizing margin:

$$\begin{aligned}
\min_{\boldsymbol{w},\,b,\,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{l}\xi_i\ , && (2)\\
\text{s.t.} \quad & y_i((\boldsymbol{w}\cdot\boldsymbol{x}_i)+b)\geq 1-\xi_i\ ,\ i=1,\cdots,l\ , && (3)\\
& \xi_i\geq 0\ ,\ i=1,\cdots,l\ , && (4)
\end{aligned}$$

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

where the constant $C(> 0)$ determines the trade-off between margin maximization and training error minimization. The dual problem of the primal problem (2) $\sim$ (4) is

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j (x_i \cdot x_j) - \sum_{i=1}^{l}\alpha_i \ , \tag{5}$$

$$\text{s.t.} \quad \sum_{i=1}^{l} y_i \alpha_i = 0 \ , \tag{6}$$

$$0 \le \alpha_i \le C \ , \ i = 1, \cdots, l \ , \tag{7}$$

where $\alpha = (\alpha_1, \alpha_2, \cdots, \alpha_l)^{\mathrm{T}}$. Suppose $\alpha^* = (\alpha_1^*, \alpha_2^*, \cdots, \alpha_l^*)^{\mathrm{T}}$ is the solution of the dual problem (5) $\sim$ (7), if there exists some j such that $0 < \alpha_j^* < C$, the solution about (w, b) of the primal problem (2) $\sim$ (4) can be calculated by the following:

$$w^* = \sum_{i=1}^{l} \alpha_i^* y_i x_i, x_j), \tag{8}$$

$$b^* = y_i - \sum_{i=1}^{l} y_i \alpha_i^* (x_i \cdot x_j). \tag{9}$$

A new point x is to be assigned with the label $f(x) = sgn((\boldsymbol{w^*} \cdot \boldsymbol{x}) + b^*)$.

An ensemble of SVM classifiers Because the data (see Table II, the number of positive points sites is much less than the number of negative points sites) is imbalanced and we cannot guarantee that a single SVM always provides the global optimal classification performance over all test examples, therefore, the ensemble SVM [8], [11] is used for classification in this paper.

An ensemble of SVM classifiers is a collection of some individual SVM classifiers. Each individual SVM has been trained independently from the randomly chosen training samples and the final label of a sample is decided by vote. That is, if the majority of the individual SVM classifiers classify the sample into positive class, this sample will be regarded as a positive sample.

III. Numerical experiment

A. dataset

The database of the Phospho.ELM (version 1109) [14] has been used as a benchmark to test the performance of many published computational models for the phosphorylation prediction. It contains a collection of experimentally verified serine (S), threonine (T), and tyrosine (Y) specific phosphorylation sites in eukaryotic proteins. The entries provide the information about the phosphorylated proteins and the exact positions of the known phosphorylated residues, which are catalyzed by a given kinase.

For a given kinase, we define a local window with each phosphorylation or non-phosphorylation site in the middle and n sequence neighbors on each side; Then the window size is $2n + 1$. Since the window size can affect the prediction accuracy, we need to find the most suitable size. In this paper, the window size is set to 11, 13, 15, 17 and 19 separately to compare with each other. It must be noted that the positive data and negative data might contain some homologous sites from homologous proteins. And the prediction accuracy would be inaccurate if the testing data are highly homologous with the training data. To avoid this fault, we remove the homologous sequences with threshold 70%. That is, if two sequence fragments have 70% identity in the corresponding positions, then only one will be reserved while the other will be discarded. The homology reducing process is separately carried out on positive and negative data. The final non-homologous datasets are summarized in Table II.

TABLE II
The Size of Positive and Negative Datasets of different window length

Protein kinase	n=11	n=13	n=15	n=17	n=19
	l_+/l_-	l_+/l_-	l_+/l_-	l_+/l_-	l_+/l_-
CDK	94/3370	94/3402	93/3391	93/3383	93/3399
CDK1	145/8265	145/8322	145/8311	144/8297	145/8329
CDK2	67/2711	68/2726	68/2724	68/2723	68/2728
CK2	214/7233	219/7323	218/7310	218/7298	222/7335
MAPK3	86/6060	86/6143	86/6121	86/6110	86/6137
GRK	37/312	37/319	37/315	37/316	37/318
PKA	313/17765	319/18066	320/17980	317/17918	321/18075
PKC	225/9946	225/10061	225/10033	225/10020	225/10089
SRC	73/853	73/864	73/863	73/859	74/867

l_+ is the number of positive samples and l_- is the number of negative samples.

B. Evaluate criterion

In order to evaluate our prediction,some criterion are necessary. We use the following four measurements to evaluate our prediction: sensitivity(Sn), specificity(Sp), accuracy(Acc), and the area under the ROC curve(AUC). They are defined as follow:

$$S_n = \frac{TP}{TP + FN},$$

$$S_p = \frac{TN}{TN + FP},$$

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

where TP, TN, FP and FN denotes the number of true positives, true negatives, false positives and false negatives, respectively. The prediction validity is often examined by observing its ROC curve because they are able to show the trade-off between sensitivity and specificity and give a complete evaluation. The area under the curve(AUC) is another important indicator, the larger, the better.

In numerical experiment, we first select a window length and an encoding scheme. Then, put it into an ensemble of SVM classifiers. In this paper, we make an ensemble of 9 classifiers, and each classifiers will give its prediction. The kernel used in each individual SVM is the Gauss kernel and the kernel parameter σ and the penal parameter C in individual SVM is chosen form $2^{-5} \sim 2^5$ respectively. According to

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

comparing results from the numerical experiment, we record those parameter with the largest AUC.

C. Results

For all datasets with different window size, we conduct ten-fold cross validation 5 times and compare the average AUC and S_n, S_p and A_{cc}. For short of page, only parts of the results, including the average AUC with different window size and the ROC curves, are shown in this paper. Next, we analysis the performance of each encoding scheme on all datasets.

From Table III and Fig.1 to Fig.9 , we can see that BinCTF preforms better than CTF and CTF2. The best AUC of BinCTF is the result on the 'CDK1' data, the mean AUC of different window size is 0.955 which is 8% higher then the best AUC of CTF2 on 'CK2'and 12.3%higher than the best AUC of CTF on 'CK2'. The performance of BinCTF is better than Binary on four datasets, 'CDK', 'MAPK3', 'GRK' and 'PKC', the lowest AUC of BinCTF is about 0.1% to 1% higher then the AUC of Binary. On 'CDK1', 'CDK2', 'CK2' and 'SRC', BinCTF performs worse, the AUC is about 1% lower than the AUC of Binary. On 'PKA', two methods perform the same. Comparing BinCTF with BinCTF2, we find that they performs almost the same. The difference of the average AUC between them is no more than 0.9%. We also compare CTF with CTF2 and we find that the CTF2 performs better than CTF on some datasets, such as 'GRK', 'PKA', 'PKC', 'SRC' 'CDK2'. On the other datasets, CTF2 performs worse than CTF. The difference of average AUC between CTF and CTF2 is about 1%. From Table III, we can also find out that feature selection has significant meaning on prediction. Every encoding scheme becomes better when feature selection is applied. The average AUC of t-BinCTF is about 0.03 higher than BinCTF which means 3% improvement, and so does t-BinCTF2. t-CTF2 has an improvement which is about 7.5% on average than CTF, but it is still lower than t-BinCTF and t-BinCTF2. The most remarkable improvement in the using of feature selection is on 'SRC'. Before that, the best AUC prediction on 'SRC' is 0.712, and when feature selection is applied, the best AUC prediction it has is 0.897. Therefore, we conclude that feature selection based on t-test should be widely applied in this prediction process!

D. Comparison with other Prediction Tools

To further evaluate the performance of our methods, we compared it with three existing tools, NetPhosK 1.0[20], GPS 2.1[21], and Musite[22]. For the sake of room, only the three well known kinases family: PKA, CK2, and MAPK3 family are used for comparison. We adjust the prediction thresholds to set the specificity levels as close as possible to 99.0, 98.0, 97.0, 95.0, and 90.0% and compared the corresponding sensitivities. Only t-BinCTF2 and t-BinCTF are used for comparison, since they have the best performance among the methods that we come up with in this paper. From Table IV we can find out that t-BinCTF and BinCTF2 work better than previous methods at some specificity(Sp) level, and have some worse results compare to those methods either. (The best performed result

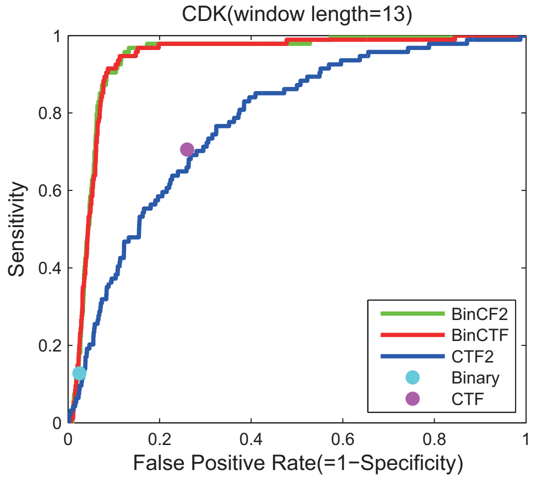

Fig. 1. Roc curve for CDK

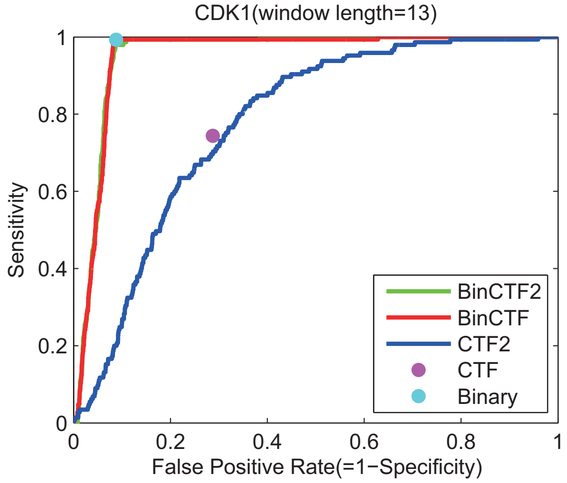

Fig. 2. Roc curve for CDK1

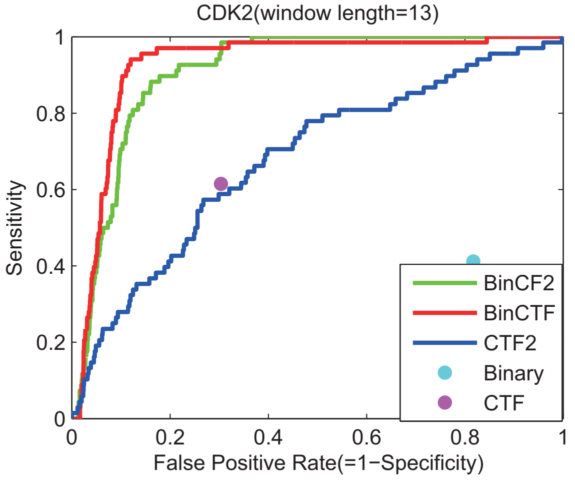

Fig. 3. Roc curve for CDK2

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

TABLE III
THE AUC OF DIFFERENT ENCODING SCHEME

method	length	CDK	CDK1	CDK2	CK2	MAPK3	GRK	PKA	PKC	SRC	mean
Binary	11	0.953	0.965	0.943	0.911	0.942	0.874	0.949	0.890	0.717	0.905
	13	0.939	0.962	0.946	0.903	0.938	0.894	0.937	0.880	0.735	0.904
	15	0.948	0.963	0.947	0.913	0.937	0.883	0.952	0.896	0.704	0.905
	17	0.938	0.962	0.947	0.912	0.934	0.879	0.951	0.894	0.696	0.901
	19	0.888	0.960	0.949	0.916	0.935	0.869	0.951	0.894	0.695	0.895
	mean	0.933	0.962	0.946	0.911	0.937	0.880	0.948	0.891	0.709	
CTF	11	0.801	0.813	0.759	0.871	0.786	0.837	0.892	0.850	0.583	0.799
	13	0.795	0.795	0.735	0.864	0.774	0.792	0.884	0.849	0.576	0.785
	15	0.779	0.776	0.775	0.873	0.762	0.831	0.873	0.857	0.586	0.790
	17	0.744	0.770	0.746	0.878	0.750	0.807	0.867	0.854	0.621	0.782
	19	0.764	0.761	0.748	0.875	0.718	0.815	0.854	0.850	0.623	0.779
	mean	0.777	0.783	0.753	0.872	0.758	0.816	0.874	0.852	0.598	
BinCTF	11	0.946	0.958	0.937	0.907	0.941	0.888	0.949	0.892	0.711	0.903
	13	0.946	0.957	0.937	0.910	0.941	0.919	0.950	0.895	0.698	0.906
	15	0.943	0.954	0.935	0.908	0.941	0.867	0.948	0.899	0.682	0.897
	17	0.941	0.954	0.935	0.907	0.937	0.883	0.946	0.894	0.669	0.896
	19	0.944	0.954	0.931	0.911	0.929	0.869	0.945	0.893	0.692	0.896
	mean	0.944	0.955	0.935	0.909	0.938	0.885	0.948	0.895	0.690	
BinCTF2	11	0.946	0.958	0.936	0.907	0.941	0.890	0.949	0.891	0.712	0.903
	13	0.946	0.956	0.935	0.910	0.938	0.857	0.950	0.894	0.692	0.897
	15	0.943	0.955	0.937	0.907	0.940	0.890	0.949	0.896	0.667	0.898
	17	0.942	0.955	0.932	0.905	0.937	0.872	0.947	0.893	0.652	0.893
	19	0.942	0.954	0.928	0.908	0.938	0.887	0.946	0.897	0.680	0.898
	mean	0.944	0.956	0.934	0.907	0.939	0.879	0.948	0.894	0.681	
CTF2	11	0.804	0.816	0.769	0.874	0.790	0.837	0.892	0.853	0.591	0.803
	13	0.791	0.795	0.757	0.877	0.771	0.860	0.886	0.854	0.593	0.798
	15	0.779	0.776	0.775	0.873	0.762	0.831	0.873	0.857	0.586	0.790
	17	0.772	0.770	0.754	0.876	0.750	0.807	0.867	0.854	0.621	0.786
	19	0.764	0.761	0.748	0.875	0.718	0.815	0.854	0.850	0.623	0.779
	mean	0.782	0.783	0.761	0.875	0.758	0.830	0.874	0.853	0.603	
t-BinCTF	11	0.957	0.966	0.960	0.917	0.952	0.880	0.954	0.905	0.879	0.930
	13	0.961	0.968	0.959	0.921	0.952	0.900	0.954	0.910	0.883	0.934
	15	0.959	0.967	0.965	0.921	0.960	0.906	0.954	0.908	0.897	0.937
	17	0.960	0.967	0.968	0.923	0.957	0.906	0.956	0.908	0.884	0.937
	19	0.965	0.967	0.972	0.929	0.958	0.911	0.955	0.913	0.886	0.939
	mean	0.960	0.967	0.965	0.922	0.956	0.901	0.954	0.909	0.886	
t-BinCTF2	11	0.958	0.967	0.959	0.917	0.950	0.884	0.954	0.905	0.877	0.930
	13	0.961	0.966	0.961	0.922	0.953	0.901	0.954	0.910	0.879	0.934
	15	0.961	0.967	0.964	0.921	0.959	0.908	0.955	0.907	0.890	0.937
	17	0.961	0.968	0.964	0.920	0.957	0.906	0.956	0.908	0.866	0.934
	19	0.965	0.968	0.967	0.927	0.961	0.909	0.956	0.911	0.894	0.940
	mean	0.961	0.967	0.963	0.921	0.956	0.902	0.955	0.908	0.881	
t-CTF2	11	0.885	0.852	0.856	0.879	0.834	0.856	0.897	0.872	0.820	0.861
	13	0.869	0.834	0.838	0.883	0.832	0.884	0.896	0.871	0.812	0.858
	15	0.839	0.822	0.842	0.882	0.834	0.891	0.887	0.870	0.807	0.852
	17	0.839	0.815	0.825	0.881	0.812	0.887	0.883	0.868	0.767	0.842
	19	0.849	0.812	0.840	0.884	0.816	0.889	0.872	0.864	0.785	0.846
	mean	0.856	0.827	0.840	0.882	0.825	0.881	0.887	0.869	0.798	

in each specificity level (column) for each kinase or kinase family is highlighted in yellow.) In any case, the prediction performance shows that our methods are comparable with other kinase-specific prediction tools at least.

IV. CONCLUSIONS

In this paper, we have proposed three new encoding scheme, BinCTF, CTF2 and BinCTF2, and make feature selection on each method. The numerical results show that BinCTF and BinCTF2 preforms much better than CTF ecoding. While, the CTF2 performs not always better than CTF. We conclude that using CTF encoding scheme needn't consider the dummy amino acid which has no really chemical functions. Furthermore, the better performance of BinCTF indicates that we should consider not only the specific amino acid but also the chemical and physical characters of the amino acids when developing new encoding scheme. Besides, feature selection is proved has remarkable meaning in prediction, the average AUC has improved a lot on each methods, so this process is highly recommended. We also compare the performance of different window size when predicting phosphorylation sites. The results show that the window size should be set as 11 or 13, for over long will lead to decline on prediction accuracy.

In this paper, we use ensemble SVM as the classifier for predicting the PTM sites. Future work can focus on compare these encoding schemes using different classifier, such as random forest, BayesSVM, decision tree and Conditional random fields (CRFs) etc.

V. ACKNOWLEDGMENT

This paper is supported by Chinese University scientific fund(No.2011JS039).

978-1-4673-4398-5/12/$31.00 ©2012 IEEE

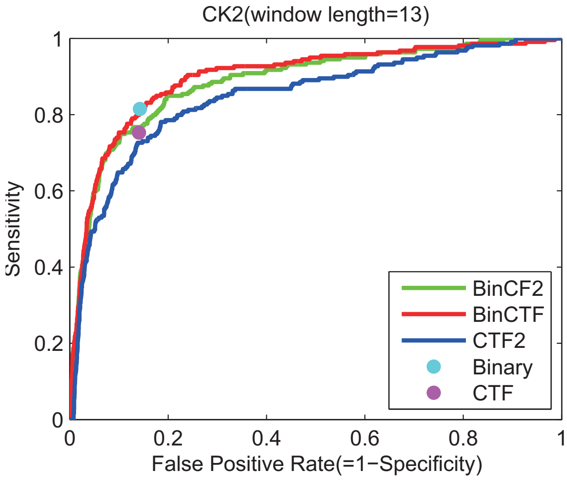

Fig. 4. Roc curve for CK2

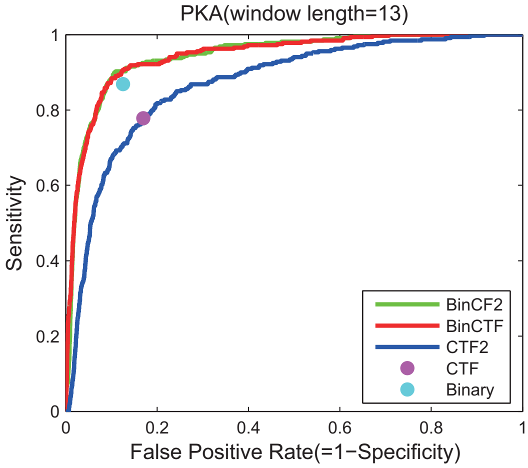

Fig. 5. Roc curve for PKA

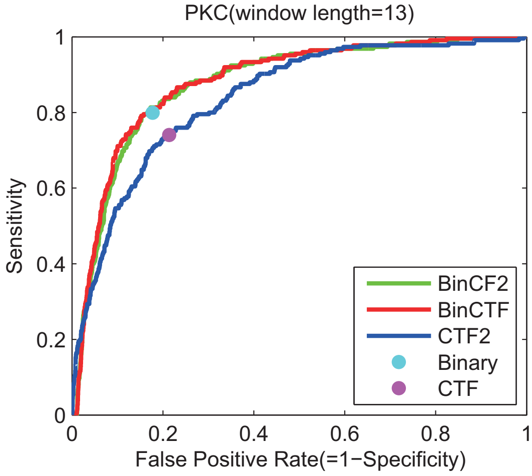

Fig. 6. Roc curve for PKC

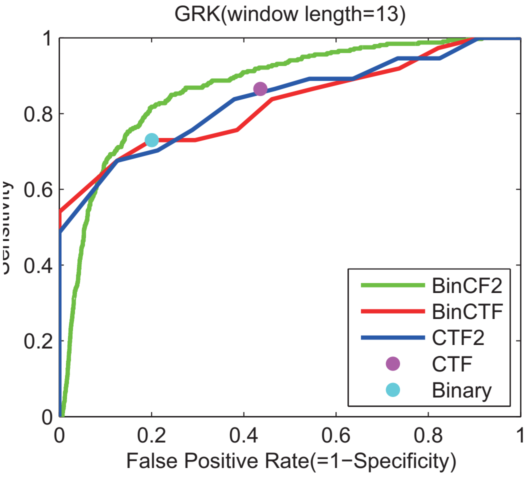

Fig. 7. Roc curve for GRK

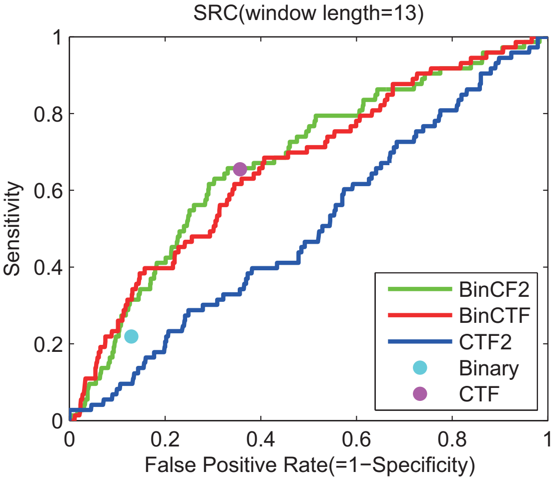

Fig. 8. Roc curve for SRC

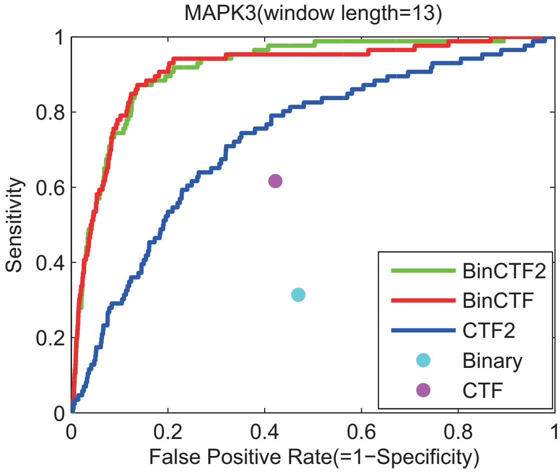

Fig. 9. Roc curve for MAPK3

TABLE IV
PERFORMANCE COMPARISON WITH EXISTING KINASE-SPECIFIC PREDICTION TOOLS

	Sp(%)	99.00	98.00	97.00	95.00	90.00
PKA						
t-BinCTF	Sn(%)	30.72	52.35	62.70	73.35	87.15
t-BinCTF2	Sn(%)	34.80	50.47	63.32	73.98	84.30
GPS2.1	Sn(%)	49.57	58.97	67.52	72.65	83.76
NetPhosK 1.0	Sn(%)	28.04	38.62	48.68	56.08	72.49
Musite	Sn(%)	47.01	58.55	69.23	74.36	85.47
CK2						
t-BinCTF	Sn(%)	20.55	33.33	45.66	61.64	74.89
t-BinCTF2	Sn(%)	18.26	30.59	43.38	61.64	76.26
GPS2.1	Sn(%)	49.56	61.95	68.58	74.34	82.74
NetPhosK 1.0	Sn(%)	37.70	51.31	55.50	62.30	74.35
Musite	Sn(%)	48.67	60.62	66.81	72.12	81.42
MAPK3						
t-BinCTF	Sn(%)	16.28	30.23	46.51	58.14	84.88
t-BinCTF2	Sn(%)	23.26	33.72	41.86	56.98	87.21
GPS2.1	Sn(%)	24.89	40.27	52.04	71.04	81.00
Musite	Sn(%)	27.15	38.46	47.06	63.35	81.90

REFERENCES

[1] Kim J. H.,Lee J.,Oh B.,Kimm K.,Koh I. Prediction of phosphorylation sites using SVMs. *Bioinformatics*. 2004. 20(17):3179-3184.

[2] Yu,C.S., Chen,Y.C., Lu,C.H. and Hwang,J.K. Prediction ofprotein subcellular localization. *Proteins*. 2006.64:643-651.

[3] Wong Y. H.,Lee T. Y.,Liang H. K.,Huang C. M.,Wang T. Y.,Yang Y. H.,Chu C. H.,Huang H. D.,Ko M. T.,Hwang J. K. KinasePhos 2.0: a web server for identifying protein kinase-specific phosphorylation sites based on sequences and coupling patterns. *Nucleic Acids Res*. 2007. 35(2): 588-594.

[4] Yoo P. D.,Ho Y. S.,Zhou B. B.,Zomaya A. Y. SiteSeek: post-translational modification analysis using adaptive locality-effective kernel methods and new profiles. *BMC Bioinformatics*. 2008. 9: 272-289.

[5] Shen J.,Zhang J.,Luo X.,Zhu W.,Yu K.,Chen K.,Li Y.,Jiang H. Predicting protein-protein interactions based only on sequences information. *Proc. Natl. Acad. Sci. U S A*. 2007. 104(11):4337-4341.

[6] Chang,W.C., Lee,T.Y., Shien,D.M., Hsu,J.B., Horng,J.T., Shu,P.C., Wang,T.Y., Pan,R.L., Incorporating supportvector machine for identifying protein tyrosine sites.J.Comput.Chem. 2009. 30(15):2526-2537.

[7] Sheldahl L.C., Park M., Malbon C.C., Moon R.T. Protein kinase C is differentially stimulated by Wnt and Frizzled homologs in a G-protein-dependent manner. *Current Biology*. 1999. 9(13):695-698

[8] Xu Y.,Wang X. B.,Ding J.,Wu L. Y., Deng N. Y. Lysine acetylation sites prediction using an ensemble of support vector machine classifiers. *Theor. Biol*., 2010. 264:130-135.

[9] Henikoff S.,Henikoff J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U S A*. 1992. 89(22):10915-10919.

[10] Tan J.Y., Wu L.Y., Li Y.X., Deng N.Y. Comparison of different encoding scheme on Prediction of Post-translational Modification Sites. In press.

[11] Hyun-Chul Kim,Shaoning Pang, Hong-Mo Je . Constructing support vector machine ensemble.*Volume 36, Issue 12*.December 2003 Pages 2757C2767.

[12] Pang, S.N., Kim, D., Bang, S.Y., Fraud detection using support vector machine ensemble. *ICONIP2001*.2001 pp. 1344C1349.

[13] Wang L., Zhu J., Zou H. Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*2008, 23: 2507-2517.

[14] http://phospho.elm.eu.org/cgimodel.py

[15] She J.W., Zhang J., Luo X.M. Predicting proteinCprotein interactions based only on sequences information.*Proceedings of the National Academy of Sciences*, 2007, 104:4337-4341.

[16] Chang,W.C., Lee,T.Y., Shien,D.M., Hsu,J.B., Horng,J.T., Shu,P.C., Wang,T.Y., Pan,R.L., *Incorporating support vector machine for identifying protein tyrosine sites.* J.Comput.Chem.2009 30:2526-2537.

[17] Vapnik, V. *The Nature of Statistical Learning Theory* . Springer-Verlag, New York.1995.

[18] Burges, C. A tutorial on support vector machines for pattern recognition. *Data Mining Knowl. Dsicov.*1998, 2:121-167.

[19] Yan R., Liu Y., Jin R., Hauptmann A. ON predicting rare classes with SVM ensembles in scene classification. *IEEE International Conference on Acoustics, Speech and Sinal Processing* 2003.

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

[21] Xue, Y., Ren, J., Gao, X., Jin, C., Wen, L., and Yao, X.*GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy.*Proteomics, 2008, 7:1598-1608

[22] Jianjiong Gao, Jay J. Thelen, A. Keith Dunker, and Dong Xu.*Musite, a Tool for Global Prediction of General and Kinase-Specific Phosphorylation Sites.* Molecular & Cellular Proteomics. 2010. 9(12):2586-600.