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/

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:

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

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

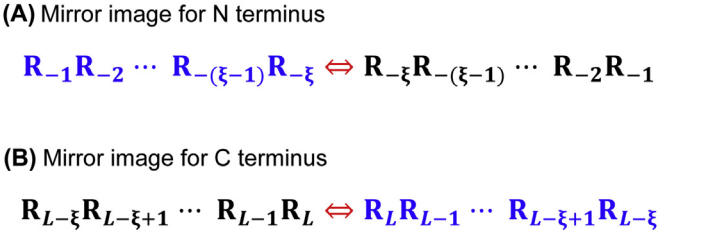

Fig.1. Schematic illustration to show the mirror images of the ξ residues for the N terminus (A) and the C terminus (B). The red symbol ⇔ represents a mirror, and the real peptide segment is colored in black, whereas its mirror image is colored in blue. (For interpretation of the references to color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