Analytical Biochemistry 550 (2018) 1–7

Contents lists available at ScienceDirect

Analytical Biochemistry

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

Prediction of lysine glutarylation sites by maximum relevance minimum redundancy feature selection

Zhe Ju[a,*], Jian-Jun He[b]

[a] College of Science, Shenyang Aerospace University, 110136, People's Republic of China
[b] College of Information and Communication Engineering, Dalian Minzu University, 116600, People's Republic of China

ARTICLE INFO

Keywords:
Post-translational modification
Glutarylation
Feature extraction
Incremental feature selection
Support vector machine

ABSTRACT

Lysine glutarylation is new type of protein acylation modification in both prokaryotes and eukaryotes. To better understand the molecular mechanism of glutarylation, it is important to identify glutarylated substrates and their corresponding glutarylation sites accurately. In this study, a novel bioinformatics tool named GlutPred is developed to predict glutarylation sites by using multiple feature extraction and maximum relevance minimum redundancy feature selection. On the one hand, amino acid factors, binary encoding, and the composition of k-spaced amino acid pairs features are incorporated to encode glutarylation sites. And the maximum relevance minimum redundancy method and the incremental feature selection algorithm are adopted to remove the redundant features. On the other hand, a biased support vector machine algorithm is used to handle the imbalanced problem in glutarylation sites training dataset. As illustrated by 10-fold cross-validation, the performance of GlutPred achieves a satisfactory performance with a Sensitivity of 64.80%, a Specificity of 76.60%, an Accuracy of 74.90% and a Matthew's correlation coefficient of 0.3194. Feature analysis shows that some k-spaced amino acid pair features play the most important roles in the prediction of glutarylation sites. The conclusions derived from this study might provide some clues for understanding the molecular mechanisms of glutarylation.

Introduction

Some of the crucial protein post-translational modifications (PTMs), such as acetylation, butyrylation, malonylation, succinylation, 2-hydroxyisobutyrylation and crotonylation, can occur at the ε-amino groups of specific lysine residues [1]. These protein lysine acylations play an important role in regulating various biological functions of proteins [2–6]. Recently, a new type of lysine acylation, glutarylation, has been identified in both eukaryotic and prokaryotic cells. Lysine glutarylation is a dynamic and evolutionarily conserved PTM, which impacts metabolic processes and other mitochondrial functions [7]. Similar to acetylation and succinylation, lysine glutarylation is highly enriched in metabolic processes, such as cellular respiration, fatty acid metabolism and amino acid metabolism [6,7]. Tan et al. [7] also found that glutarylation of carbamoyl phosphate synthase 1 (CPS1) inhibits its activity but can be reversed by SIRT5. Moreover, previous studies have shown that mitochondrial protein glutarylation was markedly elevated in a mouse model of glutaric acidemia type I, suggesting that glutarylation was closely related to this disease. Therefore, deciphering the biological function and the underlying molecular mechanisms of lysine glutarylation might be beneficial in the treatment of glutaric acidemia type I. Although some functions of lysine glutarylation have begun to be revealed, the role of glutarylation in cellular regulation remains poorly characterized relative to succinylation and acetylation.

To better understand the molecular mechanisms of glutarylation, a fundamental step is the identification of glutarylation sites. Although several large-scale proteomics methods such as mass spectrometry [7,8] have been applied to detect glutarylation sites, these experimental methods are time-consuming and labor-intensive. The majority of lysine glutarylated substrates and glutarylation sites still remain largely unknown. Therefore, it is urgent to develop computational methods to identify the potential glutarylated proteins and the corresponding glutarylation sites. Similar computational methods have been developed for predicting various lysine acylation sites, such as acetylation sites [9–11], succinylation sites [12,13], propionylation sites [14] and crotonylation sites [15,16].

Abbreviations: PTM, post-translational modification; AAF, amino acid factors; BE, binary encoding; CKSAAP, composition of *k*-spaced amino acid pairs; mRMR, maximum relevance minimum redundancy; IFS, incremental feature selection; SVM, support vector machine; Sn, sensitivity; Sp, specificity; ACC, accuracy; MCC, Matthew's correlation coefficient; AUC, area under receiver operating characteristic curve

* Corresponding author. #37 Daoyi South Street, Shenyang, People's Republic of China.
E-mail addresses: juzhe1120@hotmail.com (Z. Ju), jianjunhe@live.com (J.-J. He).

https://doi.org/10.1016/j.ab.2018.04.005
Received 17 January 2018; Received in revised form 5 April 2018; Accepted 6 April 2018
Available online 08 April 2018
0003-2697/ © 2018 Elsevier Inc. All rights reserved.

To develop an accurate predictor for identifying glutarylation sites, it is important to find an effective encoding method for representing the sequence context around the glutarylation sites. In addition, the number of glutarylation sites is much smaller than non-glutarylation sites, which meant the training dataset is highly unbalanced. To address these two issues, a novel predictor called GlutPred was proposed to predict glutarylation sites using multiple feature extraction and a biased support vector machine algorithm. Amino acid factors (AAF), binary encoding (BE), and the composition of k-spaced amino acid pairs (CKSAAP) were combined to encode glutarylation site peptides. The maximum relevance minimum redundancy (mRMR) method [17] and the incremental feature selection (IFS) algorithm were used to remove the redundant features and improve the prediction performance of the predictor. To overcome the disadvantage of highly unbalanced training dataset, a biased support vector machine (SVM) [18] was used to construct a robust classifier for the prediction of glutarylation sites. The biased SVM was able to effectively reduce the imbalance of training set by assigning a large penalty factor for glutarylation sites and a small penalty factor for non-glutarylation sites. The experimental results showed that the area under receiver operating characteristic curve (AUC) of GlutPred was 0.7806 on the training set evaluated by 10-fold cross-validation and 0.7663 on the independent test set. Our analysis showed that the occurrence frequency of k-spaced amino acid pairs surrounding the lysine-centered sites was important for the prediction of glutarylation sites. The conclusions derived from this study might provide some useful information for studying the mechanisms of glutarylation and the related experimental validations.

Materials and methods

Dataset

All training data were retrieved from the PLMD database [19], which included 211 glutarylated proteins. To reduce redundancy, the proteins whose similarities were greater than 40% were removed by the CD-HIT program [20,21]. We then obtained the non-redundant dataset consisting of 187 glutarylated proteins with 646 glutarylation sites. After removing the redundancies, we randomly selected 20 proteins from these proteins to construct the independent testing dataset and the remaining 167 proteins were used to construct the training dataset. Finally, the training dataset consisted of 167 proteins with 590 glutarylation sites and 3498 non-glutarylation lysine sites. The independent testing set consisted of 20 proteins with 56 glutarylation sites and 428 non-glutarylation lysine sites.

The sliding window method was used to represent every lysine residue K in the training dataset because glutarylation only occurs in lysine residues. In this study, the window size was selected as 35. The selection methodology to obtain the best window size will be discussed in Section 3.1. Then, every sample of the training and testing datasets was represented as a peptide segment of length 35 with 17 residues upstream and 17 residues downstream of the lysine residue K. To ensure the unified length of each training peptide, a virtual residue 'X' was used to fill the corresponding positions where there were no sufficient residues. The training set and independent testing set were provided in Supplementary material S1.

Feature construction

After a preliminary test, we found that the amino acid factors (AAF), binary encoding (BE) and the composition of k-spaced amino acid pairs (CKSAAP) encoding were more suitable for representing the protein sequence around the glutarylation sites than other encoding methods, such as pseudo amino acid composition [22] or split amino acid composition [23] (see Supplementary material S2). Therefore, these three features were incorporated to encode the glutarylation sites.

Amino acid factors

The AAIndex database contains various physicochemical properties of amino acids. Five multidimensional patterns of attributes were generated from the AAIndex database using multivariate statistical analyses: the attributes reflected the polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge, respectively [24]. These five transformed attributes are the so-called amino acid factors (AAF). As an effective feature encoding technique, AAF has been successfully applied to many biology problems [25–27]. Here, the AAF was used to encode every training peptide as 35*5 = 175-dimensional feature vectors.

Binary encoding

Binary encoding (BE) can reflect the amino acid position and composition information of the residues surrounding the lysine-centered site. For the 21 amino acids (20 native amino acids and 1 virtual amino acid 'X') which were ordered as 'ACDEFGHIKLMNPQRSTVWYX', each amino acid residue in the training peptide was encoded by a 21-dimensional binary vector. For example, amino acid 'A' is encoded as (100000000000000000000), 'C' is encoded as (010000000000000000000) and 'X' is encoded as (000000000000000000001). Therefore, every training peptide was represented as a 35*21 = 735-dimensional feature vector according to the BE approach.

Composition of k-spaced amino acid pairs

The CKSAAP encoding calculates the frequencies of occurrence of the *k*-spaced amino acid pairs in a given peptide, which can reflect the short linear motif information in a sequence fragment [28]. The *k*-spaced amino acid pair means two amino acids separated by any *k* amino acids. For example, the CKSAAP encoding of a peptide for $k = 1$ is a 441-dimensional feature vector defined as:

$$(N_{AxA}/N_{Total},\ N_{AxC}/N_{Total}, \ldots, N_{XxX}/N_{Total})_{441} \tag{1}$$

where '*x*' represents any one of the 21 amino acids, and N_{Total} represents the total number of 1-spaced amino acid pairs. Here, CKSAAP with k = 0, 1, 2, 3 and 4 was utilized to encode the training peptides as 2205-dimensional feature vectors.

The feature space

To capture more useful information from the protein sequences, the AAF, BE and CKSAAP encoding methods were combined to encode the training peptides in this study. Overall, each training peptide was represented by 35*5 + 35*21 + 2205 = 3115 features.

mRMR method

In this study, the maximum relevance and minimum redundancy (mRMR) feature selection method [17] was used to remove the irrelevant and redundant features. In the mRMR method, each feature is ranked based not only on its relevance to the target variable, but also on its redundancy in the feature set. Mutual information (MI) is applied to measure both redundancy and relevance in the mRMR method. MI is defined as follows:

$$I(X, Y) = \iint P(x, y) \log \frac{p(x, y)}{p(x)p(y)} \tag{2}$$

where X, Y are vectors, $p(x, y)$ is the joint probabilistic density, $p(x)$ and $p(y)$ are the marginal probabilistic densities. Given M data points drawn from the joint probability distribution (x_i, y_i), $i = 1, 2, \ldots, M$, the joint and marginal densities can be estimated by the Gaussian kernel estimator as follows:

$$p\left(x, y\right) = \frac{1}{M} \sum \frac{1}{2\pi h^2} e^{-\frac{1}{2h^2}\left((x-x_i)^2+(y-y_i)^2\right)} \tag{3}$$

Download English Version:

https://daneshyari.com/en/article/7556813

Download Persian Version:

https://daneshyari.com/article/7556813

Daneshyari.com