ISuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components...

Article *in* Analytical Biochemistry · December 2015

DOI: 10.1016/j.ab.2015.12.009

Analytical Biochemistry 497 (2016) 48–56

Contents lists available at ScienceDirect

Analytical Biochemistry

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

iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset

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

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

ARTICLE INFO

Article history:
Received 10 November 2015
Received in revised form
2 December 2015
Accepted 11 December 2015
Available online 23 December 2015

Keywords:
Lysine succinylation
Sequence-coupling model
PseAAC
Optimize training dataset
Target cross-validation

ABSTRACT

Succinylation is a posttranslational modification (PTM) where a succinyl group is added to a Lys (K) residue of a protein molecule. Lysine succinylation plays an important role in orchestrating various biological processes, but it is also associated with some diseases. Therefore, we are challenged by the following problem from both basic research and drug development: given an uncharacterized protein sequence containing many Lys residues, which one of them can be succinylated, and which one cannot? With the avalanche of protein sequences generated in the postgenomic age, the answer to the problem has become even more urgent. Fortunately, the statistical significance experimental data for succinylated sites in proteins have become available very recently, an indispensable prerequisite for developing a computational method to address this problem. By incorporating the sequence-coupling effects into the general pseudo amino acid composition and using KNNC (*K*-nearest neighbors cleaning) treatment and IHTS (inserting hypothetical training samples) treatment to optimize the training dataset, a predictor called iSuc-PseOpt has been developed. Rigorous cross-validations indicated that it remarkably outperformed the existing method. A user-friendly web-server for iSuc-PseOpt has been established at http://www.jci-bioinfo.cn/iSuc-PseOpt, where users can easily get their desired results without needing to go through the complicated mathematical equations involved.

One of the most efficient biological mechanisms for expanding the genetic code and for regulating cellular physiology is the posttranslational modification (PTM) of proteins [1,2]. Owing to the importance of PTM in basic research and drug development, many efforts have been made with the aim of predicting various PTM sites in proteins (see, e.g., Refs. [3–10] and two review articles [11,12] published recently).

The lysine residue in proteins can undergo many types of PTMs, such as methylation, acetylation, biotinylation, ubiquitination, ubiquitin-like modifications, propionylation, and butyrylation, leading to the remarkable complexity of PTM networks.

Recently, a new type of PTM, called lysine succinylation, was identified by mass spectrometry and protein sequence alignment. It has been shown that lysine succinylation responds to different physiological conditions and is evolutionary conserved [13]. In 2013, Park and coworkers [14] identified 2565 succinylation sites from 779 proteins and revealed that lysine succinylation has potential impacts on enzymes involved in mitochondrial metabolism, including amino acid degradation, tricarboxylic acid (TCA) cycle, and fatty acid metabolism [14]. Lysine succinylation also occurs in histones, suggesting that it may play an important role in regulating chromatin structures and functions as well [15,16]. Accordingly, identification of lysine succinylation sites in proteins is no doubt a crucial topic in cellular physiology and pathology, which can

Abbreviations: PTM, posttranslational modification; PseAAC, pseudo amino acid composition; SVM, support vector machine; KNN, K-nearest neighbors; RF, random forest; 30-D, 30-dimensional; KNNC, K-nearest neighbors cleaning; IHTS, inserting hypothetical training samples; Acc, overall accuracy; MCC, Mathew's correlation coefficient; Sn, sensitivity; Sp, specificity; ROC, receiver operating characteristic; AUC, area under the curve.

* Corresponding author. Gordon Life Science Institute, Boston, MA 02478, USA.
** Corresponding author. Computer Department, Jingdezhen Ceramic Institute, Jingdezhen, 333043, China.
*** Corresponding author. Gordon Life Science Institute., Boston, MA 02478, USA.
E-mail addresses: jjh163yx@163.com (J. Jia), xxiao@gordonlifescience.org (X. Xiao), kcchou@gordonlifescience.org (K.-C. Chou).

http://dx.doi.org/10.1016/j.ab.2015.12.009
0003-2697/© 2015 Elsevier Inc. All rights reserved.

provide very useful information for both biomedical research and drug development.

It is time-consuming and expensive to determine the succinylation residues by purely using the experimental techniques alone. In particular, facing the explosive growth of protein sequences in the postgenomic age, it is highly critical to develop computational tools for timely and effectively identifying the succinylation sites in proteins.

Actually, some computational methods have been proposed (see, e.g., Ref. [17]) for the aforementioned purpose. However, because of 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 current study was initiated in an attempt to address this problem by developing a more powerful predictor via incorporating a vectorized sequence-coupling model [18] into the general form of pseudo amino acid composition (PseAAC) [19].

As shown in a series of recent publications [20–27] in compliance with Chou's five-step rule [19], to establish a really useful sequence-based statistical predictor for a biological system, we should logically follow the five guidelines below and make them crystal clear: (i) how to construct or select a valid benchmark dataset to train and test the predictor, (ii) how to formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted, (iii) how to introduce or develop a powerful algorithm (or engine) to operate the prediction, (iv) how to properly perform cross-validation tests to objectively evaluate its anticipated accuracy, and (v) how to establish a user-friendly webserver that is accessible to the public. Below, we address the aforementioned five procedures one by one.

Materials and methods

Benchmark dataset

The benchmark dataset used in this study was derived from the CPLM, a protein lysine modification database [28]. It contains 2521 lysine succinylation sites and 24,128 non-succinylation sites determined from 896 proteins [28]. All of the corresponding protein sequences were derived from the UniProt [29] database. For facilitating description later, Chou's peptide formulation was adopted. It was used for studying signal peptide cleavage sites [30], HIV protease cleavage sites [18], and protein–protein interaction [25]. According to Chou's scheme, a potential succinylation site-containing peptide sample can be generally expressed by

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

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, $\mathrm{R}_{+\xi}$ represents the ξ-th downstream amino acid residue, and so forth. The $(2\xi+1)$-tuple peptide sample $\mathrm{P}_{\xi}(\mathbb{K})$ can be further classified into the following categories:

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

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

As elaborated in a comprehensive review [31], there is no need at all to separate a benchmark dataset into a training dataset and a testing dataset if the predictor to be developed will be tested by the jackknife test or the subsampling (K-fold) cross-validation test because the outcome obtained in this way is actually from a combination of many different independent dataset tests. Thus, the benchmark dataset $\mathbb{S}_{\xi}$ for the current study can be formulated as

$$\mathbb{S}_{\xi} = \mathbb{S}_{\xi}^{+} \cup \mathbb{S}_{\xi}^{-}, \tag{3}$$

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

Because the length of peptide sample $\mathrm{P}_{\xi}(\mathbb{K})$ is $2\xi+1$ (see Eq. (1)), the benchmark dataset with a different ξ value will contain peptide segments with a different number of amino acid residues, as illustrated below:

$$\text{The length of peptide samples in } \mathbb{S}_{\xi} \\ = \begin{cases} 19 \text{ amino acid esidues,} & \text{if } \xi = 9 \\ 23 \text{ amino acid esidues,} & \text{if } \xi = 11 \\ 27 \text{ amino acid esidues,} & \text{if } \xi = 13 \\ 31 \text{ amino acid esidues,} & \text{if } \xi = 15 \\ 35 \text{ amino acid esidues,} & \text{if } \xi = 17 \\ \vdots & \vdots \end{cases}. \tag{4}$$

The detailed procedures to construct $\mathbb{S}_{\xi}$ are as follows. First, as done in Ref. [32], slide the $(2\xi+1)$-tuple peptide window along each of the 896 protein sequences taken from Ref. [28], and only those peptide segments that have K (Lys or lysine) at the center (see Eq. (1)) were collected. Second, 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). Third, the peptide segment samples obtained in this way were put into the positive subset $\mathbb{S}_{\xi}^{+}$ if their centers have been experimentally annotated as the succinylation sites; otherwise, they were put into the negative subset $\mathbb{S}_{\xi}^{-}$. Fourth, using the CD-HIT software [33], 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 a series of benchmark datasets with different ξ values.

But preliminary tests had indicated that it would be most promising when $\xi = 15$. Accordingly, for further study below, instead of Eq. (3) we shall consider

$$\mathbb{S}_{\xi=15} = \mathbb{S}_{\xi=15}^{+} \cup \mathbb{S}_{\xi=15}^{-}, \tag{5}$$

where the benchmark dataset $\mathbb{S}_{\xi=15}$ contains 4720 $(2\xi+1) = 31$-tuple peptide samples, of which 1167 belong to the positive subset $\mathbb{S}_{\xi=15}^{+}$ and 3553 belong to the negative subset $\mathbb{S}_{\xi=15}^{-}$. For readers' convenience, the detailed sequences of the aforementioned positive and negative samples are given in Supporting

(A) Mirror image for N terminus

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

(B) Mirror image for C terminus

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

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

Information S1 and Supporting Information S2, respectively, in the online supplementary material.

Sample formulation with general PseAAC

With the avalanche of biological sequence generated in the postgenomic age, one of the most challenging problems in computational biology and biomedicine is how to formulate a biological sequence or sample with a discrete model or vector yet still keep a considerable sequence pattern or characteristic. This is because all of the existing machine-learning algorithms, such as NN (neural network), SVM (support vector machine), KNN (*K*-nearest neighbors), and RF (random forest), can handle only vector but not sequence samples, as elaborated in a recent comprehensive review [11]. However, a vector defined in a discrete model or framework may totally lose all of the sequence pattern information. To avoid completely losing the sequence pattern information for proteins, the pseudo amino acid composition [34,35] or PseAAC [36] was proposed. Ever since the concept of pseudo amino acid composition or Chou's PseAAC [37–39] was proposed, it has rapidly penetrated into many biomedicine and drug development areas [40] and nearly all of the areas of computational proteomics (see, e.g., Refs. [41–50] as well as a long list of references cited in Ref. [51] and a recent review article [52]). Because it has been widely and increasingly used, recently three powerful open access software programs, called PseAAC-Builder [37], propy [38], and PseAAC-General [51], were established; the former two are for generating various modes of Chou's special PseAAC, whereas the latter one is for those of Chou's general PseAAC [19], including not only all of 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 Ref. [19]), Gene Ontology mode (see Eqs. (11) and (12) of Ref. [19]), and Sequential Evolution or PSSM mode (see Eqs. (13) and (14) of Ref. [19]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, three web-servers [53–55] were developed for generating various pseudo components for DNA/RNA sequences as well. In particular, recently a powerful web-server called Pse-in-One [56] has been established that can be used to generate any desired pseudo components in a vector for protein/peptide and DNA/RNA sequences according to users' needs.

It is obvious from Eq. (1) that when $\xi = 15$ the corresponding peptide contains $(2\,\xi + 1) = 31$ amino acid residues; that is, it can be reduced to

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_{14}\mathrm{R}_{15}\mathbb{K}\mathrm{R}_{16}\mathrm{R}_{17}\cdots\mathrm{R}_{30}. \tag{6}$$

Thus, according to the general form of PseAAC [19], the samples in the positive subset $\mathbb{S}^+_{\xi=15}$ and negative subset $\mathbb{S}^-_{\xi=15}$ of Eq. (5) can be respectively formulated as

$$\mathbf{P}^+ = \left[\Psi_1^+\ \Psi_2^+\ \cdots\ \Psi_u^+\cdots\ \Psi_\Omega^+\right]^\mathrm{T} \tag{7}$$

and

$$\mathbf{P}^- = \left[\Psi_1^-\ \Psi_2^-\ \cdots\ \Psi_u^-\ \cdots\ \Psi_\Omega^-\right]^\mathrm{T}, \tag{8}$$

where T is the transpose operator and Ω is an integer to reflect the dimension of the PseAAC vector. The value of Ω, as well as the components Ψ_u^+ and Ψ_u^- $(u = 1,\ 2,\ \cdots,\ \Omega)$ therein, will depend on how to extract the desired information from the peptide samples in Eq. (6). In this study, to make P^+ better reflect the intrinsic correlation with the lysine succinylation sites, the components in Eq. (7) are defined by the following sequence-coupling factors via the conditional probability approach as originally proposed in Refs. [18,57] for predicting the HIV protease cleavage sites in proteins:

$$\Psi_u^+ = \begin{cases} p^+(\mathrm{R}_1|\mathrm{R}_2), & \text{if } u = 1 \\ p^+(\mathrm{R}_2|\mathrm{R}_3), & \text{if } u = 2 \\ \vdots & \vdots \\ p^+(\mathrm{R}_{15}), & \text{if } u = 15 \\ p^+(\mathrm{R}_{16}), & \text{if } u = 16 \\ \vdots & \vdots \\ p^+(\mathrm{R}_{29}|\mathrm{R}_{28}), & \text{if } u = 29 \\ p^+(\mathrm{R}_{30}|\mathrm{R}_{29}), & \text{if } u = 30 \end{cases} \quad \Omega = 30, \tag{9}$$

where $p^+(\mathrm{R}_1|\mathrm{R}_2)$ is the conditional probability of amino acid R_1 occurring at the first position given that its right neighbor in the peptide sample (cf. Eq. (6)) is R_2, $p^+(\mathrm{R}_2|\mathrm{R}_3)$ is the conditional probability of amino acid R_2 occurring at the second position given that its right neighbor is R_3, and so forth. Note that in the above equation only $p^+(\mathrm{R}_{15})$ and $p^+(\mathrm{R}_{16})$ are of nonconditional probability given that the left neighbor of R_{15} and the right neighbor of R_{15} are always K or Lys. All of these probability values can be easily derived from the positive benchmark dataset given in Supporting Information S1 in the supplementary material as done in Ref. [18]. Similarly, the components in Eq. (8) are defined by

$$\Psi_u^- = \begin{cases} p^-(\mathrm{R}_1|\mathrm{R}_2), & \text{if } u = 1 \\ p^-(\mathrm{R}_2|\mathrm{R}_3), & \text{if } u = 2 \\ \vdots & \vdots \\ p^-(\mathrm{R}_{15}), & \text{if } u = 15 \\ p^-(\mathrm{R}_{16}), & \text{if } u = 16 \\ \vdots & \vdots \\ p^-(\mathrm{R}_{29}|\mathrm{R}_{28}), & \text{if } u = 29 \\ p^-(\mathrm{R}_{30}|\mathrm{R}_{29}), & \text{if } u = 30 \end{cases} \quad \Omega = 30, \tag{10}$$

where the probability values are derived from the corresponding negative benchmark dataset as given in Supporting Information S2.

Inspired by the concept of discriminant function that has been successfully used by many previous investigators to predict the specificity of GalNAc transferase [58], cysteine S-nitrosylation sites [4], HIV protease cleavage sites [59], hydroxyproline and hydroxylysine [7], tight turns and their types [60], and nitrotyrosine sites [8], here we use the discriminant PseAAC vector to represent a peptide sample; that is, **P** of Eq. (6) is finally formulated as a 30-D (30-dimensional) vector given by

$$\mathbf{P} = [\Psi_1\ \Psi_2\ \cdots\ \Psi_u\ \cdots\ \Psi_{30}]^\mathrm{T}, \tag{11}$$

where

$$\Psi_u = \left(\Psi_u^+ - \Psi_u^-\right), \quad u = 1, 2, \ldots, 30. \tag{12}$$

Optimizing imbalanced training datasets

In the current benchmark dataset $\mathbb{S}_{\xi=15}$ (Eq. (5)), the negative subset $\mathbb{S}^-_{\xi=15}$ is much larger than the corresponding positive subset $\mathbb{S}^+_{\xi=15}$, as can be seen by the following equation:

$$\mathbb{S}_{\xi=15}(4720) = \mathbb{S}^+_{\xi=15}(1167) \cup \mathbb{S}^-_{\xi=15}(3553), \tag{13}$$

where the figures in parentheses denote the sample numbers taken from the “Benchmark dataset” section above. As we can see from the above equation, the number of negative samples is nearly three times the size of the positive samples for the benchmark dataset.

Although this might reflect the real world in which the non-succinylation sites are always the majority compared with the succinylation sites, 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 sites [23,61,62]. Actually, what is really the most intriguing information for us is the information about the succinylation sites. Therefore, it is important to find an effective approach to optimize the unbalanced training dataset and minimize this kind of bias consequence. To realize this, we took the following procedures.

First, we used the KNNC (*K*-nearest neighbors cleaning) treatment to remove some redundant negative samples from the negative subset so as to reduce its statistical noise. The detailed process is as follows. For each of the samples in the negative subset, find its *K*-nearest neighbors, where *K* is an integer approximately equal to the ratio between the negative samples and positive ones. In the current case, $K = 3 \approx 3553/1167$. If one of its three nearest neighbors belongs to the positive subset, remove the negative sample from the negative subset.

Second, we used the IHTS (inserting hypothetical training samples) treatment to add some hypothetical positive samples into the positive subset so as to enhance the ability in identifying the interactive pairs. For details of how to generate the hypothetical training samples, see the Monte Carlo samples expanding approach in Refs. [18,63], the seed propagation approach in Ref. [64], or the SMOTE (synthetic minority oversampling technique) approach in Ref. [65].

After the above two treatments, we can change the original highly skewed training dataset to a balanced training dataset with its positive subset and negative subset having exactly the same size.

It is instructive to point out that the hypothetical samples generated via the IHTS treatment can be expressed only by their feature vectors as defined in Eq. (7) but not the real peptide segment samples as given in Supporting Information S1 of the supplementary material. Nevertheless, it would be perfectly reasonable to do so because the data directly used to train a predictor were actually the samples' feature vectors but not their sequence codes. This is the key to optimize an imbalanced benchmark dataset in the current study, and the rationale of such an interesting approach is further elucidated later.

Random forest operation engine

The random forest algorithm is a powerful algorithm that has been used in many areas of computational biology (see, e.g., Refs. [3,25,66–69]). Detailed procedures and formulation of RF have been very clearly described in Ref. [70], and so there is no need to repeat them here. The RF algorithm with MATLAB code was downloaded from https://code.google.com/p/randomforest-matlab/.

For the current study, all of the involved peptide samples are converted into their 30-D vectors according to the definition of Eq. (11), followed by entering them into the RF classifier. In addition, the classifier's output will indicate whether the center residue K of the query peptide is a succinylation site or a non-succinylation site.

The predictor established in this way is called "iSuc-PseOpt," where "i" stands for "identify," "Suc" stands for "succinylation site," "Opt" stands for "optimizing" training dataset, and "Pse" stands for "pseudo" components.

Results and discussion

As mentioned in the introductory paragraphs, one of the important guidelines in developing a predictor is how to objectively and properly evaluate its anticipated success rates [19]. To realize this, we need to consider two things: one is what metrics should be adopted to quantitatively measure the prediction accuracy, and the other is what test method should be applied to calculate the metrics values. Below, we address the two problems.

A set of four metrics

For measuring the success rates for this kind of binary classifications, a set of four metrics is usually used in the literature: (i) overall accuracy or Acc, (ii) Mathew's correlation coefficient or MCC, (iii) sensitivity or Sn, and (iv) specificity or Sp (see, e.g., Ref. [71]). Unfortunately, their conventional formulations are not quite intuitive and easy-to-be-understood for most experimental scientists, particularly the one for MCC. Quite interesting, however, by using Chou's symbols and derivation in the study of signal peptides [72], the aforementioned four metrics can be converted into a set of equations given by

$$\begin{cases} \mathrm{Sn} = 1 - \dfrac{N_{-}^{+}}{N^{+}} & 0 \le \mathrm{Sn} \le 1 \\[2ex] \mathrm{Sp} = 1 - \dfrac{N_{+}^{-}}{N^{-}} & 0 \le \mathrm{Sp} \le 1 \\[2ex] \mathrm{Acc} = \ \Lambda = 1 - \dfrac{N_{-}^{+} + N_{+}^{-}}{N^{+} + N^{-}} & 0 \le \mathrm{Acc} \le 1 \\[2ex] \mathrm{MCC} = \dfrac{1 - \left(\dfrac{N_{-}^{+}}{N^{+}} + \dfrac{N_{+}^{-}}{N^{-}} \right)}{\sqrt{\left(1 + \dfrac{N_{+}^{-} - N_{-}^{+}}{N^{+}} \right) \left(1 + \dfrac{N_{-}^{+} - N_{+}^{-}}{N^{-}} \right)}} & -1 \le \mathrm{MCC} \le 1 \end{cases} , \tag{14}$$

where N^{+} represents the total number of succinylation sites investigated, N_{-}^{+} represents the number of true succinylation sites incorrectly predicted to be of non-succinylation sites, N^{-} represents the total number of non-succinylation sites investigated, and N_{+}^{-} represents the number of non-succinylation sites incorrectly predicted to be of succinylation sites.

According to Eq. (14), it is very intuitive to see the following. When $N_{-}^{+} = 0$, meaning that none of the true succinylation sites is incorrectly predicted to be of non-succinylation sites, we have sensitivity $\mathrm{Sn} = 1$. When $N_{-}^{+} = N^{+}$, meaning that all of the succinylation sites are incorrectly predicted to be of non-succinylation sites, we have sensitivity $\mathrm{Sn} = 0$. Likewise, when $N_{+}^{-} = 0$, meaning that none of the non-succinylation sites is incorrectly predicted to be of succinylation sites, we have specificity $\mathrm{Sp} = 1$. When $N_{+}^{-} = N^{-}$, meaning that all of the non-succinylation sites are incorrectly predicted to be of succinylation sites, we have 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 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 overall accuracy $\mathrm{Acc} = 0$ and $\mathrm{MCC} = -1$. When $N_{-}^{+} = N^{+}/2$ and $N_{-}^{+} = N^{-}/2$, we have $\mathrm{Acc} = 0.5$ and $\mathrm{MCC} = 0$, meaning no better than random guessing. Therefore, using Eq. (14) has made the meanings of sensitivity, specificity, overall accuracy, and Mathew's correlation coefficient crystal clear and much easier to understand, particularly for the meaning of MCC, as concurred recently by many investigators (see, e.g., Refs. [23–27,62,69,73–77]).

It should be pointed out, however, that the set of equations defined in Eq. (14) is valid only for the single-label systems. For the multi-label systems, whose emergence has become more frequent in system biology [78–80] and system medicine [81], a completely different set of metrics as defined in Ref. [82] is needed.

Cross-validation and target cross-validation

With the metrics for quantitatively measuring the predictor's quality, the next thing is what validation method should be adopted to derive their 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 [83]. 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 Ref. [19] 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., Refs. [41–44,47,80,84,85]). However, to reduce the computational time, in this study we adopted the *K*-fold cross-validation, as done by most investigators with SVM and RF algorithms as the prediction engine.

When conducting the *K*-fold cross-validation for the current predictor iSuc-PseOpt, however, some special consideration is needed. This is because a dataset, after optimization by the KNNC and IHTS treatments, may miss many experimental negative samples and contain some hypothetical positive samples. It would be fine to use such a dataset to train a predictor, but not for validation. This is because the validation should be performed for all of the experimental data in the benchmark dataset but not on the added hypothetical samples or only on the data in the reduced negative subset. To ensure this, a special cross-validation, the so-called target cross-validation, has been introduced here. During the target cross-validation process for the positive samples, only the experiment-confirmed samples are singled out as the targets (or test samples) for validation; however, during the target cross-validation process for the negative samples, even all the excluded experimental data must be taken into account. The detailed procedures of the target *K*-fold cross-validation are as follows (without losing the generality, let us consider $K = 10$).

Step 1.

Before optimizing the original benchmark dataset in Eq. (13), both its positive and negative subsets were randomly divided into 10 parts with about the same size.

Step 2.

One of the 10 sets was singled out as the testing dataset, and the remaining 9 sets were selected as the training dataset.

Step 3.

The training set was optimized using the KNNC and IHTS treatments as described in the "Optimizing imbalanced training datasets" section in Materials and Methods. After such a process, the original imbalanced training dataset would become a balanced one; that is, its positive subset and negative subset would contain the same number of samples.

Step 4.

The aforementioned balanced dataset was used to train the operation engine, followed by applying the iSuc-PseOpt predictor to calculate the prediction scores for the testing dataset, which had been singled out in step 2 before the optimized treatment and, hence, contained the experiment-confirmed samples only.

Step 5.

The scores obtained in this way were substituted into Eq. (14) to calculate Sn, Sp, Acc, and MCC.

Step 6.

Steps 2 to 5 were repeated until all 10 divided sets had been singled out one by one for testing validation.

Step 7.

An average of the metrics scores was taken over the 10-round tests.

It is instructive to emphasize again that it is absolutely reasonable to use the above target cross-validation steps to compare the current predictor with the existing ones. This is because all of the predictors concerned were tested using exactly the same experiment-confirmed samples and all of the added hypothetical samples had been completely excluded from the testing datasets.

Comparison with the existing method

The success rates achieved by the iSuc-PseOpt predictor via the *K*-fold target cross validation on the benchmark dataset (see Supporting Information S1 and Supporting Information S2 in supplementary material) derived from the 896 proteins [28] are given in Table 1. For facilitating comparison, also listed are the corresponding rates achieved by iSuc-PseAAC [17], the only peer counterpart in the area of predicting the lysine succinylation sites in the aforementioned 896 proteins. As we can see from the table, iSuc-PseOpt remarkably outperformed iSuc-PseAAC in all four metrics, indicating that, in comparison with the existing method, the proposed new predictor has better sensitivity, specificity, overall accuracy, and stability.

Graphs are a useful vehicle for studying complicated biological systems because they can provide intuitive insights, as demonstrated by a series of previous studies (see, e.g., Refs. [86–92]). To provide an intuitive comparison, the graph of receiver operating characteristic (ROC) [93,94] was adopted to show the improvement of iSuc-PseOpt over iSuc-PseAAC. The blue graphic line in Fig. 2 is the ROC curve for the iSuc-PseAAC predictor, whereas the red

Table 1
Comparison of proposed predictor with the existing method.

Cross-validation	Method	Sn (%)[a]	Sp (%)[a]	Acc (%)[a]	MCC[a]	AUC[b]
10-fold	iSuc-PseAAC[c]	50.65	89.67	80.02	0.4320	0.7820
	iSuc-PseOpt[d]	68.80	96.48	87.44	0.7084	0.9468
8-fold	iSuc-PseAAC[c]	50.25	89.65	79.91	0.4280	0.7820
	iSuc-PseOpt[d]	69.30	96.56	87.71	0.7137	0.9470
6-fold	iSuc-PseAAC[c]	49.95	89.70	79.87	0.4260	0.7810
	iSuc-PseOpt[d]	69.38	96.86	87.86	0.7193	0.9475

[a] See Eq. (14).
[b] The area under the curve of Fig. 1. The greater the AUC value is, the better the corresponding predictor will be [93,94].
[c] The predictor developed in Ref. [17].
[d] The predictor proposed in this article.

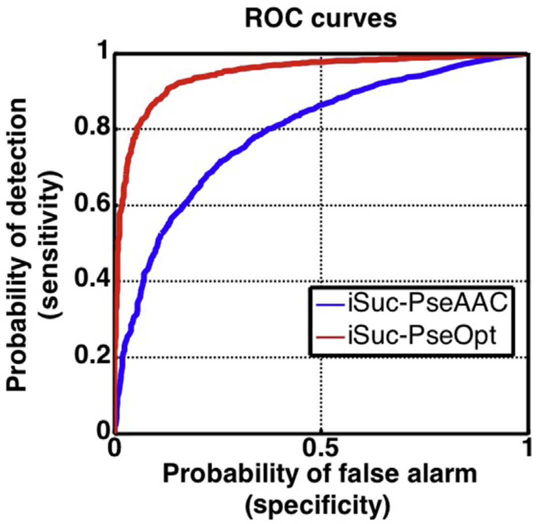


Fig.2. Intuitive graphs of ROC curves [93,94] to show the performance of iSuc-PseAAC [17] and iSuc-PseOpt. See the main text for further explanation. (For interpretation of the references to color in the text description of this figure, the reader is referred to the web version of this article.)

graphic line is that for the proposed predictor iSuc-PseOpt. 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 [93,94]. As we can see from Fig. 2, the area under the red curve is remarkably greater than that under the blue curve, indicating that the proposed predictor is indeed better than iSuc-PseAAC [17]. Therefore, we anticipate that iSuc-PseOpt may become a useful high-throughput tool in this important area or, at the very least, will play a complementary role to the existing method.

Why could the proposed method be so powerful? The reasons are as follows. First, the KNNC and IHTS treatments have been introduced to optimize the training datasets so as to avoid many misprediction events caused by the highly imbalanced training datasets used in Ref. [17]. Second, the coupling effects among the amino acids around the target sites have been taken into account via the conditional probability as done in Refs. [18,58,60,95].

Web-server and user guide

To enhance the value of its practical applications, a web-server for iSuc-PseOpt has been established at http://www.jci-bioinfo.cn/iSuc-PseOpt. Furthermore, to maximize the convenience for the majority of experimental scientists, a step-by-step guide is provided below.

Step 1.

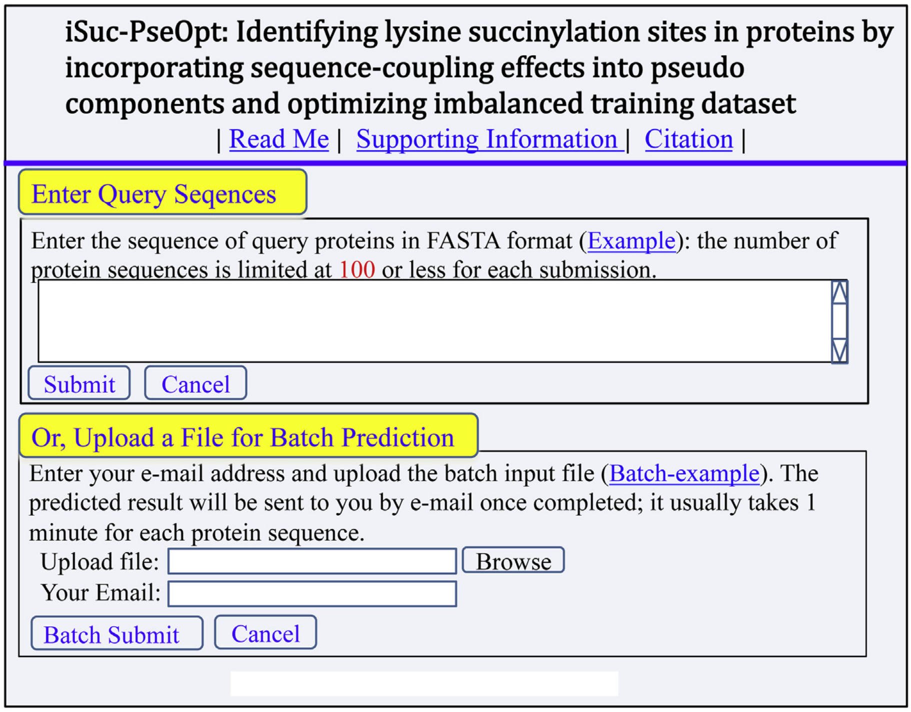
Opening the web-server at http://www.jci-bioinfo.cn/iSuc-PseOpt, you will see the top page of iSuc-PseOpt on your computer screen, as shown in Fig. 3. Click on the “Read Me” button to see a brief introduction about the iSuc-PseOpt predictor.

Step 2.

Either type or copy/paste the query protein sequences into the input box at the center of Fig. 3. The input sequence should be in FASTA format. For examples of sequences in FASTA format, click the “Example” button right above the input box.

Step 3.

Click on the “Submit” button to see the predicted result. For example, if you use the two query protein sequences in the “Example” window as the input, approximately 20 s after your submitting you will see the following on the screen of your computer: (1) Sequence-1 contains 234 amino acid residues, of which 6 are highlighted with red, meaning being of succinylation sites. (2)


iSuc-PseOpt: Identifying lysine succinylation sites in proteins by incorporating sequence-coupling effects into pseudo components and optimizing imbalanced training dataset

| Read Me | Supporting Information | Citation |

Enter Query Seqences

Enter the sequence of query proteins in FASTA format (Example): the number of protein sequences is limited at 100 or less for each submission.

Or, Upload a File for Batch Prediction

Enter your e-mail address and upload the batch input file (Batch-example). The predicted result will be sent to you by e-mail once completed; it usually takes 1 minute for each protein sequence.

Upload file:
Your Email:

Fig.3. Semi-screenshot of the top page for the web-server iSuc-PseOpt at http://www.jci-bioinfo.cn/iSuc-PseOpt.

Sequence-2 contains 417 residues, of which 9 are highlighted with red, meaning being of succinylation sites. All of these predicted results are fully consistent with experimental observations except for residues 3 and 141 in sequence-1 and residue 285 in sequence-2, which are overpredicted.

Step 4.

As shown in the lower panel of Fig. 3, you may also choose the batch prediction by entering your e-mail address and your desired batch input file (in FASTA format) via the "Browse" button. To see the sample of batch input file, click on the "Batch-example" button.

Step 5.

Click on the "Supporting Information" button to download the benchmark dataset used in this study.

Step 6.

Click on the "Citation" button to find the relevant articles that document the detailed development and algorithm of iSuc-PseOpt.

Conclusion

It is a very effective approach to optimize the training dataset via the KNNC and IHTS treatments for enhancing the success rates in predicting the lysine succinylation sites. This is because the training datasets extracted directly from the database [28] are usually extremely skewed and unbalanced, with the negative subset being overwhelmingly larger than the positive subset. In addition, it is important to consider the coupling effects of the amino acids around the potential succinylation sites.

We anticipate that the iSuc-PseOpt web-server presented in this article will become a very useful high-throughput tool for identifying lysine succinylation sites or, at the very least, will become a complementary tool to the existing prediction method in this area.

It has not escaped our notice that the approaches introduced here, such as optimizing the training dataset and incorporating the sequence-coupling effects, can also be used to address many other important problems in computational proteomics.

Acknowledgments

The authors thank the two anonymous reviewers, whose constructive comments were very helpful in strengthening the presentation of this article. This work was partially supported by the National Nature Science Foundation of China (61261027, 31260273, 31560316, and 31560316), the Natural Science Foundation of Jiangxi Province (20122BAB211033, 20122BAB201044, and 20132BAB201053), and 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 data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ab.2015.12.009.

References

[1] E.S. Witze, W.M. Old, K.A. Resing, N.G. Ahn, Mapping protein post-translational modifications with mass spectrometry, Nat. Methods 4 (2007) 798–806.

[2] C.T. Walsh, S. Garneau-Tsodikova, G.J. Gatto, Protein posttranslational modifications: the chemistry of proteome diversifications, Angew. Chem. Int. Ed. 44 (2005) 7342–7372.

[3] Y. Xu, J. Ding, L.Y. Wu, iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition, PLoS One 8 (2013) e55844.

[4] Y. Xu, X.J. Shao, L.Y. W, iSNO-AAPair: incorporating amino acid pairwise coupling into PseAAC for predicting cysteine S-nitrosylation sites in proteins, PeerJ 1 (2013) e171, http://dx.doi.org/10.7717/peerj.171.

[5] C. Jia, X. Lin, Z. Wang, Prediction of protein S-nitrosylation sites based on adapted normal distribution bi-profile Bayes and Chou's pseudo amino acid composition, Int. J. Mol. Sci. 15 (2014) 10410–10423.

[6] W.R. Qiu, X. Xiao, W.Z. Lin, iMethyl-PseAAC: identification of protein methylation sites via a pseudo amino acid composition approach, Biomed. Res. Int. 2014 (2014), http://dx.doi.org/10.1155/2014/947416.

[7] Y. Xu, X. Wen, X.J. Shao, iHyd-PseAAC: predicting hydroxyproline and hydroxylysine in proteins by incorporating dipeptide position-specific propensity into pseudo amino acid composition, Int. J. Mol. Sci. 15 (2014) 7594–7610.

[8] Y. Xu, X. Wen, L.S. Wen, L.Y. Wu, iNitro-Tyr: prediction of nitrotyrosine sites in proteins with general pseudo amino acid composition, PLoS One 9 (2014) e105018.

[9] J. Zhang, X. Zhao, P. Sun, Z. Ma, PSNO: predicting cysteine S-nitrosylation sites by incorporating various sequence-derived features into the general form of Chou's PseAAC, Int. J. Mol. Sci. 15 (2014) 11204–11219.

[10] W.R. Qiu, X. Xiao, W.Z. Lin, iUbiq-Lys: Prediction of lysine ubiquitination sites in proteins by extracting sequence evolution information via a grey system model, J. Biomol. Struct. Dyn. 33 (2015) 1731–1742.

[11] K.C. Chou, Impacts of bioinformatics to medicinal chemistry, Med. Chem. 11 (2015) 218–234.

[12] Y. Xu, Recent progress in predicting posttranslational modification sites in proteins, Curr. Top. Med. Chem. 16 (2016) 591–603.

[13] Z. Zhang, M. Tan, Z. Xie, L. Dai, Y. Chen, Y. Zhao, Identification of lysine succinylation as a new post-translational modification, Nat. Chem. Biol. 7 (2011) 58–63.

[14] J. Park, Y. Chen, D.X. Tishkoff, C. Peng, M. Tan, L. Dai, Z. Xie, Y. Zhang, B.M.M. Zwaans, M.E. Skinner, SIRT5-mediated lysine desuccinylation impacts diverse metabolic pathways, Mol. Cell 50 (2013) 919–930.

[15] Z. Xie, J. Dai, L. Dai, M. Tan, Z. Cheng, Y. Wu, J.D. Boeke, Y. Zhao, Lysine succinylation and lysine malonylation in histones, Mol. Cell. Proteomics 11 (2012) 100–107.

[16] J. Du, Y. Zhou, X. Su, J.J. Yu, S. Khan, H. Jiang, J. Kim, J. Woo, J.H. Kim, B.H. Choi, Sirt5 is a NAD-dependent protein lysine demalonylase and desuccinylase, Science 334 (2011) 806–809.

[17] Y. Xu, Y.-X. Ding, J. Ding, Y.-H. Lei, L.-Y. Wu, N.-Y. Deng, iSuc-PseAAC: Predicting lysine succinylation in proteins by incorporating peptide position-specific propensity, Sci. Rep. 5 (2015) 10184, http://dx.doi.org/10.1038/srep10184.

[18] K.C. Chou, A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins, J. Biol. Chem. 268 (1993) 16938–16948.

[19] K.C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, J. Theor. Biol. 273 (2011) 236–247.

[20] W. Chen, P.M. Feng, H. Lin, iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition, Nucleic Acids Res. 41 (6) (2013) e68.

[21] W.R. Qiu, X. Xiao, iRSpot-TNCPseAAC: identify recombination spots with trinucleotide composition and pseudo amino acid components, Int. J. Mol. Sci. 15 (2014) 1746–1766.

[22] H. Lin, E.Z. Deng, H. Ding, iPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo *k*-tuple nucleotide composition, Nucleic Acids Res. 42 (2014) 12961–12972.

[23] Z. Liu, X. Xiao, W.R. Qiu, iDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition, Anal. Biochem. 474 (2015) 69–77 (also: Data in Brief 4(2015) 87–89).

[24] W. Chen, P. Feng, H. Ding, iRNA-Methyl: Identifying N6-methyladenosine sites using pseudo nucleotide composition, Anal. Biochem. 490 (2015) 26–33 (also: Data in Brief 5(2015) 376–378).

[25] J. Jia, Z. Liu, X. Xiao, 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 (2015) 47–56.

[26] B. Liu, L. Fang, R. Long, iEnhancer-2L: A two-layer predictor for identifying enhancers and their strength by pseudo *k*-tuple nucleotide composition, Bioinformatics (2015), http://dx.doi.org/10.1093/bioinformatics/btv604.

[27] B. Liu, L. Fang, S. Wang, X. Wang, Identification of microRNA precursor with the degenerate *K*-tuple or *K*mer strategy, J. Theor. Biol. 385 (2015) 153–159.

[28] Z. Liu, Y. Wang, T. Gao, Z. Pan, H. Cheng, Q. Yang, Z. Cheng, A. Guo, J. Ren, Y. Xue, CPLM: A database of protein lysine modifications, Nucleic Acids Res. 42 (2014) D531–D536.

[29] UniProt Consortium, The universal protein resource (UniProt) in 2010, Nucleic Acids Res. 38 (2010) D142–D148.

[30] K.C. Chou, Using subsite coupling to predict signal peptides, Protein Eng. 14 (2001) 75–79.

[31] K.C. Chou, H.B. Shen, Recent progress in protein subcellular location prediction, Anal. Biochem. 370 (2007) 1–16.

[32] H.B. Shen, Signal-CF: A subsite-coupled and window-fusing approach for predicting signal peptides, Biochem. Biophys. Res. Commun. 357 (2007) 633–640.

[33] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: Accelerated for clustering the next-generation sequencing data, Bioinformatics 28 (2012) 3150–3152.

[34] K.C. Chou, Prediction of protein cellular attributes using pseudo amino acid composition, Proteins 43 (2001) 246–255 (Erratum: 44 (2001) 60).

[35] K.C. Chou, Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes, Bioinformatics 21 (2005) 10–19.

[36] K.C. Chou, Pseudo amino acid composition and its applications in bioinformatics, proteomics, and system biology, Curr. Proteomics 6 (2009) 262–274.

[37] P. Du, X. Wang, C. Xu, Y. Gao, PseAAC-Builder: A cross-platform stand-alone program for generating various special Chou's pseudo-amino acid compositions, Anal. Biochem. 425 (2012) 117–119.

[38] D.S. Cao, Q.S. Xu, Y.Z. Liang, Propy: A tool to generate various modes of Chou's PseAAC, Bioinformatics 29 (2013) 960–962.

[39] S.X. Lin, J. Lapointe, 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. Eng. 6 (2013) 435–442.

[40] W.Z. Zhong, S.F. Zhou, Molecular science for drug development and biomedicine, Int. J. Mol. Sci. 15 (2014) 20072–20078.

[41] Z.U. Khan, M. Hayat, M.A. Khan, Discrimination of acidic and alkaline enzyme using Chou's pseudo amino acid composition in conjunction with probabilistic neural network model, J. Theor. Biol. 365 (2015) 197–203.

[42] A. Dehzangi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, A. Sattar, Gram-positive and gram-negative protein subcellular localization by incorporating evolutionary-based descriptors into Chou's general PseAAC, J. Theor. Biol. 364 (2015) 284–294.

[43] K.C. Chou, Y.D. Cai, Prediction of membrane protein types by incorporating amphipathic effects, J. Chem. Inf. Model 45 (2005) 407–413.

[44] H.B. Shen, K.C. Chou, Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells, Biopolymers 85 (2007) 233–240.

[45] X. Wang, W. Zhang, Q. Zhang, G.Z. Li, MultiP-SChlo: Multi-label protein subchloroplast localization prediction with Chou's pseudo amino acid composition and a novel multi-label classifier, Bioinformatics 31 (2015) 2639–2645.

[46] S. Ahmad, M. Kabir, M. Hayat, Identification of heat shock protein families and J-protein types by incorporating dipeptide composition into Chou's general PseAAC, Comput. Methods Programs Biomed. 122 (2015) 165–174.

[47] G.L. Fan, X.Y. Zhang, Y.L. Liu, Y. Nang, H. Wang, DSPMP: Discriminating secretory proteins of malaria parasite by hybridizing different descriptors of Chou's pseudo amino acid patterns, J. Comput. Chem. 36 (2015) 2317–2327.

[48] C. Huang, J.Q. Yuan, Simultaneously identify three different attributes of proteins by fusing their three different modes of Chou's pseudo amino acid compositions, Protein Pept. Lett. 22 (2015) 547–556.

[49] M. Mandal, A. Mukhopadhyay, U. Maulik, 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 (2015) 331–344.

[50] V. Sanchez, A.M. Peinado, J.L. Perez-Cordoba, A.M. Gomez, A new signal characterization and signal-based Chou's PseAAC representation of protein sequences, J. Bioinform. Comput. Biol. 13 (2015), http://dx.doi.org/10.1142/S0219720015500249.

[51] P. Du, S. Gu, Y. Jiao, 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 (2014) 3495–3506.

[52] W. Chen, H. Lin, K.C. Chou, Pseudo nucleotide composition or PseKNC: An effective formulation for analyzing genomic sequences, Mol. Biosyst. 11 (2015) 2620–2634.

[53] W. Chen, T.Y. Lei, D.C. Jin, H. Lin, PseKNC: A flexible web-server for generating pseudo *K*-tuple nucleotide composition, Anal. Biochem. 456 (2014) 53–60.

[54] W. Chen, X. Zhang, J. Brooker, H. Lin, PseKNC-General: A cross-platform package for generating various modes of pseudo nucleotide compositions, Bioinformatics 31 (2015) 119–120.

[55] B. Liu, F. Liu, L. Fang, X. Wang, 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 (2015) 1307–1309.

[56] B. Liu, F. Liu, X. Wang, J. Chen, Pse-in-One: A web server for generating various modes of pseudo components of DNA, RNA, and protein sequences, Nucleic Acids Res. 43 (2015) W65–W71.

[57] K.C. Chou, Prediction of human immunodeficiency virus protease cleavage sites in proteins, Anal. Biochem. 233 (1996) 1–14.

[58] K.C. Chou, A sequence-coupled vector-projection model for predicting the specificity of GalNAc-transferase, Protein Sci. 4 (1995) 1365–1383.

[59] A.L. Tomasselli, I.M. Reardon, R.L. Heinrikson, Predicting HIV protease cleavage sites in proteins by a discriminant function method, Proteins 24 (1996) 51–72.

[60] K.C. Chou, Prediction of tight turns and their types in proteins, Anal. Biochem. 286 (2000) 1–16.

[61] Y. Sun, A.K. Wong, M.S. Kamel, Classification of imbalanced data: A review, Int. J. Pattern Recogn. Artif. Intell. 23 (2009) 687–719.

[62] X. Xiao, J.L. Min, W.Z. Lin, Z. Liu, 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 (2015) 2221–2233.

[63] C.T. Zhang, Monte Carlo simulation studies on the prediction of protein folding types from amino acid composition, Biophys. J. 63 (1992) 1523–1529.

[64] C.T. Zhang, An analysis of protein folding type prediction by seed-propagated sampling and jackknife test, J. Protein Chem. 14 (1995) 583–593.

[65] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2011) 321–357.

[66] K.K. Kandaswamy, P.N. Suganthan, S. Sridharan, G. Pugalenthi, AFP-Pred: A random forest approach for predicting antifreeze proteins from sequence-derived properties, J. Theor. Biol. 270 (2011) 56–62.

[67] W.Z. Lin, J.A. Fang, X. Xiao, iDNA-Prot: identification of DNA binding proteins using random forest with grey model, PLoS One 6 (9) (2011) e24756.

[68] G. Pugalenthi, K.K. Kandaswamy, P. Kolatkar, RSARF: prediction of residue solvent accessibility from protein sequence using random forest method, Protein Pept. Lett. 19 (2012) 50–56.

[69] J. Jia, Z. Liu, X. Xiao, B. Liu, 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), http://dx.doi.org/10.1080/07391102.2015.1095116.

[70] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.

[71] J. Chen, H. Liu, J. Yang, Prediction of linear B-cell epitopes using amino acid pair antigenicity scale, Amino Acids 33 (2007) 423–428.

[72] K.C. Chou, Prediction of protein signal sequences and their cleavage sites, Proteins 42 (2001) 136–139.

[73] W. Chen, P.M. Feng, E.Z. Deng, iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition, Anal. Biochem. 462 (2014) 76–83.

[74] W. Chen, P.M. Feng, H. Lin, iSS-PseDNC: identifying splicing sites using pseudo dinucleotide composition, Biomed. Res. Int. 2014 (2014) 623149, http://dx.doi.org/10.1155/2014/623149.

[75] H. Ding, E.Z. Deng, L.F. Yuan, L. Liu, iCTX-Type: A sequence-based predictor for identifying the types of conotoxins in targeting ion channels, Biomed. Res. Int. 2014 (2014) 286419, http://dx.doi.org/10.1155/2014/286419.

[76] B. Liu, L. Fang, F. Liu, X. Wang, Identification of real microRNA precursors with a pseudo structure status composition approach, PLoS One 10 (3) (2015) e0121501.

[77] B. Liu, L. Fang, F. Liu, X. Wang, iMiRNA-PseDPC: MicroRNA precursor identification with a pseudo distance–pair composition approach, J. Biomol. Struct. Dyn. 34 (2016) 223–235.

[78] Z.C. Wu, X. Xiao, iLoc-Hum: using accumulation–label scale to predict subcellular locations of human proteins with both single and multiple sites, Mol. Biosyst. 8 (2012) 629–641.

[79] W.Z. Lin, J.A. Fang, X. Xiao, iLoc-Animal: A multi-label learning classifier for predicting subcellular localization of animal proteins, Mol. Biosyst. 9 (2013) 634–644.

[80] X. Xiao, Z.C. Wu, 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 (2011) 42–51.

[81] X. Xiao, P. Wang, W.Z. Lin, iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types, Anal. Biochem. 436 (2013) 168–177.

[82] K.C. Chou, Some remarks on predicting multi-label attributes in molecular biosystems, Mol. Biosyst. 9 (2013) 1092–1100.

[83] C.T. Zhang, Prediction of protein structural classes, Crit. Rev. Biochem. Mol. Biol. 30 (1995) 275–349.

[84] Y.D. Cai, Prediction and classification of protein subcellular location: sequence-order effect and pseudo amino acid composition, J. Cell. Biochem. 90 (2003) 1250–1260 (Addendum: 91 (2004) 1085).

[85] H.B. Shen, J. Yang, Euk-PLoc: an ensemble classifier for large-scale eukaryotic protein subcellular location prediction, Amino Acids 33 (2007) 57–67.

[86] S. Forsen, Graphical rules for enzyme-catalyzed rate laws, Biochem. J. 187 (1980) 829–835.

[87] G.P. Zhou, M.H. Deng, An extension of Chou's graphic rules for deriving enzyme kinetic equations to systems involving parallel reaction pathways, Biochem. J. 222 (1984) 169–176.

[88] K.C. Chou, Graphic rules in steady and non-steady enzyme kinetics, J. Biol. Chem. 264 (1989) 12074–12079.

[89] I.W. Althaus, J.J. Chou, A.J. Gonzales, F.J. Kezdy, D.L. Romero, P.A. Aristoff, W.G. Tarpley, F. Reusser, Kinetic studies with the nonnucleoside HIV-1 reverse transcriptase inhibitor U-88204E, Biochemistry 32 (1993) 6548–6554.

[90] Z.C. Wu, X. Xiao, 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 (2010) 29–34.

[91] W.Z. Lin, X. Xiao, Wenxiang: A web-server for drawing wenxiang diagrams, Nat. Sci. 3 (2011) 862–865.

[92] G.P. Zhou, The disposition of the LZCC protein residues in wenxiang diagram provides new insights into the protein–protein interaction mechanism, J. Theor. Biol. 284 (2011) 142–148.

[93] J.A. Fawcett, An introduction to ROC analysis, Pattern Recogn. Lett. 27 (2005) 861–874.

[94] J. Davis, M. Goadrich, The relationship between precision–recall and ROC curves, in: Proceedings of the 23rd International Conference on Machine Learning, ACM Press, New York, 2006, pp. 233–240.

[95] K.C. Chou, Prediction of signal peptides using scaled window, Peptides 22 (2001) 1973–1979.