Research Paper

iPhos-PseEn: Identifying phosphorylation sites in proteins by fusing different pseudo components into an ensemble classifier

Wang-Ren Qiu[1,2], Xuan Xiao[1,3], Zhao-Chun Xu[1], Kuo-Chen Chou[3,4,5]

[1]Computer Department, Jingdezhen Ceramic Institute, Jingdezhen, China

[2]Department of Computer Science and Bond Life Science Center, University of Missouri, Columbia, MO, USA

[3]Gordon Life Science Institute, Boston, MA, USA

[4]Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah, Saudi Arabia

[5]Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu, Sichuan, China

Correspondence to: Wang-Ren Qiu, **email:** qiuone@163.com
Xuan Xiao, **email:** xxiao@gordonlifescience.org
Kuo-Chen Chou, **email:** kcchou@gordonlifescience.org

Keywords: *protein phosphorylation, pseudo components, random forests, ensemble classifier*

Received: *April 05, 2016* ***Accepted:*** *May 23, 2016* ***Published:*** *June 13, 2016*

ABSTRACT

Protein phosphorylation is a posttranslational modification (PTM or PTLM), where a phosphoryl group is added to the residue(s) of a protein molecule. The most commonly phosphorylated amino acids occur at serine (S), threonine (T), and tyrosine (Y). Protein phosphorylation plays a significant role in a wide range of cellular processes; meanwhile its dysregulation is also involved with many diseases. Therefore, from the angles of both basic research and drug development, we are facing a challenging problem: for an uncharacterized protein sequence containing many residues of S, T, or Y, which ones can be phosphorylated, and which ones cannot? To address this problem, we have developed a predictor called iPhos-PseEn by fusing four different pseudo component approaches (amino acids' disorder scores, nearest neighbor scores, occurrence frequencies, and position weights) into an ensemble classifier via a voting system. Rigorous cross-validations indicated that the proposed predictor remarkably outperformed its existing counterparts. For the convenience of most experimental scientists, a user-friendly web-server for iPhos-PseEn has been established at http://www.jci-bioinfo.cn/iPhos-PseEn, by which users can easily obtain their desired results without the need to go through the complicated mathematical equations involved.

INTRODUCTION

Cancer and many other major diseases are often caused by varieties of subtle modifications in biological sequences, typically by various types of post-translational modification (PTM or PTLM) in protein [1, 2], post-replication modification (PTRM) in DNA [3] and post-transcription modification (PTCM) in RNA [4]. In order to reveal the pathological mechanisms of these diseases and find new and revolutionary strategies to treat them, many efforts have been made with the aim to identify the possible modified sites in protein (see, e.g., [5–14], DNA [15, 16], and RNA sequences [17, 18]).

Protein phosphorylation is one of the most-studied post-translational modification (PTM or PTLM) that can alter the structural conformation of a protein, causing it to become activated, deactivated, or modifying its function. The most commonly phosphorylated amino acids are serine (S-type), threonine (T-type), and tyrosine (Y-type).

In human cells, phosphorylation also plays a critical role in the transmission of signals controlling a diverse array of cellular functions, such as cell growth, survival differentiation, and metabolism; while its dysregulation is implicated in many diseases. Therefore, information of phosphorylation sites in proteins is significant for both basic research and drug development.

Many efforts have been made to identify the protein phosphorylation. These methods include mass spectroscopy [19, 20], phosphor-specific antibody [21], etc. Unfortunately, these experimental techniques are both time-consuming and expensive. Facing the explosive growth of protein sequences merging in post genomic age, it is highly desired to develop computational methods for effectively identifying the phosphorylation sites in proteins.

Actually, by using computational approaches such as artificial neural networks, hidden Markov models, and support vector machines, some prediction method were developed based on various different features including disorder scores, KNN scores, amino acid frequency [22, 23], and attribute grouping and position weight amino acid composition [24].

In view of its importance and urgency, it is certainly worthwhile to further improve the prediction quality by introducing some novel approaches as elaborated below.

According to the Chou's 5-step rule [25] and demonstrated in a series of recent publications [11, 12, 18, 26–30], to develop a really useful sequence-based predictor for a biological system, we should stick to the following five guidelines and make them crystal clear: (1) how to construct or select a valid benchmark dataset to train and test the predictor; (2) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their essential correlation with the target concerned; (3) how to introduce or develop a powerful algorithm (or engine) to run the prediction; (4) how to properly conduct cross-validation tests to objectively evaluate the anticipated accuracy; (5) how to provide a web-server and user guide to make people very easily to get their desired results. Below, we are to address the five procedures one-by-one. However, their order may be changed in order to match the rubric style of Oncotarget.

RESULTS AND DISCUSSION

A new ensemble web-server predictor

By fusing four different pseudo component approaches, a new ensemble classifier, named iPhos-PseEn, has been established for predicting phosphorylation sites in proteins.

Success rates and comparison with the existing methods

The success rates achieved by the iPhos-PseEn predictor via the 5-fold cross validation for S-, T- and Y-type phosphorylation are given in Table 1, where for facilitating comparison the corresponding rates by Musite [22] and PWAAC [24] are also listed. As we can see from the table, compared with its counterparts, iPhos-PseEn is remarkably better than its counterparts in predicting all the three phosphorylation types as measured with all the four metrics, clearly indicating that the proposed predictor not only can achieve higher sensitivity, specificity, and overall accuracy but is also much more stable. As shown from the table, compared with Sp, the improvement in Sn is relatively less significant. This is quite normal because the metrics Sn and Sp are used to measure a predictor from two different angles and hence they are actually constrained with each other [28, 31, 32].

Graphical approach is a useful vehicle for analyzing complicated biological systems as demonstrated by a series of previous studies (see, e.g., [33–40]). Here, to provide an intuitive comparison, the graph of Receiver Operating Characteristic (ROC) [41, 42] was utilized to show the advantage of iPhos-PseEn over the Musite [22] and PWAAC [24]. In Figure 1 the green and red graphic lines are the ROC curves for the Musite and PWAAC, respectively; while the blue graphic line for the proposed predictor iPhos-PseEn. The area under the ROC curve is called AUC (area under the curve). The greater the AUC value is, the better the predictor will be [41, 42]. As we can see from Figure 1, the area under the blue curve is remarkably greater than that under the red or green line, once again indicating that the proposed predictor is indeed much better than Musite and PWAAC predictors. Therefore, it is anticipated that iPhos-PseEn will become a useful high throughput tool in this important area, or at the very least, play a complementary role to the existing methods.

Why could the proposed method enhance the prediction quality so significantly? The key is the following. Many important features, which have been proved being closely correlated with phosphorylation sites by previous investigators, such as disorder, nearest neighbor scores, amino acid occurrence frequency, and amino acid position weight, are fused into an ensemble classifier via the general PseAAC approach, as will be elaborated in the Materials and Methods section.

Web server and user guide

As pointed out in two recent review papers [16, 65], a prediction method with its web-server available would practically much more useful. The web-server for iPhos-PseEn has been established. Moreover, to maximize the convenience for users, a step-by-step guide is provided below.

(1) Opening the web-server at http://www.jci-bioinfo.cn/iPhos-PseEn, you will see the top page of iPhos-PseEn on your computer screen, as shown in Figure 2. Click on the Read Me button to see a brief introduction about this predictor.

(2) Either type or copy/paste your query protein sequences into the input box at the center of Figure 2. The input sequences should be in the FASTA

Table 1: A comparison of the proposed predictor with the existing methods based on the 5-fold cross-validation on exactly the same benchmark dataset

Prediction method	Metrics	Type of phosphorylation		
		S	T	Y
Musite[a]	Acc (%)[d]	67.22	77.11	71.60
PWAAC[b]		67.89	66.65	63.04
iPhos-PseEn[c]		79.76	79.88	76.28
Musite[a]	MCC[d]	0.2538	0.2960	0.2472
PWAAC[b]		0.2342	0.2079	0.1720
iPhos-PseEn[c]		0.3901	0.3444	0.3244
Musite[a]	Sn (%)[d]	76.63	68.26	69.58
PWAAC[b]		71.74	69.23	67.70
iPhos-PseEn[c]		79.64	71.51	76.18
Musite[a]	Sp (%)[d]	66.28	77.94	71.79
PWAAC[b]		67.51	66.40	62.61
iPhos-PseEn[c]		79.78	80.68	76.29

[a]The method developed by Gao et al. [22].
[b]The method developed by Huang et al. [24].
[c]The method proposed in this paper.
[d]See Eq.14 for the definition of metrics.

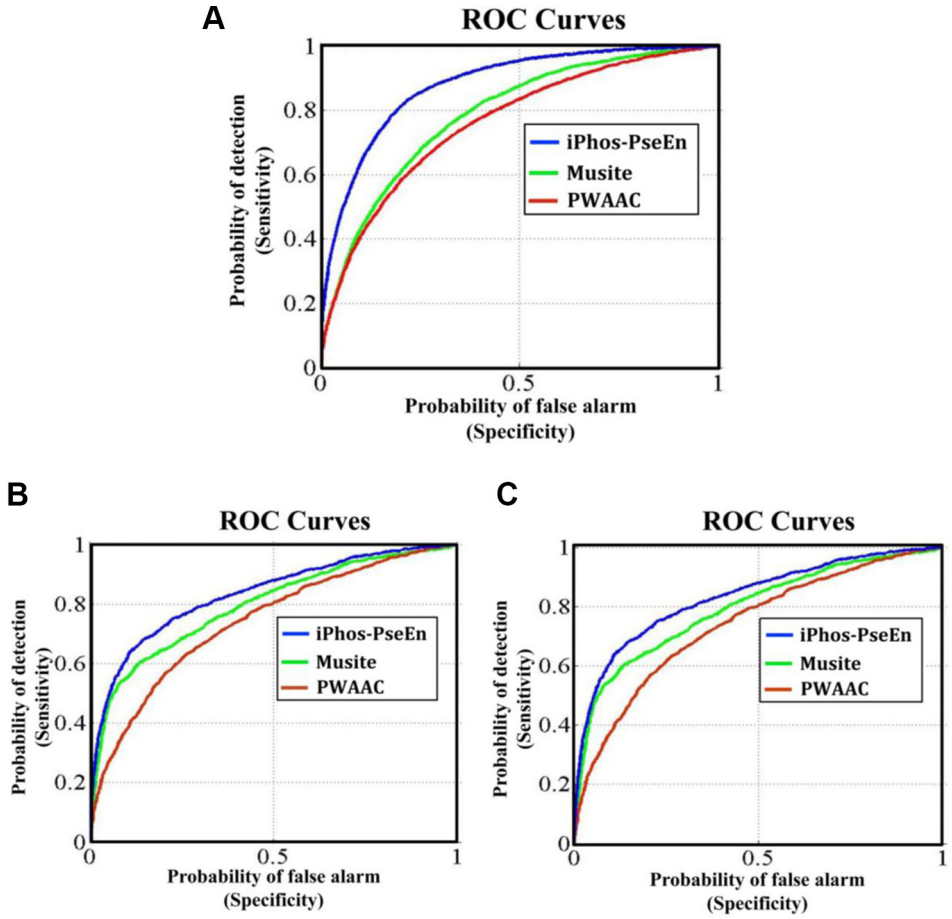

Figure 1: The intuitive graphs of ROC curves to show the performance of Musite, PWAAC, iPhos-PseEn, respectively, for the case of the center residue ⊛ is (A) S, (B) T, and (C) Y. See the main text for further explanation.

format. For the examples of sequences in FASTA format, click the Example button right above the input box.

(3) Select the phosphorylation type concerned: check on the S, T, or Y button to predict phosphoserine, phosphothreonine, or phosphotyrosine, respectively.

(4) Click on the Submit button to see the predicted result. For example, if you use the Sequence_S in the Example window as the input and check on the S button, after 20 seconds or so since your submitting, you will see the following on your screen: Sequence_S contains 11 S residues, of which 2 are predicted to be of phosphorylation site and they are at the sequence positions 2 and 37. If you use the Sequence_T as the input and check on the T button, you will see: Sequence_T contains 11 T residues, of which 5 are of phosphorylation site and at positions 12, 113, 118, 123, and 136. If you use the Sequence_Y as the input and check on the Y button, you will see: Sequence_Y contains 12 Y residues, of which 3 are of phosphorylation site and at the positions 4, 119 and 199. Compared with experimental observations, the above $(11+11+12)=34$ Predicted results contain 1 false negative result (N_{-}^{+}) that is located at 9th S residues in sequence_S, and 4 false positive results (N_{+}^{-}) that are located at the 11th, 123th, 136th residues in Sequence_T as well as the 119th Y residue in Sequence_Y. In other words, the total number of phosphorylation sites involved in the above predictions is $N^{+}=3+2+2=7$, while the total number of non-phosphorylation sites investigated is $N^{-}=8+9+10=27$. Substituting these data into Eq.14, we have $\mathrm{Sn}=85.71\%$, , Sp = 85.19% $\mathrm{Acc}=85.29\%$, and Mcc = 0.6292 quite consistent with the rates in Table 1 obtained by iPhos-PseEn on the benchmark dataset via the 5-fold cross validation test.

(5) As shown on the lower panel of Figure 2, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format of course) via the Browse button. To see the sample of batch input file, click on the button Batch-example.

(6) Click the Supporting Information button to download the benchmark dataset used in this study.

(7) Click the Citation button to find the relevant papers that document the detailed development and algorithm for iPhos-PseEn.

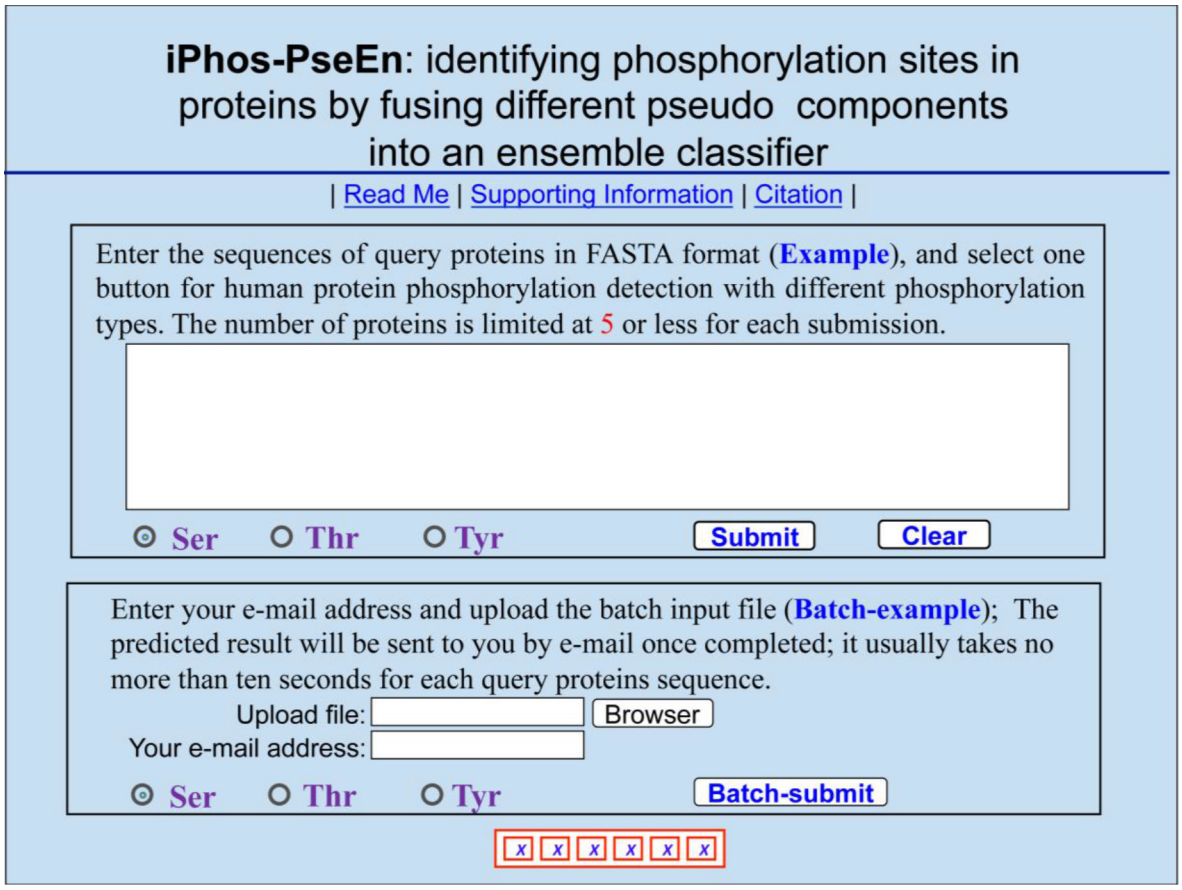

Figure 2: A semi-screenshot to show the top-page of the iPhos-PseEn web-server at http://www.jci-bioinfo.cn/iPhos-PseEn.

MATERIAL AND METHODS

Benchmark dataset

To ensure a high quality, the benchmark dataset used in this study was constructed based on UniProtKB/Swiss-Prot database (released September 2015) at http://www.ebi.ac.uk/uniprot/ according to following the procedures: (1) Open the web site at http://www.uniprot.org/, followed by clicking the button "Advanced". (2) Select "PTM/Processing" and "Modified residue [FT]" for "Fields". (3) Select "Any experimental assertion" for "Evidence". (4) Type "human" for "Term" to do search. (5) Collected were only those proteins that consist of 50 and more amino acid residues to exclude fragments. (6) The proteins thus obtained were subject to a screening operation to remove those sequences that had $\geq 50\%$ pairwise sequence identity to any other.

After strictly following the aforementioned procedures, we finally obtained 1,770 proteins, of which 638 are non-phosphorylated proteins and 1,132 are phosphorylated proteins. The latter contain 845 phosphoserine proteins, 386 phosphothreonine proteins, and 249 phosphotyrosine proteins. Note that some of phosphorylated proteins may be with multi-label, meaning they may belong to more than one type.

For facilitating description later, the Chou's peptide formulation was adopted. The formulation was used to investigate the signal peptide cleavage sites [43], nitrotyrosine sites [9], methylation sites [7], enzyme specificity [44], protein-protein interactions [45], hydroxyproline and hydroxylysine sites [8], and protein-protein binding sites [46]. According to Chou's scheme, a potential phosphorylation site-containing peptide sample can be generally expressed by

$$\mathbf{P}_{\xi}(\circledast) = \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\circledast\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\xi-1)}\mathrm{R}_{+\xi} \quad (1)$$

where the symbol $\circledast$ denotes the single amino acid code S, T, or Y, the subscript ξ is an integer, $\mathrm{R}_{-\xi}$ represents the ξ-th upstream amino acid residue from the center, the $\mathrm{R}_{+\xi}$ the ξ-th downstream amino acid residue, and so forth (Figure 3). The $(2\xi+1)$-tuple peptide sample $\mathbf{P}_{\xi}(\circledast)$ can be further classified into the following two categories:

$$\mathbf{P}_{\xi}(\circledast) \in \begin{cases} \mathbf{P}_{\xi}^{+}(\circledast), & \text{if its center is a phosphorylation site} \\ \mathbf{P}_{\xi}^{-}(\circledast), & \text{other wise} \end{cases} \quad (2)$$

where $\mathbf{P}_{\xi}^{+}(\circledast)$ denotes a true phosphorylation segment with S, T, or Y at its center, $\mathbf{P}_{\xi}^{-}(\circledast)$ denotes a corresponding false phosphorylation segment, and the symbol $\in$ means "a member of" in the set theory.

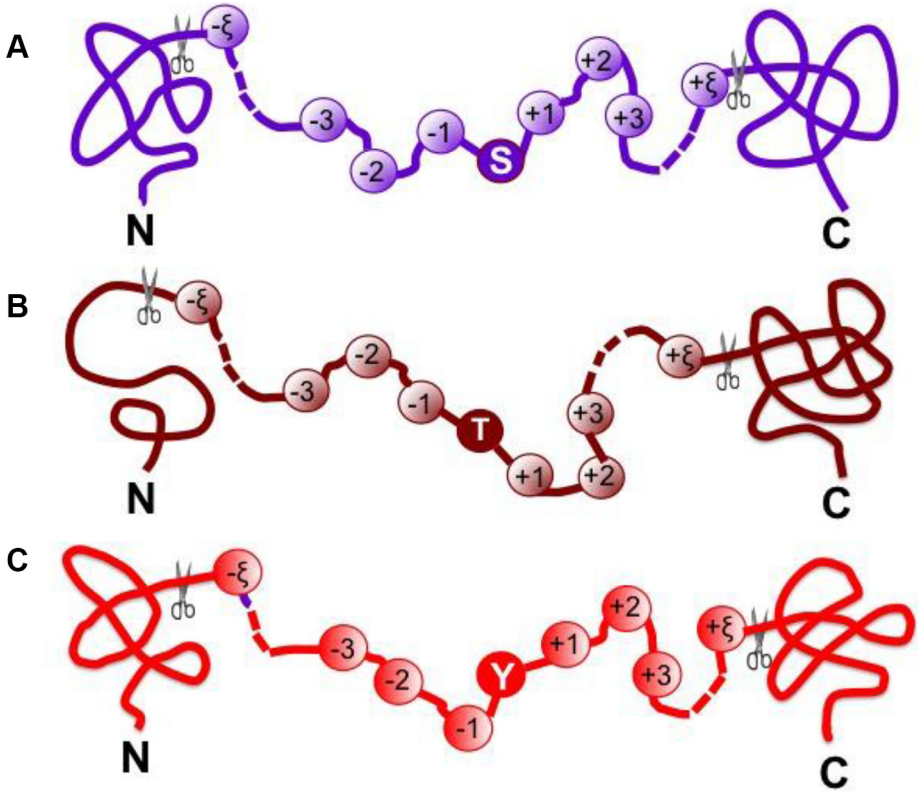

Figure 3: A schematic drawing to show the peptide model $\mathbf{P}_{\xi}(\circledast)$ when (A) $\circledast = \mathbf{S}$, (B) $\circledast = \mathbf{T}$, and (C) $\circledast = \mathbf{Y}$. See Eq.3 as well as the relevant text for further explanation.

In literature the benchmark dataset usually consists of a training dataset and a testing dataset: the former is used for training a model; while the latter, testing the model. But as pointed out in a comprehensive review [47], there is no need to artificially separate a benchmark dataset into the two parts if the prediction model is analyzed with the jackknife test or subsampling (K-fold) cross-validation because the outcome thus obtained is actually from a combination of many different independent dataset tests. Therefore, the benchmark dataset $\mathbb{S}_\xi(\circledast)$ for the current study can be formulated as

$$\begin{cases} \mathbb{S}_\xi(\mathrm{S}) = \mathbb{S}_\xi^+(\mathrm{S}) \bigcup \mathbb{S}_\xi^-(\mathrm{S}), & \text{when } \circledast = \mathrm{S} \\ \mathbb{S}_\xi(\mathrm{T}) = \mathbb{S}_\xi^+(\mathrm{T}) \bigcup \mathbb{S}_\xi^-(\mathrm{T}), & \text{when } \circledast = \mathrm{T} \\ \mathbb{S}_\xi(\mathrm{Y}) = \mathbb{S}_\xi^+(\mathrm{Y}) \bigcup \mathbb{S}_\xi^-(\mathrm{Y}), & \text{when } \circledast = \mathrm{Y} \end{cases} \quad (3)$$

where the positive subset $\mathbb{S}_\xi^+(\circledast)$ only contains the samples of true phosphorylation segments $\mathbf{P}_\xi^+(\circledast)$, and the negative subset $\mathbb{S}_\xi^-(\circledast)$ only contains the samples of false phosphorylation segments $\mathbf{P}_\xi^-(\circledast)$ (see Eq.2); while $\bigcup$ represents the symbol for "union" in the set theory.

The detailed procedures to construct the benchmark dataset are as follows. (1) As done in [48], slide the $(2\xi+1)$-tuple peptide window along each of the aforementioned 1,770 protein sequences, and collected were only those peptide segments that have S, T, and Y at the center. (2) If the upstream or downstream in a protein sequence was less than ξ or greater than $L-\xi$ where L is the length of the protein sequence concerned, the lacking residue was filled with the same residue of its closest neighbor. (3) The peptide segment samples thus obtained were put into the positive subset $\mathbb{S}_\xi^+(\circledast)$ if their centers have been experimentally annotated as the phosphorylation sites; otherwise, into the negative subset $\mathbb{S}_\xi^-(\circledast)$. (4) To reduce redundancy, all those peptide samples were removed if they had pairwise sequence identity with any other.

Note that the length of peptide samples and their number thus generated would depend on the ξ value. Many tests by previous investigators [22–24], however, had indicated that it would be most promising when $\xi = 6$ or the sample's length was $2\xi+1=13$. Accordingly, hereafter we only consider the case of $\xi = 6$; i.e., the samples with 13 amino acid residues. Thus, the benchmark datasets thus obtained for $\mathbb{S}_{\xi=6}(\mathrm{S})$, $\mathbb{S}_{\xi=6}(\mathrm{T})$, and $\mathbb{S}_{\xi=6}(\mathrm{Y})$ are given in Supporting Information S1, S2, and S3, respectively. Listed in Table 2 is a summary of their sizes.

Incorporate extracted features into general pseudo amino acid composition

With the avalanche of biological sequence generated in the post-genomic age, one of the most important problems in computational biology is how to formulate a biological sequence with a discrete model or a vector, yet still considerably keep its sequence order information or essential feature. This is because all the existing machine-learning algorithms can only handle vector but not sequence samples, as elaborated in [16].

To address this problem, the pseudo amino acid composition [49, 50] or PseAAC was proposed. Ever since the concept of pseudo amino acid composition or Chou's PseAAC [51–53] was proposed, it has rapidly penetrated into many biomedicine and drug development areas [54–56] and nearly all the areas of computational proteomics (see, e.g., [57–63] as well as a long list of references cited in [64, 65]).

Because it has been widely and increasingly used, recently three powerful open access soft-wares, called 'PseAAC-Builder' [51], 'propy' [52], and 'PseAAC-General' [64], were established: the former two are for generating various modes of Chou's special PseAAC; while the 3rd one for those of Chou's general PseAAC [25], including not only all the special modes of feature vectors for proteins but also the higher level feature vectors such as "Functional Domain" mode (see Eqs.9-10 of [25]), "Gene Ontology" mode (see Eqs.11-12 of [25]), and "Sequential Evolution" or "PSSM" mode (see Eqs.13-14 of [25]). Inspired by the successes of using PseAAC to deal with protein/peptide sequences, three web-servers [66–68] were developed for generating various feature vectors for DNA/RNA sequences. Particularly, recently a powerful web-server called Pse-in-One [69] has been developed that can be used to generate any desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of users' studies.

According to the general PseAAC [25], the peptide sequence of Eq.1 or Eq.4 can be formulated as

$$\mathbf{P}_{\xi=6}(\circledast) = \left[\Psi_1 \;\; \Psi_2 \;\; \cdots \;\; \Psi_u \;\; \cdots \;\; \Psi_\Omega\right]^{\mathbf{T}} \quad (4)$$

where the components $\Psi_u (u = 1,2,\cdots,\Omega)$ will be defined by how to extract useful features from the relevant protein/peptide sequence, and **T** is the transpose operator.

Disorder Score (DS)

Disorder score is a feature to measure the stability of the local structure. Although disordered region does not have fixed three-dimensional structure in proteins, its

Table 2: Summary of phosphorylation site samples in the benchmark dataset[a]

Subset	Phosphorylation type and number of samples		
	$\circledast$ = S	$\circledast$ = T	$\circledast$ = Y
Positive $\mathbb{S}^{+}_{\xi=6}(\circledast)$	4,317	923	743
Negative $\mathbb{S}^{-}_{\xi=6}(\circledast)$	43,532[b]	9,739[c]	8,061[d]

[a]See Eqs.1-3 and the relevant text for further explanation.
[b]Of the negative samples, 21,564 from the 845 phosphoserine proteins and the 21,968 from the 638 non-phosphorylated proteins.
[c]Of the negative samples, 4,307 from the 386 phosphothreonine proteins and the 5,432 from the 638 non-phosphorylated proteins.
[d]Of the negative samples, 3,968 from the 249 phosphotyrosine proteins and the 4,362 from the 638 non-phosphorylated proteins.

functional importance has been increasingly recognized [70–73]. It was recently used for identifying protein methylation sites [7]. Particularly, it has been observed that the phosphorylation sites have a strong tendency to be located in disordered regions [74]. Using the VSL2 program [75], the disorder score of each amino acid residues in a protein can be calculated and expressed by

$$\mathbb{DS} = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,20} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,20} \\ \vdots & \vdots & \cdots & \vdots \\ d_{k,1} & d_{k,2} & \cdots & d_{k,20} \\ \vdots & \vdots & \vdots & \vdots \\ d_{L,1} & d_{L,2} & \cdots & d_{L,20} \end{bmatrix} \quad (5)$$

where $d_{k,j}$ is the DS score of the k-th amino acid residue $(k = 1,2,\cdots,L)$ when its type is $(j = 1,2,\cdots,20)$. Thus, to reflect the disorder information, the PseAAC of Eq.4 was defined by

$$\mathbf{P}_{\mathrm{DS}} = \left[@_1 \quad @_2 \quad \cdots \quad @_{13} \right]^{\mathrm{T}} \quad (6)$$

where the components are taken from the disorder score matrix of Eq.5 according to the constituent amino acids in Eq.1 as well as their positions in the relevant protein.

K Nearest Neighbor Score (KNNS)

Local sequence clusters often exist around phosphorylation sites because the PTM samples in a same family usually share similar patterns. To reflect this kind of patterns, the PseAAC of Eq.4 was defined by

$$\mathbf{P}_{\mathrm{KNNS}} = \left[\kappa_1 \quad \kappa_2 \quad \kappa_3 \quad \kappa_4 \quad \kappa_5 \right]^{\mathrm{T}} \quad (7)$$

as done in [22, 23] via the BLOSUM62 matrix [76].

Amino Acid Occurrence Frequency (AAOF)

To reflect the amino acid occurrence frequency, the component in Eq.4 are defined by a 20-D vector; i.e.,

$$\mathbf{P}_{\mathrm{AAOF}} = \left[f_1 \quad f_2 \quad \cdots \quad f_{20} \right]^{\mathrm{T}} \quad (8)$$

where f_1 is the occurrence frequency of amino acid A in the relevant 13-tuple peptide sample, f_2 is the occurrence frequency of amino acid C, and so forth (according to the alphabetical order of the single-letter codes for 20 native amino acids).

Position Weight Amino Acid Composition (PWAAC)

Position weight amino acid composition can reveal the sequence-order information around some PTM sites, and it had been used in identifying viral protein phosphorylation sites [24] as well as methylation sites [77]. To reflect this kind of information, the PseAAC of Eq.4 was defined by

$$\mathbf{P}_{\mathrm{PWAAC}} = \left[\mathrm{c}_1 \quad \mathrm{c}_2 \quad \mathrm{c}_3 \quad \mathrm{c}_4 \quad \mathrm{c}_5 \right]^{\mathrm{T}} \quad (9)$$

where

$$\mathrm{c}_i = \frac{1}{\xi(\xi+1)} \sum_{j=-\xi}^{\xi} \delta_{ij} \left(j + \frac{|j|}{\xi} \right) \qquad (j = -\xi,\cdots,\xi) \quad (10)$$

where ξ is the same as in Eq.1, and

$$\delta_{ij} = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{otherwise} \end{cases} \quad (11)$$

Operation engine

Random forests algorithm

Widely used in various areas of computational biology (see, e.g. [11, 12, 27, 45, 46, 78–80]), the random forests (RF) algorithm is a powerful algorithm. Its detailed formulation has been clearly described in [81], and hence there is no need to repeat here.

As shown above, by using DS, KNNS, AAOF, and PWAAC, the sample of Eq.1 can be defined by four different PseAAC vectors, as indicated in Eqs.6, 7, 8, and 9, respectively. Accordingly, we have four different basic RF predictors; i.e.,

$$\begin{cases} \text{RF(1)}, & \text{when the sample is based on DS or Eq.6} \\ \text{RF(2)}, & \text{when the sample is based on KNNS or Eq.7} \\ \text{RF(3)}, & \text{when the sample is based on AAOF or Eq.8} \\ \text{RF(4)}, & \text{when the sample is based on PWAAC or Eq.9} \end{cases} \quad (12)$$

Ensemble random forests

As demonstrated by a series of previous studies, such as signal peptide prediction [82, 83], membrane protein type classification [84, 85], protein subcellular location prediction [86–88], protein fold pattern recognition [89], enzyme functional classification [90], protein-proteins interaction prediction [45], and protein-protein binding site identification [46], the ensemble predictor formed by fusing an array of individual predictors via a voting system can generate much better prediction quality.

Here, the ensemble predictor is formed by fusing the aforementioned four different individual RF predictor of Eq.12; i.e.,

$$\mathbb{RF}^E = RF(1) \forall RF(2) \forall RF(3) \forall RF(4) = \forall_{i=1}^{4} RF(i) \quad (13)$$

where $\mathbb{RF}^E$ denotes the ensemble predictor, and the symbol $\forall$ denotes the fusing operator [47]. In the current study, the concrete fusion process can be described as follows. For a query sample of Eq.1, it would be in turn predicted by RF(1), RF(2), RF(3) and RF(4), respectively. If most outcomes indicated that it belonged to phosphorylation segment, its central residue $\circledast$ was predicted to be phosphorylation site; otherwise, non-phosphorylation site. If there was a tie, the result could be randomly picked between the two. But this kind of tie case rarely happened. For more detailed about this, see a comprehensive review [47] where a crystal clear elucidation with a set of elegant equations are given and hence there is no need to repeat here.

The predictor established via the above procedures is called "iPhos-PseEn", where "i" stands for identify", "Phos" for "phosphorylation site", and "Pse" for "pseudo components", and "En" for "ensemble". Depicted in Figure 4 is a flowchart to show how the ensemble predictor is working.

As mentioned in Introduction, among the five guidelines in developing a useful predictor, one of them is how to objectively evaluate its anticipated success rates [25]. To fulfil this, the following two things need to consider: one is what metrics should be used to measure the predictor's quality; the other is what kind of test method should be taken to derive the metrics rates. Below, let us to address such two problems.

Metrics used to reflect the success rates

A set of four metrics are usually used in literature to measure the quality of a predictor: (1) overall accuracy or Acc; (2) Mathew's correlation coefficient or MCC; (3) sensitivity or Sn; and (4) specificity or Sp [91]. But the conventional formulations for the four metrics are not intuitive, and most experimental scientists feel hard to understand them, particularly for the MCC. Fortunately, if using the symbols introduced by Chou [92] in studying the signal peptides, the set of four metrics can be written as the following forms [5, 93]:

$$\begin{cases} \text{Sn} = 1 - \dfrac{N_-^+}{N^+} & 0 \le \text{Sn} \le 1 \\ \text{Sp} = 1 - \dfrac{N_+^-}{N^-} & 0 \le \text{Sp} \le 1 \\ \text{Acc} = \Lambda = 1 - \dfrac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \text{Acc} \le 1 \\ \text{MCC} = \dfrac{1 - \left(\dfrac{N_-^+}{N^+} + \dfrac{N_+^-}{N^-} \right)}{\sqrt{\left(1 + \dfrac{N_+^- - N_-^+}{N^+} \right) \left(1 + \dfrac{N_-^+ - N_+^-}{N^-} \right)}} & -1 \le \text{MCC} \le 1 \end{cases} \quad (14)$$

where N^+ represents the total number of true-phosphorylation samples investigated, whereas N_-^+ the number of phosphorylation samples incorrectly predicted to be of false- phosphorylation sample; N^- the total number of false-phosphorylation samples, whereas N_+^- the number of false-phosphorylation samples incorrectly predicted to be of true- phosphorylation sample.

According to Eq.14, the following are crystal clear. (1) When $N_-^+ = 0$ meaning none of the true-phosphorylation samples is incorrectly predicted to be of false-phosphorylation sample, we have the sensitivity $\text{Sn} = 1$; whereas $N_-^+ = N^+$ meaning that all the true-phosphorylation samples are incorrectly predicted to be of false-phosphorylation sample, we have the sensitivity $\text{Sn} = 0$. (2) When $N_+^- = 0$ meaning none of the false-phosphorylation samples is incorrectly predicted to be of true-phosphorylation sample, we have the specificity $\text{Sp} = 1$; whereas $N_+^- = N^-$ meaning that all the false-phosphorylation samples are incorrectly predicted to be of true-phosphorylation sample, we have the specificity $\text{Sp} = 0$. (3) When $N_-^+ = N_+^- = 0$ meaning that none of the true-phosphorylation samples in the positive dataset and none of the false-phosphorylation samples in the negative dataset is incorrectly predicted, we have the overall accuracy $\text{Acc} = 1$ and; $\text{MCC} = 1$ whereas $N_-^+ = N^+$ and

$N_+^- = N^-$ meaning that all the true-phosphorylation samples in the positive dataset and all the false-phosphorylation samples in the negative dataset are incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 0$ and $\mathrm{MCC} = -1$. (4) When $N_-^+ = N^+ / 2$ and, $N_+^- = N^- / 2$ we have $\mathrm{Acc} = 0.5$ and $\mathrm{MCC} = 0$ meaning no better than random guessing.

As we can see from the above discussion, the set of metrics formulated in Eq.14 has made the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient much more intuitive and easier-to-understand, particularly for the meaning of MCC, as unanimously concurred and practically applied by many authors in a series of recent publications (see, e.g., [11, 12, 18, 26, 27, 45, 94–102]).

Note that, of the four metrics in Eq.14, the most important are the Acc and MCC: the former reflects the overall accuracy of a predictor; while the latter, its stability in practical applications. The metrics Sn and Sp are used to measure a predictor from two opposite angles. When, and only when, both Sn and Sp of a tested predictor are higher than those of the other tested predictor, we can say the former predictor is better than the latter one.

Also, it is instructive to point out that the set of equations given in Eq.14 is valid for the single-label systems only. As for the multi-label systems whose emergence has become increasingly often in the system biology [103-105] and system medicine [106], a completely different set of metrics is needed as elucidated in [107].

Cross-validation

With a set of intuitive evaluation metrics clearly defined, the next step is what kind of validation method should be used to derive the metrics values.

The following three cross-validation methods are often used in literature: (1) independent dataset test, (2) subsampling (or K-fold cross-validation) test, and (3) jackknife test [108]. Of these three, however, the jackknife test is deemed the least arbitrary that can always yield a unique outcome for a given benchmark dataset as elucidated in [25]. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., [57–59, 109–115]).

In this study, however, to reduce the computational time, we adopted the 5-fold cross-validation method, as done by many investigators with SVM as the prediction engine. Given below is a more rigorous description of 5-fold cross-validation on a benchmark dataset $\mathbb{S}$.

First, randomly divided the benchmark dataset $\mathbb{S}$ into five groups $\mathbb{S}_1$, $\mathbb{S}_2$, $\mathbb{S}_3$, $\mathbb{S}_4$, and $\mathbb{S}_5$, with each having approximately the same number of samples not only for the main-set level but also for all the sub-set levels considered, as can be formulated by

$$\mathbb{S}_1 \triangleq \mathbb{S}_2 \triangleq \mathbb{S}_3 \triangleq \mathbb{S}_4 \triangleq \mathbb{S}_5 \qquad (15)$$

where the symbol $\triangleq$ means that the divided datasets are about the same in size, and so are their subsets [27]. Next,

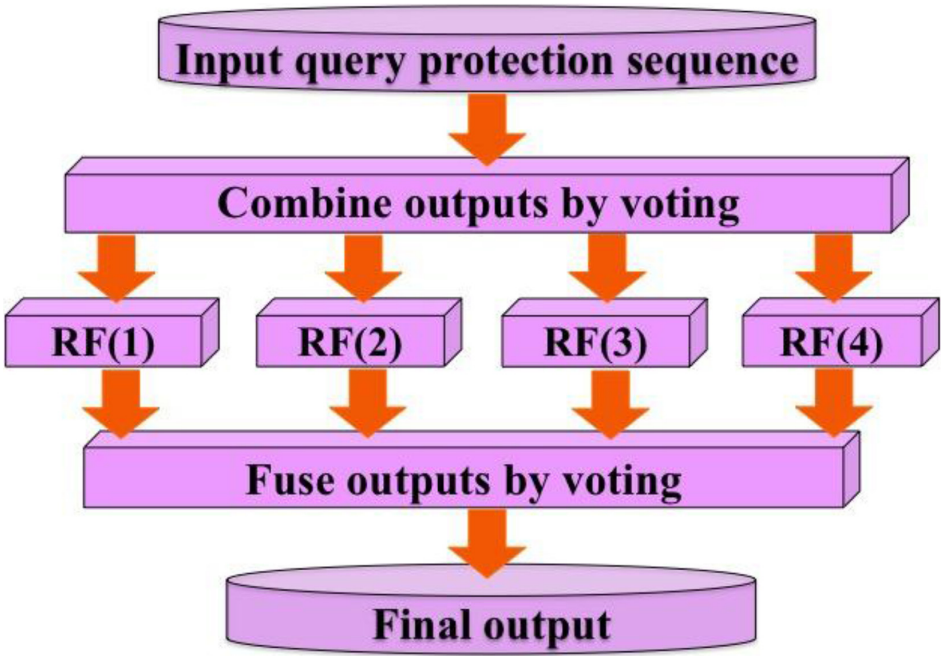

Figure 4: A flow chart to show how the four individual random forest predictors are fused into an ensemble classifier via a voting system. See Eqs.12-13 as well as the relevant text for further explanation.

each of the five sub-benchmark datasets was singled out one-by-one and tested by the model trained with the remaining four sub-benchmark datasets. The cross-validation process was repeated for five times, with their average as the final outcome. In other words, during the process of 5-fold cross-validation, both the training dataset and testing dataset were actually open, and each sub-benchmark datasets was in turn moved between the two. The 5-fold cross-validation test can exclude the "memory" effect, just like conducting 5 different independent dataset tests.

As we can see from Table 2 or Supporting Information S1, S2, and S3, the negative subset $\mathbb{S}^{-}_{\xi=6}(\circledast)$ is much larger than the positive subset $\mathbb{S}^{+}_{\xi=6}(\circledast)$. The ratio is about 10:1 for all the three types of phosphorylation. Although this might reflect the real world in which the non-phosphorylation sites are always the majority compared with the phosphorylation ones, a predictor trained by such a highly skewed benchmark dataset would inevitably have the bias consequence that many phosphorylation sites might be mispredicted as non-phosphorylation ones. To deal with this kind of situation, we randomly divide the negative subset into ten groups with each having about the same size. Thus, for each of the three types of phosphorylation, we have ten benchmark datasets in which the positive and negative samples are about the same. It was based on each of such ten datasets that the 5-fold cross-validation was performed, followed by taking an average for the final score.

CONCLUSIONS

The iPhos-PseEn predictor is a new bioinformatics tool for identifying the phosphorylation sites in proteins. Compared with the existing predictors in this area, its prediction quality is much better, with remarkably higher sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient. For the convenience of most experimental scientists, we have provided its web-server and a step-by-step guide, by which users can easily obtain their desired results without the need to go through the detailed mathematics.

We anticipate that iPhos-PseEn will become a very useful high throughput tool, or at the very least, a complementary tool to the existing methods for predicting the protein phosphorylation sites.

Online Supporting Information

Please refer to "Supporting information S1, Supporting information S2, Supporting information S3" in Supplementary Materials

ACKNOWLEDGMENTS AND FUNDING

The authors wish to thank the six anonymous reviewers for their constructive comments, which were very useful for strengthening the presentation of this paper. This work was partially supported by the National Nature Science Foundation of China (Nos. 61261027, 31260273, 61300139, 31560316), the Natural Science Foundation of Jiangxi Province, China (No. 20142BAB207013), the Scientific Research plan of the Department of Education of JiangXi Province(GJJ14640), the Visiting Scholars Program of State Scholarship Fund (201508360047). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. Foster MW, Hess DT, Stamler JS. Protein S-nitrosylation in health and disease: a current perspective. Trends Mol Med. 2009; 15:391–404.
2. Uehara T, Nakamura T, Yao D, Shi ZQ, Gu Z, Ma Y, Masliah E, Nomura Y, Lipton SA. S-nitrosylated protein-disulphide isomerase links protein misfolding to neurodegeneration. Nature. 2006; 441:513–517.
3. Kobayashi Y, Absher DM, Gulzar ZG, Young SR, McKenney JK, Peehl DM, Brooks JD, Myers RM, Sherlock G. DNA methylation profiling reveals novel biomarkers and important roles for DNA methyltransferases in prostate cancer. Genome Res. 2011; 21:1017–1027.
4. Cantara WA, Crain PF, Rozenski J, McCloskey JA, Harris KA, Zhang X, Vendeix FA, Fabris D, Agris PF. The RNA Modification Database, RNAMDB: 2011 update. Nucleic Acids Res. 2011; 39:D195–201.
5. Xu Y, Ding J, Wu LY. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE. 2013; 8:e55844.
6. Xu Y, Shao XJ, Wu LY, Deng NY. iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins. PeerJ. 2013; 1:e171.
7. Qiu WR, Xiao X, Lin WZ. iMethyl-PseAAC: Identification of Protein Methylation Sites via a Pseudo Amino Acid Composition Approach. Biomed Res Int (BMRI). 2014; 2014:947416.
8. Xu Y, Wen X, Shao XJ. iHyd-PseAAC: Predicting hydroxyproline and hydroxylysine in proteins by

incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int J Mol Sci. 2014; 15:7594–7610.

9. Xu Y, Wen X, Wen LS, Wu LY. iNitro-Tyr: Prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE. 2014; 9:e105018.
10. Qiu WR, Xiao X, Lin WZ. iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model. J Biomol Struct Dyn. 2015; 33:1731–1742.
11. Jia J, Liu Z, Xiao X, Liu B. iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset. Anal Biochem. 2016; 497:48–56.
12. Jia J, Liu Z, Xiao X, Liu B. pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach. J Theor Biol. 2016; 394:223–230.
13. Xu Y. Recent progress in predicting posttranslational modification sites in proteins. Curr Top Med Chem. 2016; 16:591–603.
14. Jia J, Liu Z, Xiao X, Liu B, Chou KC. iCar-PseCp: identify carbonylation sites in proteins by Monto Carlo sampling and incorporating sequence coupled effects into general PseAAC. Oncotarget. 2016; doi:10.18632/oncotarget.9148.
15. Liu Z, Xiao X, Qiu WR. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition. Anal Biochem. (also, Data in Brief, 2015, 4: 87–89). 2015; 474:69–77.
16. Chou KC. Impacts of bioinformatics to medicinal chemistry. Med Chem. 2015; 11:218–234.
17. Chen W, Feng P, Ding H. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition. Anal Biochem. (also, Data in Brief, 2015; 5: 376–378). 2015; 490:26–33.
18. Liu Z, Xiao X, Yu DJ, Jia J, Qiu WR. pRNAm-PC: Predicting N-methyladenosine sites in RNA sequences via physical-chemical properties. Anal Biochem. 2016; 497:60–67.
19. Herren AW, Weber DM, Rigor RR, Margulies KB, Phinney BS, Bers DM. CaMKII Phosphorylation of Na(V)1.5: Novel *in Vitro* Sites Identified by Mass Spectrometry and Reduced S516 Phosphorylation in Human Heart Failure. J Proteome Res. 2015; 14:2298–2311.
20. Tanaka K, Soeda M, Hashimoto Y, Takenaka S, Komori M. Identification of phosphorylation sites in Hansenula polymorpha Pex14p by mass spectrometry. FEBS Open Bio. 2013; 3:6–10.
21. Kaufmann H, Bailey JE, Fussenegger M. Use of antibodies for detection of phosphorylated proteins separated by two-dimensional gel electrophoresis. Proteomics. 2001; 1:194–199.
22. Gao J, Thelen JJ, Dunker AK, Xu D. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. Mol Cell Proteomics. 2010; 9:2586–2600.
23. Yao Q, Gao J, Bollinger C, Thelen JJ, Xu D. Predicting and analyzing protein phosphorylation sites in plants using musite. Frontiers in plant science. 2012; 3:186.
24. Huang SY, Shi SP, Qiu JD, Liu MC. Using support vector machines to identify protein phosphorylation sites in viruses. J Mol Graphics Modell. 2015; 56:84–90.
25. Chou KC. Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review). J Theor Biol. 2011; 273:236–247.
26. Liu B, Fang L, Long R, Lan X. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition Bioinformatics. 2016; 32:362–389.
27. Jia J, Liu Z, Xiao X, Liu B. iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecules. 2016; 21:95.
28. Liu B, Long R. iDHS-EL: Identifying DNase I hypersensi-tivesites by fusing three different modes of pseudo nucleotide composition into an en-semble learning framework. Bioinformatics. 2016; doi:10.1093/bioinformatics/btw186.
29. Xiao X, Ye HX, Liu Z, Jia JH, Chou KC. iROS-gPseKNC: predicting replication origin sites in DNA by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. Oncotarget. 2016; doi:10.18632/oncotarget.9057.
30. Qiu WR, Sun BQ, Xiao X. iPhos-PseEvo: Identifying human phosphorylated proteins by incorporating evolutionary information into general PseAAC via grey system theory Molecular Informatics. 2016; doi:10.1002/minf.201600010.
31. Chou KC. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem. 1993; 268:16938–16948.
32. Liu B, Wang S, Long R. iRSpot-EL: identify recombination spots with an ensemble learning approach. Bioinformatics. 2016; in press.
33. Chou KC, Forsen S. Graphical rules for enzyme-catalyzed rate laws. Biochem J. 1980; 187:829–835.
34. Zhou GP, Deng MH. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem J. 1984; 222:169–176.
35. Chou KC. Graphic rules in steady and non-steady enzyme kinetics. J Biol Chem. 1989; 264:12074–12079.
36. Althaus IW, Chou JJ, Gonzales AJ, Diebel MR, Kezdy FJ, Romero DL, Tarpley WG, Reusser F. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry. 1993; 32:6548–6554.
37. Althaus IW, Gonzales AJ, Kezdy FJ, Romero DL, Aristoff PA, Tarpley WG, Reusser F. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J Biol Chem. 1993; 268:14875–14880.
38. Wu ZC, Xiao X. 2D-MH: A web-server for generating graphic representation of protein sequences based on the physicochemical properties of their constituent amino acids. J Theor Biol. 2010; 267:29–34.

39. Chou KC, Lin WZ, Xiao X. Wenxiang: a web-server for drawing wenxiang diagrams. Natural Science. 2011; 3:862–865
40. Zhou GP. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. J Theor Biol. 2011; 284:142–148.
41. Fawcett JA. An Introduction to ROC Analysis. Pattern Recognition Letters. 2005; 27:861–874.
42. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning: ACM), pp. 2006; 233–240.
43. Chou KC. Using subsite coupling to predict signal peptides. Protein Eng. 2001; 14:75–79.
44. Chou KC. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Sci. 1995; 4:1365–1383.
45. Jia J, Liu Z, Xiao X. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J Theor Biol. 2015; 377:47–56.
46. Jia J, Liu Z, Xiao X, Liu B. Identification of protein-protein binding sites by incorporating the physicochemical properties and stationary wavelet transforms into pseudo amino acid composition (iPPBS-PseAAC). J Biomol Struct Dyn. 2015; doi:10.1080/07391102.2015.1095116.
47. Chou KC, Shen HB. Review: Recent progresses in protein subcellular location prediction. Anal Biochem. 2007; 370:1–16.
48. Chou KC. Prediction of signal peptides using scaled window. Peptides. 2001; 22:1973–1979.
49. Chou KC. Prediction of protein cellular attributes using pseudo amino acid composition. Proteins Struct Funct Genet. (Erratum: ibid, 2001, Vol44, 60). 2001; 43:246–255.
50. Chou KC. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics. 2005; 21:10–19.
51. Du P, Wang X, Xu C, Gao Y. PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal Biochem. 2012; 425:117–119.
52. Cao DS, Xu QS, Liang YZ. propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics. 2013; 29:960–962.
53. Lin SX, Lapointe J. Theoretical and experimental biology in one—A symposium in honour of Professor Kuo-Chen Chou's 50th anniversary and Professor Richard Giegé's 40th anniversary of their scientific careers. J Biomedical Science and Engineering (JBiSE). 2013; 6:435–442.
54. Zhong WZ, Zhou SF. Molecular science for drug development and biomedicine. Int J Mol Sci. 2014; 15:20072–20078.
55. Chou KC. An unprecedented revolution in medicinal science (doi:10.3390/MOL2NET-1-b040). Proceedings of the MOL2NET (International Conference on Multidisciplinary Sciences) 2015; 1:1–10
56. Zhou GP, Zhong WZ. Perspectives in Medicinal Chemistry. Curr Top Med Chem. 2016; 16:381–382.
57. Khan ZU, Hayat M, Khan MA. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J Theor Biol. 2015; 365:197–203.
58. Dehzangi A, Heffernan R, Sharma A, Lyons J, Paliwal K, Sattar A. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J Theor Biol. 2015; 364: 284–294.
59. Kumar R, Srivastava A, Kumari B, Kumar M. Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. J Theor Biol. 2015; 365:96–103.
60. Mondal S, Pai PP. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J Theor Biol. 2014; 356:30–35.
61. Wang X, Zhang W, Zhang Q, Li GZ. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. Bioinformatics. 2015; 31:2639–2645.
62. Kabir M, Hayat M. iRSpot-GAEnsC: identifing recombination spots via ensemble classifier and extending the concept of Chou's PseAAC to formulate DNA samples. Curr Mol Genet Genomics. MGG. 2016; 291:285–296.
63. Ahmad K, Waris M, Hayat M. Prediction of Protein Submitochondrial Locations by Incorporating Dipeptide Composition into Chou's General Pseudo Amino Acid Composition. J Membr Biol. 2016:10.1007/s00232-00015-09868-00238.
64. Du P, Gu S, Jiao Y. PseAAC-General: Fast building various modes of general form of Chou's pseudo-amino acid composition for large-scale protein datasets. Int J Mol Sci. 2014; 15:3495–3506.
65. Chen W, Lin H, Chou KC. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol BioSyst. 2015; 11:2620–2634.
66. Chen W, Lei TY, Jin DC, Lin H. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal Biochem. 2014; 456:53–60.
67. Chen W, Zhang X, Brooker J, Lin H. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics. 2015; 31:119–120.
68. Liu B, Liu F, Fang L, Wang X. repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects. Bioinformatics. 2015; 31:1307–1309.
69. Liu B, Liu F, Wang X, Chen J, Fang L, Chou KC. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 2015; 43:W65–W71.

70. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol. 1999; 293:321–331.
71. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z. Intrinsic disorder and protein function. Biochemistry. 2002; 41:6573–6582.
72. Yoon MK, Venkatachalam V, Huang A, Choi BS, Stultz CM, Chou JJ. Residual structure within the disordered C-terminal segment of p21(Waf1/Cip1/Sdi1) and its implications for molecular recognition. Protein Sci. 2009; 18:337–347.
73. Huang T, He ZS, Cui WR, Cai YD, Shi XH. A Sequence-based Approach for Predicting Protein Disordered Regions. Protein and Peptide Letters. 2013; 20:243–248.
74. Iakoucheva LM, Radivojac P, Brown CJ, O'Connor TR, Sikes JG, Obradovic Z, Dunker AK. The importance of intrinsic disorder for protein phosphorylation. Nucleic Acids Res. 2004; 32:1037–1049.
75. Peng K, Radivojac P, Vucetic S, Dunker AK, Obradovic Z. Length-dependent prediction of protein intrinsic disorder. BMC bioinformatics. 2006; 7:208.
76. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences of the United States of America. 1992; 89:10915–10919.
77. Shi SP, Qiu JD, Sun XY, Suo SB, Huang SY, Liang RP. PMeS: prediction of methylation sites based on enhanced feature encoding scheme. PloS one. 2012; 7:e38772.
78. Kandaswamy KK, Moller S, Suganthan PN, Sridharan S, Pugalenthi G. AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol. 2011; 270:56–62.
79. Lin WZ, Fang JA, Xiao X. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. PLoS ONE. 2011; 6:e24756.
80. Pugalenthi G, Kandaswamy KK, Kolatkar P. RSARF: Prediction of Residue Solvent Accessibility from Protein Sequence Using Random Forest Method. Protein Pept Lett. 2012; 19:50–56.
81. Breiman L. Random forests. Machine learning. 2001; 45:5–32.
82. Chou KC, Shen HB. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem Biophys Res Comm (BBRC). 2007; 357: 633–640.
83. Shen HB. Signal-3L: a 3-layer approach for predicting signal peptide. Biochem Biophys Res Comm (BBRC). 2007; 363:297–303.
84. Shen HB. Using ensemble classifier to identify membrane protein types. Amino Acids. 2007; 32:483–488.
85. Chou KC, Shen HB. MemType-2L: A Web server for predicting membrane proteins and their types by incorporating evolution information through Pse-PSSM. Biochem Biophys Res Comm (BBRC). 2007; 360:339–345.
86. Chou KC, Shen HB. Hum-PLoc: A novel ensemble classifier for predicting human protein subcellular localization. Biochem Biophys Res Commun (BBRC). 2006; 347:150–157.
87. Shen HB. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel. 2007; 20:39–46.
88. Chou KC, Shen HB. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J Proteome Res. 2007; 6:1728–1734.
89. Shen HB, Chou KC. Ensemble classifier for protein fold pattern recognition. Bioinformatics. 2006; 22:1717–1722.
90. Shen HB. EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. Biochem Biophys Res Comm (BBRC). 2007; 364:53–59.
91. Chen J, Liu H, Yang J. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids. 2007; 33: 423–428.
92. Chou KC. Prediction of protein signal sequences and their cleavage sites. Proteins: Struct, Funct, Genet. 2001; 42:136–139.
93. Chen W, Feng PM, Lin H, Chou KC. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 2013; 41:e68.
94. Chen W, Feng PM, Deng EZ. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal Biochem. 2014; 462:76–83.
95. Chen W, Feng PM, Lin H. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed Research International (BMRI). 2014; 2014:623149.
96. Ding H, Deng EZ, Yuan LF, Liu L. iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels. BioMed Research International (BMRI). 2014; 2014:286419.
97. Lin H, Deng EZ, Ding H, Chen W, Chou KC. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 2014; 42:12961–12972.
98. Liu B, Fang L, Liu F, Wang X. Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE. 2015; 10:e0121501.
99. Xiao X, Min JL, Lin WZ, Liu Z. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. J Biomol Struct Dyn. 2015; 33:2221–2233.
100. Chen W, Ding H, Feng P, Lin H, Chou KC. iACP: a sequence-based tool for identifying anticancer peptides. Oncotarget. 2016; 7:16895–16909. doi: 10.18632/oncotarget.7815.
101. Chen W, Feng P, Ding H. Using deformation energy to analyze nucleosome positioning in genomes. Genomics. 2016; 107:69–75.
102. Liu B, Fang L, Liu F, Wang X. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-

pair composition approach. J Biomol Struct Dyn. 2016; 34:223–235.

103. Chou KC, Wu ZC, Xiao X. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol BioSyst. 2012; 8:629–641.
104. Lin WZ, Fang JA, Xiao X. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. Mol BioSyst. 2013; 9:634–644.
105. Xiao X, Wu ZC. iLoc-Virus: A multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J Theor Biol. 2011; 284:42–51.
106. Xiao X, Wang P, Lin WZ. iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal Biochem. 2013; 436:168–177.
107. Chou KC. Some Remarks on Predicting Multi-Label Attributes in Molecular Biosystems. Mol BioSyst. 2013; 9:1092–1100.
108. Chou KC, Zhang CT. Review: Prediction of protein structural classes. Crit Rev Biochem Mol Biol. 1995; 30:275–349.
109. Zhou GP. An intriguing controversy over protein structural class prediction. J Protein Chem. 1998; 17:729–738.
110. Zhou GP, Doctor K. Subcellular location prediction of apoptosis proteins. Proteins: Struct, Funct, Genet. 2003; 50:44–48.
111. Chou KC, Cai YD. Prediction of membrane protein types by incorporating amphipathic effects. J Chem Inf Model. 2005; 45:407–413.
112. Shen HB. Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers. 2007; 85:233–240.
113. Nanni L, Brahnam S, Lumini A. Prediction of protein structure classes by incorporating different protein descriptors into general Chou's pseudo amino acid composition. J Theor Biol. 2014; 360:109–116.
114. Ahmad S, Kabir M, Hayat M. Identification of Heat Shock Protein families and J-protein types by incorporating Dipeptide Composition into Chou's general PseAAC. Computer methods and programs in biomedicine. 2015; 122:165–174.
115. Liu B, Xu J, Fan S, Xu R, Jiyun Zhou J, Wang X. PseDNA-Pro: DNA-binding protein identification by combining Chou's PseAAC and physicochemical distance transformation. Molecular Informatics. 2015; 34:8–17.