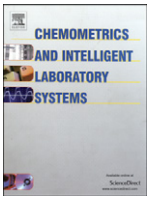
Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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



Combining random forest with multi-amino acid features to identify protein palmitoylation sites

Liang Fu [a], Hua-Lin Xie [a,*], Xiang-Rong Xu [b], Hua-Juan Yang [b], Xi-Du Nie [c]

[a] School of Chemistry and Chemical Engineering, Yangtze Normal University, Fuling 408100, PR China
[b] The First People's Hospital of Changde City, Changde 415003, PR China
[c] College of Material and Chemical Engineering, Hunan Institute of Technology, Hengyang 421002, PR China

ARTICLE INFO

Article history:
Received 18 February 2014
Received in revised form 6 April 2014
Accepted 11 April 2014
Available online 24 April 2014

Keywords:
Palmitoylation sites
Random forest (RF)
Multi amino acid features

ABSTRACT

As one of the most important and ubiquitous post-translational lipid modifications, protein palmitoylation plays significant roles in a variety of biological processes, including signaling, neuronal transmission, and membrane trafficking. Protein palmitoylation is a highly dynamic process, which regulates various protein functions. The dynamic nature of palmitoylation makes it very difficult to identify such kind modification by experimental assay methods. Therefore, using computational approaches to identify palmitoylation sites is of highly important. In this study, a new method was proposed to predict palmitoylation sites based on multi-amino acid properties and random forest (RF) algorithm. The prediction accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC) and area under the curve values (AUC) for current method were 91.85%, 88.89%, 94.67%, 0.8377, and 0.9595, respectively. These results indicated that the current method was a powerful and effective tool for identifying palmitoylation sites, which would be a complement to protein palmitoylation research. Furthermore, a free online service was established in http://sysbio.yznu.cn/Research/RandomForcast.aspx.

1. Introduction

Protein palmitoylation, also known as S-acylation, is one of the most ubiquitous post-translational modifications (PTMs), which reversibly attaching a 16-carbon saturated fatty acid as lipid palmitate (C16:0) to cysteine residues in protein substrates through thioester linkage [1–3]. In addition protein palmitoylation is a reversible lipid modification that plays important roles in cell signaling. Palmitoylation modification can increase the hydrophobicity of proteins to promote protein-membrane association [4–6]. Furthermore, palmitoylation modifies numerous proteins to control protein–protein interactions [7], intracellular trafficking [8], lipid raft targeting [9,10], and proteins' activities [11–13]. Moreover, palmitoylation has been implicated in a variety of biological and physiological processes, including signal transduction [12,13], neuronal development [3], and apoptosis [14]. It is very obvious that revealing the exact positions of palmitoylation sites in a protein sequence can elucidate many important biological processes such as protein folding, subcellular localization, protein transportation, functions, and provide useful clues for drug design and other biotechnology applications.

To date, several conventional experimental techniques (such as mass spectrometry) for understanding the mechanisms of protein palmitoylation and identifying the exact positions of palmitoylation have been employed [15–17]. Although the protein palmitoylation sites can be determined by these conventional experimental techniques, the features of substrate specificity for palmitoylation are still unclear, and most previous studies have proposed that there is no common and canonical consensus motif for palmitoylation [10]. These drawbacks make experimental methods to determinate that palmitoylation sites in proteins are still an expensive and laborious process; thus, it is highly desirable to develop a fast, automated and effective computational method to identify protein palmitoylation sites, in contrast with time-consuming and expensive experimental methods.

Some computational methods have been developed to predict protein palmitoylation sites. Zhou et al., first employed a clustering and scoring strategy (CSS) to build a model to predict palmitoylation sites in 2006 [18,19]. Xue et al., applied a Naive Bayes method to predict palmitoylation sites in 2006 [20]. Wang et al., using the composition of k-spaced amino acid pairs as the encoding scheme, proposed a predictor called CKSAAP-Palm to identify the potential palmitoylation sites [21]. Recently, Hu et al., proposed a predictor named IFS-Palm for predicting palmitoylation sites based on the amino acid sequence features. In these methods, the greatest total accuracy was 90.65% based on IFS-Palm predictor [22]. Hence, the prediction accuracies were far from ideal, it was crucially important to develop some reliable computational methods for identifying protein palmitoylation sites.

Abbreviations: (RF), Random forests; MCC, Matthews correlation coefficient; AUC, Area under the curve; ROC, Receiver operating characteristic.

* Corresponding author at: Department of Chemistry and Chemical Engineering, Yangtze Normal University, Fuling 408100, PR China. Tel./fax: +86 23 72792170.
E-mail address: hualinxie@163.com (H.-L. Xie).

http://dx.doi.org/10.1016/j.chemolab.2014.04.009
0169-7439/© 2014 Elsevier B.V. All rights reserved.

In this study, a novel approach was developed to identify palmitoylation sites by coupling multi amino acid properties with random forest (RF). Seven amino acid physicochemical properties and structural characteristic information were fused to form the feature variables for classification. The influences of sequence extracting window sizes, ratios between the number of positive sites and the number of negative sites were optimized. The proposed algorithm obtained satisfied predictive results in identifying palmitoylation sites.

2. Materials and methods

2.1. Datasets construction and preprocessing

The dataset was constructed by the following steps: Firstly, 172 proteins covering experimental palmitoylation cysteine were obtained by searching the keywords “palmitoylated cysteine” from UniProtKB/Swiss-Prot, which were not annotated as “by similarity”, “potential” or “probable”; Secondly, among these proteins, 151 proteins were used by Hu et al., [22]. And the remaining 21 proteins were new uploaded in UniProtKB/Swiss-Prot, these proteins were used as independent dataset in this study.

Then, we defined the cysteine (C) residues as central residues, and a sequence fragment with $2n + 1$ amino acids was constructed by taking n upstream residues and n downstream residues from the cysteine residue in each protein sequence. Here, the cysteine (C) palmitoylated modification sites that verified by the experimental methods were defined as positive data, while those did not verified by experimental methods were defined as negative data; lastly, training dataset contained 144 experimental palmitoylation sites (positive data) and 1268 non-experimental palmitoylation sites (negative data). Moreover, to verify our method, the remaining 21 protein sequences were selected as independent testing dataset, which contained 33 experimental palmitoylation sites and 245 non-experimental palmitoylation sites.

2.2. Feature extraction and coding

In this study, the protein amino acid properties were chosen for residue representation. A protein sequence can be represented as a series of amino acids by their single-character codes A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W and Y, formulated as

$$R_1R_2R_3R_4R_5R_6R_7R_8\ldots R_L. \tag{1}$$

Suppose $H(R_1)$ is the hydrophobic value of the 1st residue R_1, $H(R_2)$ that of the 2nd residue R_2, and so forth. In terms of these hydrophobic values the protein sequence of Eq. (1) can be converted to a digit signal. If we have chosen 10 as the sliding window size, each sequence fraction would contained 21 residues, and then could be transformed into 21 digit features. After that the mean values for all the sequence fractions were calculated. Therefore, the dimension of feature vector for each sequence was 22.

Each amino acid in peptides was encoded by 7 properties. These properties included hydrophobicity, hydrophilicity, volume of side chains, polarity, polarizability, solvent accessible, and net charge index of side chains. The amino acid properties were available at AAindex [23]. Therefore, the total number of input variables for a sequence was 154 ($22 * 7$), when n equals to 10. In the current study, the window length n was optimized from 4 to 12.

2.3. Random forest

Random forest (RF) is a classifier consisting of an ensemble of classification and regression tree-structured classifiers [24]. All trees in the forest are unpruned. RF takes advantages of two powerful machine learning techniques: bagging and random feature selection. In bagging, each tree is trained on bootstrap samples of the training data, and predictions are made by the majority vote of the trees. RF is a further development of bagging, which instead of using all features, it randomly selects a subset of features to split at each node when growing a tree. In order to assess the prediction performance of the random forest algorithm, RF performs a type of a cross-validation in parallel with the training step by using the so called OOB samples. The OOB samples were used to get an unbiased estimate of the classification error as trees were added to the forest. It was also used to get estimates of variable importance. Specifically, in the process of training, each tree is grown using some particular bootstrap samples. Since bootstrapping is a sampling method with replacement from the training data, some of the data will not be chosen to establish the training dataset or can be called “left out”, while some samples will be chosen to train the model many times. The ‘left out’ data is also called the “OOB sample”. On average, each tree is grown using about 2/3 of the training data, leaving about 1/3 samples as OOB sample. Since OOB data have not been used in the tree construction, it can be used to estimate the prediction performance. The RF algorithm implemented in the R-package randomForest was used in this study [25]. The algorithm (for both classification and regression) can be stated as follows:

1. Draw n_{tree} bootstrap samples from the original data, n_{tree} is the number of ensemble trees, in the current study n_{tree} is 500;
2. For all bootstrap samples, grow an un-pruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all variables, randomly select m_{try} variables and choose the best split among those variables (bagging can be thought as the special case of random forest when $m_{try} = p$, the number of variables). In general, m_{try} is simply a number (positive integer) between 1 and p [24]. In the current study, the value of m_{try} is 15.
3. Predict new data by aggregating the predictions of the $\mathrm{n}_{\mathrm{tree}}$ (i.e., majority votes for classification, average for regression).

Variable importance: RF, as an ensemble of trees, inherits the ability to estimate feature importance. A measure of how each feature contributes to the prediction performance of RF can be calculated in the course of the training. The important scores can be used to identify biomarkers or as a filter to remove non-informative variables. The frequently used type of RF to measure feature importance is the mean decrease in classification based on permutation. For each tree, the classification accuracy of the OOB samples is determined both with and without random permutation of the values to each variable, one by one. The prediction accuracy of after permutation is subtracted from the prediction accuracy before permutation and averaged over all trees in the forest to give the permutation importance value. In the current research, the mean decrease in classification accuracy was accepted to measure variable importance. The importance of each variable (j) can be calculated as Eq. (2)

$$\text{Importance of j} = \text{Accuracy}_{\text{j normal}} - \text{Accuracy}_{\text{j permuted}}. \tag{2}$$

2.4. Model training and evaluation

The performance of classifier classification has been evaluated by the following measures [26]:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{4}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{5}$$