Journal of Theoretical Biology 394 (2016) 223–230

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

pSuc-Lys: Predict lysine succinylation sites in proteins with PseAAC and ensemble random forest approach

Jianhua Jia [a,b,c,*], Zi Liu [a], Xuan Xiao [a,c,*], Bingxiang Liu [a], Kuo-Chen Chou [c,d,*]

[a] Computer Department, Jingdezhen Ceramic Institute, Jing-De-Zhen, 333403, China
[b] Computer science, University of Birmingham, B29 2TT, UK
[c] Gordon Life Science Institute, Boston, MA 02478, USA
[d] Center of Excellence in Genomic Medicine Research (CEGMR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

HIGHLIGHTS

- Succinylation plays an important role in regulating various biological processes.
- A novel ensemble classifier has been developed to predict protein succinylation sites.
- It was formed by fusing a series of individual random forest classifiers via a voting system.
- A user-friendly web-server has been established.

ARTICLE INFO

Article history:
Received 20 November 2015
Received in revised form
6 January 2016
Accepted 7 January 2016
Available online 22 January 2016

Keywords:
Lysine succinylation
Sequence-coupling model
General PseAAC
Random downsampling
Ensemble random forest
pSuc-Lys web-server

ABSTRACT

Being one type of post-translational modifications (PTMs), protein lysine succinylation is important in regulating varieties of biological processes. It is also involved with some diseases, however. Consequently, from the angles of both basic research and drug development, we are facing a challenging problem: for an uncharacterized protein sequence having many Lys residues therein, which ones can be succinylated, and which ones cannot? To address this problem, we have developed a predictor called **pSuc-Lys** through (1) incorporating the sequence-coupled information into the general pseudo amino acid composition, (2) balancing out skewed training dataset by random sampling, and (3) constructing an ensemble predictor by fusing a series of individual random forest classifiers. Rigorous cross-validations indicated that it remarkably outperformed the existing methods. A user-friendly web-server for **pSuc-Lys** has been established at http://www.jci-bioinfo.cn/pSuc-Lys, by which users can easily obtain their desired results without the need to go through the complicated mathematical equations involved. It has not escaped our notice that the formulation and approach presented here can also be used to analyze many other problems in computational proteomics.

1. Introduction

In vivo, one of the most efficient biological mechanisms for expanding the genetic code and for regulating cellular physiology is protein post-translational modification (PTM) (Witze et al., 2007; Walsh et al., 2005). Lysine residue in protein can be subjected to many types of PTMs, such as methylation, acetylation, biotinylation, ubiquitination, ubiquitin-like modifications, propionylation and butyrylation, and leading to different complexity of PTM networks. Recently, a new type of PTM, named lysine succinylation, was initially identified by mass spectrometry and protein sequence alignment. Further studies showed that the lysine succinylation responses to different physiological conditions and is evolutionarily conserved (Zhang et al., 2011). In 2013, Park et al. (2013) identified 2565 succinylation sites from 779 proteins and they revealed that lysine succinylation have potential impacts on enzymes involved in mitochondrial metabolism including amino acid degradation, tricarboxylic acid cycle (TCA) and fatty acid metabolism. In histones, lysine succinylation is also present, suggesting that it possibly plays an important role in regulating chromatin structures and functions (Xie et al., 2012; Du et al., 2011). Consequently, identifying the lysine succinylation sites in proteins is vitally important for cellular physiology and pathology

* Corresponding authors at: Gordon Life Science Institute, Boston, MA 02478, USA.
E-mail addresses: jjia@gordonlifescience.org (J. Jia), liuzi189836@163.com (Z. Liu), xxiao@gordonlifescience.org (X. Xiao), lbx1966@163.com (B. Liu), kcchou@gordonlifescience.org (K.-C. Chou).

http://dx.doi.org/10.1016/j.jtbi.2016.01.020
0022-5193/

that can provide very useful information for both basic research and drug development.

Identifying succinylation residues by experimental technique was mainly via mass spectrometry, which was however costly and time-consuming. Therefore, it is highly demanded to develop computational methods to deal with this problem.

Actually, some efforts have been made in this regard (Xu et al., 2015, 2015; Zhao et al., 2015). Since the importance of the topic as well as the urgency of demanding more powerful high-throughput tools in this area, further efforts are definitely needed to enhance the prediction quality. The present study was devoted to developing a more powerful predictor by the pseudo amino acid composition or PseAAC via incorporating a vectorized sequence-coupling model (Chou, 1993) into the general form of pseudo amino acid composition (PseAAC) (Chou, 2011, 2005) and ensemble random forest approach (Shen, 2006; Jia et al., 2015a).

According to the Chou's 5-step rule (Chou, 2011) and demonstrated in a series of recent publications (Chen et al., 2014b; Ding et al., 2014; Lin et al., 2014; Qiu and Xiao, 2014; Liu et al., 2015a, 2015b, 2015c; Chen et al., 2015b), to establish a really useful sequence-based statistical predictor for a biological system, we need to consider the following five procedures: (1) construct or select a valid benchmark dataset to train and test the predictor; (2) formulate the biological sequence samples with an effective mathematical expression that can effectively correlate with the target to be predicted; (3) introduce or develop a powerful algorithm (or operation engine) to calculate the prediction; (4) properly carry out cross-validation tests to objectively evaluate the anticipated accuracy; (5) establish a user-friendly web-server accessible to the public. Below, let us elaborated how to fulfill these steps one-by one.

2. Materials and methods

2.1. Benchmark dataset

The benchmark dataset used in this study was derived from CPLM (Liu et al., 2014), which is a protein lysine modification database. The database contains 2521 lysine succinylation sites and 24,128 non-succinylation sites determined from 896 proteins (Liu et al., 2014). All the protein sequences concerned were derived from the UniProt (UniProt Consortium, 2010). For facilitating description later, the Chou's peptide formulation was adopted. It was used for studying enzyme specificity (Chou, 1995), signal peptide cleavage sites (Chou, 2001a), hydroxyproline and hydroxylysine sites (Xu et al., 2014), methylation sites (Qiu et al., 2014), nitrotyrosine sites (Xu et al., 2014), protein–protein interaction (Jia et al., 2015a), and protein–protein binding sites (Jia et al., 2015b, 2016). According to Chou's scheme, a potential succinylation site-containing peptide sample can be generally expressed by

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

where the center $\mathbb{K}$ represents "lysine", the subscript ξ is an integer, $\mathrm{R}_{-\xi}$ represents the ξ-th upstream amino acid residue from the center, the $\mathrm{N}_{+\xi}$ the ξ-th downstream amino acid residue, and so forth. The $(2\xi+1)$-tuple peptide sample $\mathbf{P}_{\xi}(\mathbb{K})$ can be further classified into the following two categories:

$$\mathbf{P}_{\xi}(\mathbb{K}) \in \begin{cases} \mathbf{P}_{\xi}^{+}(\mathbb{K}), & \text{if its center is } a \text{ succinylation site} \\ \mathbf{P}_{\xi}^{-}(\mathbb{K}), & \text{otherwise} \end{cases} \quad (2)$$

where $\mathbf{P}_{\xi}^{+}(\mathbb{K})$ denotes a true succinylation segment with lysine at its center, $\mathbf{P}_{\xi}^{-}(\mathbb{K})$ a false succinylation segment with lysine at its center, and the symbol $\in$ means "a member of" in the set theory.

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 used for testing the model. But as pointed out in a comprehensive review (Chou and Shen, 2007), there is no need to artificially separate a benchmark dataset into the two parts if the prediction model is examined by 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. Thus, the benchmark dataset set $\mathbb{S}$ for the current study can be formulated as

$$\mathbb{S}_{\xi} = \mathbb{S}_{\xi}^{+} \bigcup \mathbb{S}_{\xi}^{-} \quad (3)$$

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

The detailed procedures to construct $\mathbb{S}_{\xi}$ are as follows. (1) As done in (Shen, 2007a), slide the $(2\xi+1)$-tuple peptide window along each of the 896 protein sequences taken from (Liu et al., 2014), and collected were only those peptide segments that have K (Lys or lysine) at the center (see Eq. (1)). (2) If the upstream or downstream in a protein sequence was less than ξ or greater than $L-\xi$ (L is the length of the protein sequence concerned), the lacking amino acid was filled with its mirror image (Fig. 1). (3) The peptide segment samples thus obtained were put into the positive subset $\mathbb{S}_{\xi}^{+}$ if their centers have been experimentally annotated as the succinylation sites; otherwise, into the negative subset $\mathbb{S}_{\xi}^{-}$. (4) Using the CD-HIT software (Fu et al., 2012), the aforementioned samples were further subject to a screening procedure to winnow those that had $\geq 40\%$ pairwise sequence identity to any other in a same subset. By following the above procedures, we obtained an array of benchmark datasets with different ξ values.

As we can see from Eq. (1), the length of the peptide sample $\mathbf{P}_{\xi}(\mathbb{K})$ can be expressed as

$$L(\xi) = 2\xi + 1 \quad (4)$$

i.e., the benchmark dataset $\mathbb{S}_{\xi}$ with different ξ value will contain peptide segments formed by different number of amino acid residues. Also, it will contain different number of samples as denoted by

$$N(\xi) = N^{+}(\xi) + N^{-}(\xi) \quad (5)$$

where $N^{+}(\xi)$ denotes the number of samples in the positive benchmark dataset $\mathbb{S}_{\xi}^{+}$, while $N^{-}(\xi)$ the number of samples in the negative benchmark dataset $\mathbb{S}_{\xi}^{-}$.

Preliminary tests, however, had indicated that it would be most promising when $\xi = 15$. Under such case, we have $L(\xi) = 31$, $N^{+}(\xi) = 1,167$, and $N^{-}(\xi) = 3,553$. The detailed sequences for the 1167 samples in the positive subset $\mathbb{S}_{\xi}^{+}$ are given in Supporting information S1, and those for the 3553 samples in the negative subset $\mathbb{S}_{\xi}^{-}$ given in Supporting information S2.

(a) Mirror image for N terminus

$$\mathbf{R_{-1}R_{-2}}\cdots\mathbf{R_{-(\xi-1)}R_{-\xi}} \Leftrightarrow \mathbf{R_{-\xi}R_{-(\xi-1)}}\cdots\mathbf{R_{-2}R_{-1}}$$

(b) Mirror image for C terminus

$$\mathbf{R_{L-\xi}R_{L-\xi+1}}\cdots\mathbf{R_{L-1}R_{L}} \Leftrightarrow \mathbf{R_{L}R_{L-1}}\cdots\mathbf{R_{L-\xi+1}R_{L-\xi}}$$

Fig. 1. Schematic illustration to show the mirror images of the ξ residues for (a) the C-terminus, and (b) the N-terminus. The red symbol $\Leftrightarrow$ represents a mirror, and the real peptide segment is colored in black, while its mirror image in blue. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

2.2. Use general PseAAC to formulate peptide samples

Facing the explosive growth of biological sequence generated in the post-genomic age, one of the most challenging problems in computational biology is how to formulate a biological sequence or sample with a discrete model or a vector, yet still keep considerable sequence pattern or characteristic. This is because all the existing machine-learning algorithms, such as Covariant Discriminant (CD), Neural Network (NN), Support Vector Machine (SVM), K nearest Neighbor (KNN), and Random Forest (RF), can only handle vector but not sequence samples, as elaborated in a recent comprehensive review (Chou, 2015). However, a vector defined in a discrete model may completely lose all the sequence-pattern information. To avoid this, the pseudo amino acid composition (Chou, 2005, 2001c) or PseAAC (Chou, 2009) was proposed. Ever since the concept of pseudo amino acid composition or Chou's PseAAC (Du et al., 2012; Cao et al., 2013; Lin and Lapointe, 2013) was proposed, it has rapidly penetrated into many biomedicine and drug development areas (Zhong and Zhou, 2014) and nearly all the computational proteomics areas (see, e.g., Khan et al., 2015; Dehzangi et al., 2015; Kumar et al., 2015; Mondal and Pai, 2014; Wang et al., 2015; Ahmad et al., 2015; Fan et al., 2015; Huang and Yuan, 2015; Mandal et al., 2015; Sanchez et al., 2015) as well as a long list of references cited in (Du et al., 2014) and a recent review article (Chen and Lin, 2015). Owing to its wide and increasing usage, recently three powerful open access soft-wares, called 'PseAAC-Builder' (Du et al., 2012), 'propy' (Cao et al., 2013), and 'PseAAC-General' (Du et al., 2014), 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 (Chou, 2011), 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) and (10) of Chou (2011)), "Gene Ontology" mode (see Eqs. (11) and (12) of Chou (2011)), and "Sequential Evolution" or "PSSM" mode (see Eqs. (13) and (14) of Chou (2011)). Inspired by the successes of using PseAAC to deal with protein/peptide sequences, three web-servers (Chen et al., 2014c, 2015a; Liu et al., 2015d) were consecutively developed for generating various pseudo components for DNA/RNA sequences as well. Particularly, recently a powerful web-server called Pse-in-One (Liu et al., 2015e) has been established that can be used to generate any desired pseudo components for protein/peptide and DNA/RNA sequences according to users' own need.

Based on the general PseAAC's concept, the peptide sequence of Eq. (1) can be formulated as

$$\mathbf{P}_{\xi}(\mathbb{K}) = \mathbf{P}^{+}_{\xi}(\mathbb{K}) - \mathbf{P}^{-}_{\xi}(\mathbb{K}) \tag{6}$$

where

$$\mathbf{P}^{+}_{\xi}(\mathbb{K}) = \begin{bmatrix} p^{+}_{-\xi}(\mathrm{R}_{-\xi}\,|\,\mathrm{R}_{-(\xi-1)}) \\ p^{+}_{-(\xi-1)}(\mathrm{R}_{-(\xi-1)}\,|\,\mathrm{R}_{-(\xi-2)}) \\ \vdots \\ p^{+}_{-2}(\mathrm{R}_{-2}\,|\,\mathrm{R}_{-1}) \\ p^{+}_{-1}(\mathrm{R}_{-1}) \\ p^{+}_{+1}(\mathrm{R}_{+1}) \\ p^{+}_{+2}(\mathrm{R}_{+2}\,|\,\mathrm{R}_{+1}) \\ \vdots \\ p^{+}_{+(\xi-1)}(\mathrm{R}_{+(\xi-1)}\,|\,\mathrm{R}_{+(\xi-2)}) \\ p^{+}_{+\xi}(\mathrm{R}_{+\xi}\,|\,\mathrm{R}_{+(\xi-1)}) \end{bmatrix} \tag{7}$$

and

$$\mathbf{P}^{-}_{\xi}(\mathbb{K}) = \begin{bmatrix} p^{-}_{-\xi}(\mathrm{R}_{-\xi}\,|\,\mathrm{R}_{-(\xi-1)}) \\ p^{-}_{-(\xi-1)}(\mathrm{R}_{-(\xi-1)}\,|\,\mathrm{R}_{-(\xi-2)}) \\ \vdots \\ p^{-}_{-2}(\mathrm{R}_{-2}\,|\,\mathrm{R}_{-1}) \\ p^{-}_{-1}(\mathrm{R}_{-1}) \\ p^{-}_{+1}(\mathrm{R}_{+1}) \\ p^{-}_{+2}(\mathrm{R}_{+2}\,|\,\mathrm{R}_{+1}) \\ \vdots \\ p^{-}_{+(\xi-1)}(\mathrm{R}_{+(\xi-1)}\,|\,\mathrm{R}_{+(\xi-2)}) \\ p^{-}_{+\xi}(\mathrm{R}_{+\xi}\,|\,\mathrm{R}_{+(\xi-1)}) \end{bmatrix} \tag{8}$$

In Eq. (7) $p^{+}_{-\xi}(\mathrm{R}_{-\xi}\,|\,\mathrm{R}_{-(\xi-1)})$ is the conditional probability of amino acid $\mathrm{R}_{-\xi}$ occurring at the left 1st position (see Eq. (1)) given that its closest right neighbor is $\mathrm{R}_{-(\xi-1)}$, $p^{+}_{-(\xi-1)}(\mathrm{R}_{-(\xi-1)}\,|\,\mathrm{R}_{-(\xi-2)})$ is the conditional probability of amino acid $\mathrm{R}_{(\xi-1)}$ occurring at the left 2nd position given that its closest right neighbor is $\mathrm{R}_{-(\xi-2)}$, and so forth. Note that in Eq. (5), only $p^{+}_{-1}(\mathrm{R}_{-1})$ and $p^{+}_{+1}(\mathrm{R}_{+1})$ are of non-conditional probability since the right neighbor of R_{-1} and the left neighbor of R_{+1} are always K or Lys. All these probability values can be easily derived from the positive benchmark dataset given in Supporting information S1 as done in (Chou, 1996). Likewise, the components in Eq. (8) are the same as those in Eq. (7) except for that they are derived from the negative benchmark dataset given in Supporting information S2.

2.3. Ensemble random forest algorithm

As we can see from Section 2.1 and Eq. (5), $N^{-}(\xi) \gg N^{+}(\xi)$ meaning that the negative subset $\mathbb{S}^{-}_{\xi=15}$ is much larger than the corresponding positive subset $\mathbb{S}^{+}_{\xi=15}$. Although this might reflect the real world in which the non-succinylation sites are always the majority compared with the succinylation ones, a predictor trained by such a highly skewed benchmark dataset would inevitably have the bias consequence that many succinylation sites might be mispredicted as non- succinylation ones (Liu et al., 2015c; Sun et al., 2009; Xiao et al., 2015). Actually, what is really the most desired information for us is the information about the succinylation sites. Therefore, it is important to find an effective approach to minimize this kind of bias consequence. To realize this, the ensemble random forest approach is a good choice.

The random forests (RF) algorithm is a powerful algorithm and has been used in many areas of computational biology (see, e.g. Jia et al. (2015a', 2015b), Kandaswamy et al. (2011); Lin et al. (2011); Pugalenthi et al. (2012)). The detailed procedures and formulation of RF have been very clearly described in (Breiman, 2001), and hence there is no need to repeat here.

Since most classifiers (including RF) are usually working properly for the benchmark datasets consisting of balanced subsets. For the current skewed dataset, we are to use the asymmetric bootstrap approach (Jia et al., 2011) to deal with it. The concrete procedures are as follows.

First of all, to make the working benchmark dataset become a balanced one, we randomly extract $N^{+}(\xi)$ samples from the negative subset. The negative subset thus obtained is denoted by $\mathbb{S}^{-}_{\xi}(1)$, whose size is exactly the same as $\mathbb{S}^{+}_{\xi}$ Repeat the above procedure for m times, generating an array of negative working subsets $\mathbb{S}^{-}_{\xi}(k)$ $(k = 1, 2, \cdots, m)$. Accordingly, we also have an array of working benchmark datasets denoted by

$$\mathbb{S}_{\xi}(k) = \mathbb{S}^{+}_{\xi} \bigcup \mathbb{S}^{-}_{\xi}(k) \quad (k = 1, 2, \cdots, m) \tag{10}$$

Based on each of the above working benchmark datasets, we have an individual random forest classifier denoted by $\mathbb{RF}(k)$.

Encouraged by the facts that using the ensemble classifier formed by fusing many individual classifiers can remarkably enhance the success rates in predicting protein subcellular localization (Shen, 2007b, 2006) and protein quaternary structural attribute (Shen, 2009), here we are also to develop an ensemble classifier by fusing the m individual predictors $\mathbb{RF}(k)$ through a voting system, as formulated by

$$\mathbb{RF}^{\mathrm{E}} = \mathbb{RF}(1) \forall \mathbb{RF}(2) \forall \cdots \forall \mathbb{RF}(m) = \forall_{k=1}^{m} \mathbb{RF}(k) \tag{11}$$

where $\mathbb{RF}^{\mathrm{E}}$ represents the ensemble classifier, and the symbol $\forall$ denotes the fusing operator. For the detailed procedures of how to fuse the results from the m individual predictors to reach a final outcome via the voting system, see Eqs. (30)–(35) in Chou and Shen (2007), where a crystal clear and elegant derivation was elaborated and hence there is no need to repeat here. To provide an intuitive picture, a flowchart is given in Fig. 2 to illustrate how the m individual random forest predictors are fused into the ensemble classifier. In the current study, the number of bagging times was set at 5, i.e., $m = 5$ Also, 50 trees were used for each of the individual predictors $\mathbb{RF}(1)$ $\mathbb{RF}(2)$, $\mathbb{RF}(3)$, $\mathbb{RF}(4)$, and $\mathbb{RF}(5)$ The dimension of the random subspace is the square root of the input vector's dimension (see Eq. (6)); i.e.,

$$D = \mathrm{INT}\left\{\sqrt{2\xi+1}\right\} = \mathrm{INT}\left\{\sqrt{31}\right\} \tag{12}$$

where the symbol INT means taking the integer part for the number within the brackets right after it (Shen, 2006a, 2006b).

The final predictor thus obtained is called "**pSuc-Lys**", where "p" stands for "predict", "Suc" for "succinylation", and "Lys" for "lysine" being the potential modification target.

3. Result and discussion

As pointed out is Section 1, one of the important steps in developing a predictor is how to properly evaluate its anticipated success rates (Chou, 2011). To fulfill this, we need to consider the following two aspects: one is what metrics should be used to quantitatively measure the prediction quality; the other is what validation method should be utilized to calculate the metrics values. Below, we are to address the two problems.

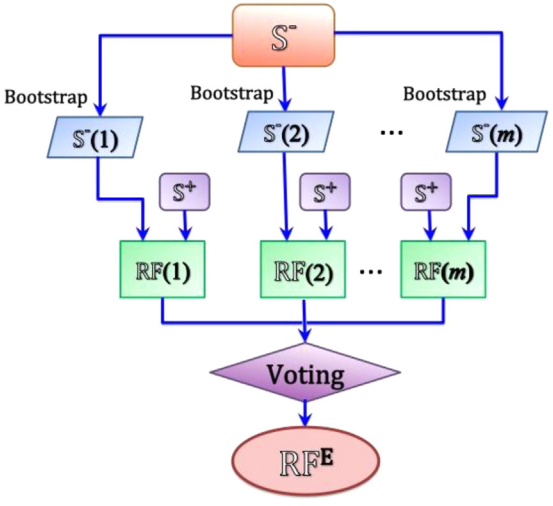

Fig. 2. A flowchart to how the ensemble classifier $\mathbb{RF}^{\mathrm{E}}$ works via a voting system. See the relevant text for further explanation.

3.1. A set of four metrics

In literature the following four metrics are usually used to measure the quality of classification: (1) overall accuracy or Acc; (2) Mathew's correlation coefficient or MCC; (3) sensitivity or Sn; and (4) specificity or Sp (see, e.g., Chen et al., (2007)). Unfortunately, their conventional formulations are not quite intuitive; most experimental scientists feel difficult to understand them, particularly for the one of MCC. Interestingly, by using the Chou's symbols and derivation in the study of signal peptides (Chou, 2001d), the aforementioned four metrics can be easily converted into a set of equations (Xu et al., 2013; Chen et al., 2013) given by

$$\begin{cases} \mathrm{Sn} = 1 - \frac{N_-^+}{N^+} & 0 \le \mathrm{Sn} \le 1 \\ \mathrm{Sp} = 1 - \frac{N_+^-}{N^-} & 0 \le \mathrm{Sp} \le 1 \\ \mathrm{Acc} = \Lambda = 1 - \frac{N_-^+ + N_+^-}{N^+ + N^-} & 0 \le \mathrm{Acc} \le 1 \\ \mathrm{MCC} = \dfrac{1 - \left(\frac{N_-^+}{N^+} + \frac{N_+^-}{N^-}\right)}{\sqrt{\left(1 + \frac{N_+^- - N_-^+}{N^+}\right)\left(1 + \frac{N_-^+ - N_+^-}{N^-}\right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} \tag{13}$$

where N^+ represents the total number of succinylation sites investigated whereas N_-^+ the number of true succinylation sites incorrectly predicted to be of non-succinylation site; N^- the total number of the non-succinylation sites investigated whereas N_+^- the number of non-succinylation sites incorrectly predicted to be of succinylation site.

According to Eq. (13), it is crystal clear to see the following. When $N_-^+ = 0$ meaning none of the true succinylation sites are incorrectly predicted to be of non-succinylation site, we have the sensitivity $\mathrm{Sn} = 1$. When $N_-^+ = N^+$ meaning that all the succinylation sites are incorrectly predicted to be of non-succinylation site, we have the sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_+^- = 0$ meaning none of the non-succinylation sites are incorrectly predicted to be of succinylation site, we have the specificity $\mathrm{Sp} = 1$; whereas $N_+^- = N^-$ meaning that all the non-succinylation sites are incorrectly predicted to be of succinylation sites, we have the specificity $\mathrm{Sp} = 0$. When $N_-^+ = N_+^- = 0$ meaning that none of succinylation sites in the positive dataset and none of the non-succinylation sites in the negative dataset are incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 1$ and $\mathrm{MCC} = 1$; when $N_-^+ = N^+$ and $N_+^- = N^-$ meaning that all the succinylation sites in the positive dataset and all the non-succinylation sites in the negative dataset are incorrectly predicted, we have the overall accuracy $\mathrm{Acc} = 0$ and $\mathrm{MCC} = -1$; whereas when $N_-^+ = N^+/2$ and $N_+^- = N^-/2$ we have $\mathrm{Acc} = 0.5$ and $\mathrm{MCC} = 0$ meaning no better than random guess. Therefore, using Eq. (13) 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 concurred recently by many investigators (see, e.g., Jia et al. (2015a), Chen et al. (2014b), Ding et al. (2014), Liu et al. (2015a', 2015b', 2015c), Chen et al. (2015b), Jia et al. (2015b), Xiao et al. (2015), Chen et al. (2014a), Liu et al. (2015f), Liu et al. (2015g)).

Note that, however, the set of equations defined in Eq. (13) is valid only for the single-label systems. For the multi-label systems whose emergence has become more frequent in system biology (Wu and Xiao, 2012; Lin et al., 2013; Xiao and Wu, 2011) and system medicine (Xiao et al., 2013), a completely different set of metrics is needed as elaborated in (Chou, 2013).

3.2. Cross-validation

With the evaluation metrics available, the next thing is what validation method should be used to derive the metrics values.

In statistical prediction, the following three cross-validation methods are often used to derive the metrics values for a predictor: independent dataset test, subsampling (or K-fold cross-

validation) test, and jackknife test (Chou and Zhang, 1995). 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 (Chou, 2011) and demonstrated by Eqs. (28)–(32) therein. Accordingly, the jackknife test has been widely recognized and increasingly used by investigators to examine the quality of various predictors (see, e.g., Khan et al. (2015); Dehzangi et al. (2015); Kumar et al. (2015); Mondal and Pai (2014); Fan et al. (2015); Shen and Yang (2007); Chou and Cai (2003); Cai (2005)). However, to reduce the computational time, in this study we adopted the *K*-fold cross-validation, as done by most investigators with SVM and random forests algorithms as the prediction engine.

3.3. Comparison with the existing methods

Listed in Table 1 are the values of the four metrics (cf. Eq. (13)) obtained by the current **pSuc-Lys** predictor using the 5-fold cross-validation on the benchmark dataset given in Supporting information S1 and Supporting information S2. For facilitating comparison, listed there are also the corresponding rates achieved by **iSuc-PseAAC** (Xu et al., 2015) and **SucPred** (Zhao et al., 2015), the two existing predictors for identifying the lysine succinylation sites in the aforementioned 896 proteins. As we can see from the table, **pSuc-Lys** remarkably outperformed **iSuc-PseAAC** and **SucPred** in Acc, MCC, and Sn, indicating that, in comparison with the existing methods, the proposed new predictor has better overall accuracy, sensitivity, and stability. Although the Sp by **SucPred** is about 1.2% higher than that by our predictor, the gap between its Sn and Sp is very large ($\approx 48\%$), implying that the results by **SucPred** contains many false negative events (Chen et al., 2007) and hence its high Sp rate is problematic.

By providing intuitive insights, the graphical approach is a useful vehicle for analyzing complicated biological systems as demonstrated by a series of previous studies (see, e.g., Forsen (1980); Zhou and Deng (1984); Chou (1989); Althaus et al. (1993b, 1993a); Wu and Xiao (2010); Lin and Xiao (2011); Zhou (2011)). To provide an intuitive comparison, the graph of Receiver Operating Characteristic (ROC) (Fawcett, 2005; Davis and Goadrich, 2006) was adopted to show the improvement of **pSuc-Lys** over the **iSuc-PseAAC** and **SucPred**. In Fig. 3 the red and green graphic lines are the ROC curves respectively for the **iSuc-PseAAC** and **SucPred** predictors, while the blue graphic line for the proposed predictor **pSuc-Lys**. 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 (Fawcett, 2005; Davis and Goadrich, 2006). As we can see from Fig. 3, the area under the blue curve is remarkably greater than that under the red or green line, clearly indicating that the proposed predictor is indeed better than **iSuc-PseAAC** (Xu et al., 2015) and **SucPred** (Zhao et al., 2015). Therefore, it is anticipated that **pSuc-Lys** may become a useful high throughput tool in this important area, or at the very least, play a complementary role to the existing methods.

Table 1
A comparison of the proposed predictor with the existing methods.

Method	Acc (%)[a]	MCC[a]	Sn (%)[a]	Sp (%)[a]	AUC[b]
iSuc-PseAAC[c]	79.98	0.4370	50.63	89.68	0.7823
SucPred[d]	85.32	0.5710	49.13	97.17	0.8933
pSuc-Lys[e]	90.83	0.7695	76.79	95.97	0.9325

[a] See Eq. (13) for the definition of metrics.
[b] The area under the curve of Fig. 3; the greater the AUC value is, the better the corresponding predictor will be (Fawcett, 2005; Davis and Goadrich, 2006).
[c] The predictor developed in (Xu et al., 2015), where the length of sliding window considered is $L(\xi) = (2\xi+1) = 2 \times 7+1 = 15$ (see Eq. (4)).
[d] The predictor developed in (Zhao et al., 2015), where the length of sliding window considered is 19.
[e] The predictor proposed in this paper; the length of sliding window used is 31.

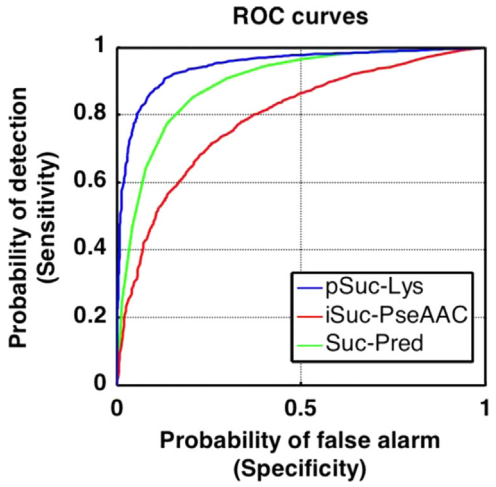

Fig. 3. The intuitive graphs of ROC curves (Fawcett, 2005; Davis and Goadrich, 2006) to show the performance of **iSuc-PseAAC** (Xu et al., 2015), **SucPred** (Zhao et al., 2015), and **pSuc-Lys**, respectively. See the main text for further explanation.

The following points may be the reasons why the proposed predictor can improve the success rates so remarkably. First, its outcome is finalized by fusing the results yielded from an array of individual predictors via a voting system, and hence more stable and reliable as proved in a series of previous studies (Shen, 2007a; 2007b; 2006b; 2009; 2007d; 2010; Chou and Shen, 2006; Liu et al., 2007; Shen, 2007d; Chou and Shen, 2008; Shen and Song, 2009). Second, each of its individual predictors is trained by a balanced benchmark dataset generated through the random sample-extraction procedure, and hence many false prediction events caused by imbalanced and skewed training datasets can be avoided as demonstrated in some recent studies (Liu et al., 2015; Xiao et al., 2015). Third, the coupling effects among the amino acids around the succinylation sites are taken into account via the conditional probability approach as done by previous investigators in successfully enhancing the prediction quality for the HIV protease cleavage sites (Chou, 1993; Zhang, 1993; Zhang and Kezdy, 1993; Zhang, 1993; Tomasselli et al., 1996), specificity of GalNAc-transferase (Chou, 1995; Zhang et al., 1995), and signal peptides and their cleavage sites (Chou, 2001a; Shen, 2007a; Chou, 2001, 2002).

3.4. Web server and user guide

As pointed out in two recent review papers (Chou, 2015; Chen and Lin, 2015), a prediction method with its web-server available will attract more users. The web-server for **pSuc-Lys** has been established. Furthermore, to maximize the convenience for most experimental scientists, a step-to-step guide is provided below.

Step 1. Opening the web-server at http://www.jci-bioinfo.cn/pSuc-Lys, you will see the top page of **pSuc-Lys** on your computer screen, as shown in Fig. 4. Click on the *Read Me* button to see a brief introduction about the predictor.
Step 2. Either type or copy/paste your query protein sequences into the input box at the center of Fig. 4. The input sequence should be in the FASTA format. For the examples of sequences in FASTA format, click the *Example* button right above the input box.
Step 3. Click the *Submit* button to get the predicted result. For example, if you use the two query protein sequences in the *Example* window as the input, after 20 s or so since your submitting, you will see the following on the screen of your computer: (1) Sequence-1 (protein B1XBY6) contains 234 amino acid

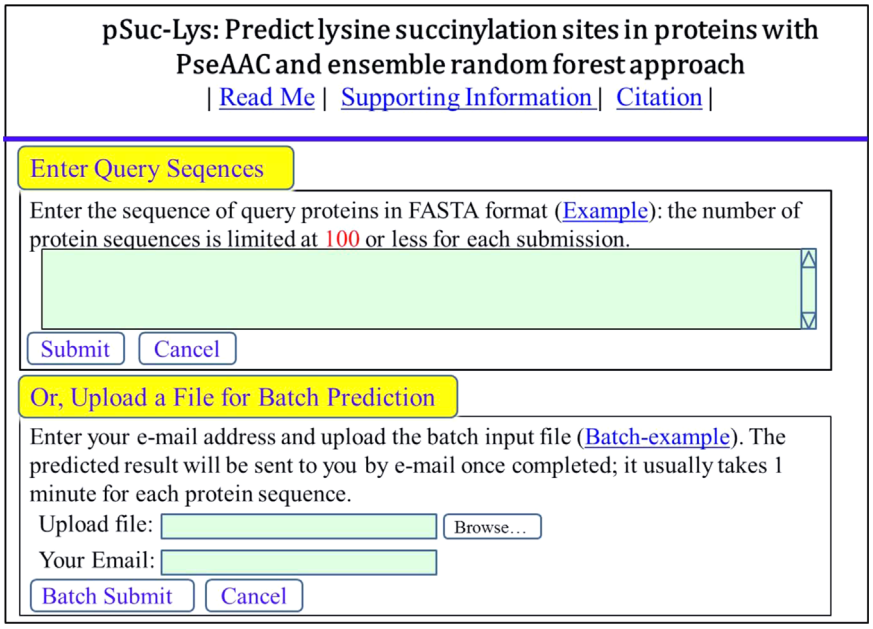

Fig. 4. A semi-screenshot for the top page of the web server **pSuc-Lys** at http://www.jci-bioinfo.cn/iSuc-Lys.

residues, of which 5 are highlighted with red, meaning belonging to succinylation sites. (2) Sequence-2 (protein B1XB26) contains 417 residues, of which 8 are highlighted with red, belonging to succinylation sites. All these predicted results are fully consistent with experimental observations except for residues 3 in sequence-1 that is overpredicted.

Step 4. As shown on the lower panel of Fig. 4, 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*.

Step 5. Click the Supporting information button to download the benchmark dataset used in this study.

Step 6. Click the *Citation* button to find the relevant papers that document the detailed development and algorithm of **pSuc-Lys**.

4. Conclusion

pSuc-Lys is a new bioinformatics tool for predicting the succinylation sites in proteins. Compared with the existing predictors in this area, **pSuc-Lys** can achieve remarkably higher success rates. 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 mathematical formulations. The reason of including them in this paper is for the integrity of the new prediction method, and for that they can be used to stimulate the development of more powerful methods for predicting other PTM sites as well.

We anticipate that **iSuc-Lys** will become a very useful high throughput tool, or at the very least, a complementary tool to the existing methods of predicting the protein succinylation sites.

Acknowledgments

The authors wish to thank the two anonymous reviewers, whose constructive comments are very helpful for strengthening the presentation of this paper. This work was partially supported by the National Natural Science Foundation of China (Nos. 61261027, 31260273, 31560316, 31560316), the Natural Science Foundation of Jiangxi Province, China (No. 20122BAB211033, 20122BAB201044, 20132BAB201053), the Scientific Research plan of the Department of Education of JiangXi Province (GJJ14640). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.jtbi.2016.01.020.

References

Ahmad, S., Kabir, M., Hayat, M., 2015. Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC. Comput. Methods Programs Biomed. 122, 165–174.
Althaus, I.W., Gonzales, A.J., Chou, J.J., Diebel, M.R., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993. The quinoline U-78036 is a potent inhibitor of HIV-1 reverse transcriptase. J. Biol. Chem. 268, 14875–14880.
Althaus, I.W., Chou, J.J., Gonzales, A.J., Kezdy, F.J., Romero, D.L., Aristoff, P.A., Tarpley, W.G., Reusser, F., 1993. Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E. Biochemistry 32, 6548–6554.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Cai, Y.D., 2005. Prediction of membrane protein types by incorporating amphipathic effects. J. Chem. Inf. Model. 45, 407–413.
Cao, D.S., Xu, Q.S., Liang, Y.Z., 2013. propy: a tool to generate various modes of Chou's PseAAC. Bioinformatics 29, 960–962.
Chen, J., Liu, H., Yang, J., 2007. Prediction of linear B-cell epitopes using amino acid pair antigenicity scale. Amino Acids 33, 423–428.
Chen, W., Lin, H., 2015. Pseudo nucleotide composition or PseKNC: an effective formulation for analyzing genomic sequences. Mol. Biosyst. 11, 2620–2634.
Chen, W., Feng, P.M., Lin, H., 2013. iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition. Nucleic Acids Res. 41, e68.
Chen, W., Feng, P.M., Lin, H., 2014. iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition. Biomed. Res. Int. 2014, 623149.
Chen, W., Feng, P.M., Deng, E.Z., 2014. iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. Anal. Biochem. 462, 76–83.
Chen, W., Lei, T.Y., Jin, D.C., 2014. PseKNC: a flexible web-server for generating pseudo K-tuple nucleotide composition. Anal. Biochem. 456, 53–60.
Chen, W., Zhang, X., Brooker, J., Lin, H., 2015. PseKNC-General: a cross-platform package for generating various modes of pseudo nucleotide compositions. Bioinformatics 31, 119–120.

Chen, W., Feng, P., Ding, H., Lin, H., 2015. iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition (also, Data in Brief, 2015, 5: 376–378). Anal. Biochem. 490, 26–33.
Chou, K.C., 1989. Graphic rules in steady and non-steady enzyme kinetics. J.Biol. Chem. 264, 12074–12079.
Chou, K.C., 1993. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J. Biol. Chem. 268, 16938–16948.
Chou, K.C., 1995. A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase. Protein Sci. 4, 1365–1383.
Chou, K.C., 1996. Review: prediction of human immunodeficiency virus protease cleavage sites in proteins. Anal. Biochem. 233, 1–14.
Chou, K.C., 2001. Prediction of protein cellular attributes using pseudo amino acid composition (Erratum: ibid., 2001, Vol. 44, p. 60). Proteins: Struct. Funct. Genet. 43, 246–255.
Chou, K.C., 2001. Prediction of protein signal sequences and their cleavage sites. Proteins: Struct. Funct. Genet. 42, 136–139.
Chou, K.C., 2001. Using subsite coupling to predict signal peptides. Protein Eng. 14, 75–79.
Chou, K.C., 2001. Prediction of signal peptides using scaled window. Peptides 22, 1973–1979.
Chou, K.C., 2002. Review: Prediction of protein signal sequences. Curr. Protein Pept. Sci. 3, 615–622.
Chou, K.C., 2005. Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes. Bioinformatics 21, 10–19.
Chou, K.C., 2009. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. Curr. Proteom. 6, 262–274.
Chou, K.C., 2011. Some remarks on protein attribute prediction and pseudo amino acid composition (50th anniversary year review). J. Theor. Biol. 273, 236–247.
Chou, K.C., 2013. Some remarks on predicting multi-label attributes in molecular biosystems. Mol. Biosyst. 9, 1092–1100.
Chou, K.C., 2015. Impacts of bioinformatics to medicinal chemistry. Med. Chem. 11, 218–234.
Chou, K.C., Zhang, C.T., 1995. Review: prediction of protein structural classes. Crit. Rev. Biochem. Mol. Biol. 30, 275–349.
Chou, K.C., Cai, Y.D., 2003. Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition (Addendum, ibid. 2004, 91, 1085). J. Cell. Biochem. 90, 1250–1260.
Chou, K.C., Shen, H.B., 2006. Predicting protein subcellular location by fusing multiple classifiers. J. Cell. Biochem. 99, 517–527.
Chou, K.C., Shen, H.B., 2007. Review: recent progresses in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Chou, K.C., Shen, H.B., 2008. ProtIdent: a web server for identifying proteases and their types by fusing functional domain and sequential evolution information. Biochem. Biophys. Res. Commun. 376, 321–325.
Davis, J., Goadrich, M., 2006. The relationship between Precision-Recall and ROC curves, Proceedings of the 23rd International Conference on Machine Learning. ACM, pp. 233–240.
Dehzangi, A., Heffernan, R., Sharma, A., Lyons, J., Paliwal, K., Sattar, A., 2015. Gram-positive and Gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC. J. Theor. Biol. 364, 284–294.
Ding, H., Deng, E.Z., Yuan, L.F., Liu, L., Lin, H., 2014. iCTX-Type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels. Biomed. Res. Int. 2014, 286419.
Du, J., Zhou, Y., Su, X., Yu, J.J., Khan, S., Jiang, H., Kim, J., Woo, J., Kim, J.H., Choi, B.H., 2011. Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase. Science 334, 806–809.
Du, P., Gu, S., Jiao, Y., 2014. 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. 15, 3495–3506.
Du, P., Wang, X., Xu, C., Gao, Y., 2012. PseAAC-Builder: a cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions. Anal. Biochem. 425, 117–119.
Fan, G.L., Zhang, X.Y., Liu, Y.L., Nang, Y., Wang, H., 2015. DSPMP: discriminating secretory proteins of malaria parasite by hybridizing different descriptors of Chou's pseudo amino acid patterns. J Comput. Chem. 36, 2317–2327.
Fawcett, J.A., 2005. An introduction to ROC Analysis. Pattern Recognit. Lett. 27, 861–874.
Forsen, S., 1980. Graphical rules for enzyme-catalyzed rate laws. Biochem. J. 187, 829–835.
Fu, L., Niu, B., Zhu, Z., Wu, S., Li, W., 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28, 3150–3152.
Huang, C., Yuan, J.Q., 2015. Simultaneously identify three different attributes of proteins by fusing their three different modes of chou's pseudo amino acid compositions. Protein Pept. Lett. 22, 547–556.
Jia, J., Liu, Z., Xiao, X., 2015a. iPPI-Esml: an ensemble classifier for identifying the interactions of proteins by incorporating their physicochemical properties and wavelet transforms into PseAAC. J. Theor. Biol. 377, 47–56.
Jia, J., Xiao, X., Liu, B., Jiao, L., 2011. Bagging-based spectral clustering ensemble selection. Pattern Recognit. Lett. 32, 1456–1467.
Jia, J., Liu, Z., Xiao, X., Liu, B., 2015b. 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. . http://dx.doi.org/10.1080/07391102.2015.1095116
Jia, J., Liu, Z., Xiao, X., Liu, B., 2016. iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. Molecules 21, 95. http://dx.doi.org/10.3390/molecules21010095.
Kandaswamy, K.K., Martinetz, T., Moller, S., Suganthan, P.N., Sridharan, S., Pugalenthi, G., 2011. AFP-Pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J. Theor. Biol. 270, 56–62.
Khan, Z.U., Hayat, M., Khan, M.A., 2015. Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model. J. Theor. Biol. 365, 197–203.
Kumar, R., Srivastava, A., Kumari, B., Kumar, M., 2015. Prediction of beta-lactamase and its class by Chou's pseudo-amino acid composition and support vector machine. J. Theor. Biol. 365, 96–103.
Lin, H., Deng, E.Z., Ding, H., Chen, W., Chou, K.C., 2014. iPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. Nucleic Acids Res. 42, 12961–12972.
Lin, S.X., Lapointe, J., 2013. 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. Biomed. Sci. and Eng. 6, 435–442.
Lin, W.Z., Xiao, X., 2011. Wenxiang: a web-server for drawing wenxiang diagrams. Nat. Sci. 3, 862–865.
Lin, W.Z., Fang, J.A., Xiao, X., 2011. iDNA-Prot: Identification of DNA Binding Proteins Using Random Forest with Grey Model. PLoS ONE 6, e24756.
Lin, W.Z., Fang, J.A., Xiao, X., 2013. iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins. Mol. Biosyst. 9, 634–644.
Liu, B., Fang, L., Wang, S., Wang, X., 2015. Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy. J. Theor. Biol. 385, 153–159.
Liu, B., Fang, L., Long, R., Lan, X., 2015. iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo k-tuple nucleotide composition. Bioinformatics . http://dx.doi.org/10.1093/bioinformatics/btv604.
Liu, B., Liu, F., Fang, L., Wang, X., 2015. 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 31, 1307–1309.
Liu, B., Fang, L., Liu, F., Wang, X., 2015. iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach. J. Biomol. Struct. Dyn. http://dx.doi.org/10.1080/07391102.2015.1014422
Liu, B., Fang, L., Liu, F., Wang, X., Chen, J., 2015. Identification of real microRNA precursors with a pseudo structure status composition approach. PLoS ONE 10, e0121501.
Liu, B., Liu, F., Wang, X., Chen, J., Fang, L., Chou, K.C., 2015. Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences. Nucleic Acids Res. 43, W65–W71.
Liu, D.Q., Liu, H., Shen, H.B., Yang, J., 2007. Predicting secretory protein signal sequence cleavage sites by fusing the marks of global alignments. Amino Acids 32, 493–496.
Liu, Z., Xiao, X., Qiu, W.R., 2015. iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition (also, Data in Brief, 2015, 4: 87–89). Anal. Biochem. 474, 69–77.
Liu, Z., Wang, Y., Gao, T., Pan, Z., Cheng, H., Yang, Q., Cheng, Z., Guo, A., Ren, J., Xue, Y., 2014. CPLM: a database of protein lysine modifications. Nucleic Acids Res. 42, D531–D536.
Mandal, M., Mukhopadhyay, A., Maulik, U., 2015. Prediction of protein subcellular localization by incorporating multiobjective PSO-based feature subset selection into the general form of Chou's PseAAC. Med. Biol. Eng. Comput. 53, 331–344.
Mondal, S., Pai, P.P., 2014. Chou's pseudo amino acid composition improves sequence-based antifreeze protein prediction. J. Theor. Biol. 356, 30–35.
Park, J., Chen, Y., Tishkoff, D.X., Peng, C., Tan, M., Dai, L., Xie, Z., Zhang, Y., Zwaans, B. M.M., Skinner, M.E., 2013. SIRT5-mediated lysine desuccinylation impacts diverse metabolic pathways. Mol. Cell 50, 919–930.
Pugalenthi, G., Kandaswamy, K.K., Kolatkar, P., 2012. RSARF: prediction of residue solvent accessibility from protein sequence using random forest method. Protein Pept. Lett. 19, 50–56.
Qiu, W.R., Xiao, X., 2014. iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components. Int. J. Mol. Sci. 15, 1746–1766.
Qiu, W.R., Xiao, X., Lin, W.Z., 2014. iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach. Biomed Res. Int. 2014, 947416.
Sanchez, V., Peinado, A.M., Perez-Cordoba, J.L., Gomez, A.M., 2015. A new signal characterization and signal-based Chou's PseAAC representation of protein sequences. J. Bioinform. Comput. Biol., 1550024.
Shen, H.B., 2006. Hum-PLoc: a novel ensemble classifier for predicting human protein subcellular localization. Biochem. Biophys. Res. Commun. 347, 150–157.
Shen, H.B., 2006. Predicting eukaryotic protein subcellular location by fusing optimized evidence-theoretic K-nearest neighbor classifiers. J. Proteom. Res. 5, 1888–1897.
Shen, H.B., 2007. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng. Des. Sel. 20, 561–567.
Shen, H.B., 2007. Signal-CF: a subsite-coupled and window-fusing approach for predicting signal peptides. Biochem. Biophys. Res. Commun. 357, 633–640.
Shen, H.B., 2007. Euk-mPLoc: a fusion classifier for large-scale eukaryotic protein subcellular location prediction by incorporating multiple sites. J. Proteome Res. 6, 1728–1734.

Shen, H.B., 2007. Virus-PLoc: A fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85, 233–240.
Shen, H.B., 2009. QuatIdent: a web server for identifying protein quaternary structural attribute by fusing functional domain and sequential evolution information. J. Proteome Res. 8, 1577–1584.
Shen, H.B., 2010. Virus-mPLoc: a fusion classifier for viral protein subcellular location prediction by incorporating multiple sites. J. Biomol. Struct. Dyn. 28, 175–186.
Shen, H.B., Yang, J., 2007. Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction. Amino Acids 33, 57–67.
Shen, H.B., Song, J.N., 2009. Prediction of protein folding rates from primary sequence by fusing multiple sequential features. J. Biomed. Sci. Eng. 2, 136–143.
Sun, Y., Wong, A.K., Kamel, M.S., 2009. Classification of imbalanced data: a review. Int. J. Pattern Recognit. Artif. Intell. 23, 687–719.
Tomasselli, A.L., Reardon, I.M., Heinrikson, R.L., 1996. Predicting HIV protease cleavage sites in proteins by a discriminant function method. Proteins: Struct. Funct. Genet. 24, 51–72.
UniProt Consortium, 2010. The universal protein resource (UniProt) in 2010. Nucleic acids Res. 38, D142–D148.
Walsh, C.T., Garneau-Tsodikova, S., Gatto, G.J., 2005. Protein posttranslational modifications: the chemistry of proteome diversifications. Angew. Chem. Int. Ed. 44, 7342–7372.
Wang, X., Zhang, W., Zhang, Q., Li, G.Z., 2015. MultiP-SChlo: multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier. Bioinformatics 31, 2639–2645.
Witze, E.S., Old, W.M., Resing, K.A., Ahn, N.G., 2007. Mapping protein post-translational modifications with mass spectrometry. Nat. Methods 4, 798–806.
Wu, Z.C., Xiao, X., 2010. 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. 267, 29–34.
Wu, Z.C., Xiao, X., 2012. iLoc-Hum: Using accumulation-label scale to predict subcellular locations of human proteins with both single and multiple sites. Mol. Biosyst. 8, 629–641.
Xiao, X., Wu, Z.C., 2011. iLoc-Virus: a multi-label learning classifier for identifying the subcellular localization of virus proteins with both single and multiple sites. J. Theor. Biol. 284, 42–51.
Xiao, X., Wang, P., Lin, W.Z., Jia, J.H., 2013. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. Anal. Biochem. 436, 168–177.
Xiao, X., Min, J.L., Lin, W.Z., Liu, Z., 2015. iDrug-Target: predicting the interactions between drug compounds and target proteins in cellular networking via the benchmark dataset optimization approach. J. Biomol. Struct. Dyn. 33, 2221–2233.
Xie, Z., Dai, J., Dai, L., Tan, M., Cheng, Z., Wu, Y., Boeke, J.D., Zhao, Y., 2012. Lysine succinylation and lysine malonylation in histones. Mol. Cell. Proteom. 11, 100–107.
Xu, H.D., Shi, S.P., Wen, P.P., Qiu, J.D., 2015. SuccFind: a novel succinylation sites online prediction tool via enhanced characteristic strategy. Bioinformatics.
Xu, Y., Ding, J., Wu, L.Y., 2013. iSNO-PseAAC: Predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. PLoS ONE 8, e55844.
Xu, Y., Wen, X., Shao, X.J., Deng, N.Y., 2014. iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition. Int. J. Mol. Sci. 15, 7594–7610.
Xu, Y., Wen, X., Wen, L.S., Wu, L.Y., Deng, N.Y., 2014. iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition. PLoS ONE 9, e105018.
Xu, Y., Ding, Y.-X., Ding, J., Lei, Y.-H., Wu, L.-Y., Deng, N.-Y., 2015. iSuc-PseAAC: predicting lysine succinylation in proteins by incorporating peptide position-specific propensity. Sci. Rep. 5.
Zhang, C.T., 1993. An alternate-subsite-coupled model for predicting HIV protease cleavage sites in proteins. Protein Eng. 7, 65–73.
Zhang, C.T., 1993. Studies on the specificity of HIV protease: an application of Markov chain theory. J. Protein Chem. 12, 709–724.
Zhang, C.T., Kezdy, F.J., 1993. A vector approach to predicting HIV protease cleavage sites in proteins. Proteins: Struct. Funct. Genet. 16, 195–204.
Zhang, C.T., Kezdy, F.J., Poorman, R.A., 1995. A vector projection method for predicting the specificity of GalNAc-transferase. Proteins: Struct. Funct. Genet. 21, 118–126.
Zhang, Z., Tan, M., Xie, Z., Dai, L., Chen, Y., Zhao, Y., 2011. Identification of lysine succinylation as a new post-translational modification. Nat. Chem. Biol. 7, 58–63.
Zhao, X., Ning, Q., Chai, H., Ma, Z., 2015. Accurate in silico identification of protein succinylation sites using an iterative semi-supervised learning technique. J. Theor. Biol. 374, 60–65.
Zhong, W.Z., Zhou, S.F., 2014. Molecular science for drug development and biomedicine. Int. J. Mol. Sci. 15, 20072–20078.
Zhou, G.P., 2011. The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein-protein interaction mechanism. J. Theor. Biol. 284, 142–148.
Zhou, G.P., Deng, M.H., 1984. An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways. Biochem. J. 222, 169–176.