Neurocomputing 128 (2014) 267–272

Contents lists available at ScienceDirect

Neurocomputing

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

Predicting pupylation sites in prokaryotic proteins using pseudo-amino acid composition and extreme learning machine

Yong-Xian Fan [a,b], Hong-Bin Shen [a,*]

[a] Department of Automation, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China
[b] School of Computer Science and Engineering, Guilin University of Electronic Technology, Guilin 541004, China

ARTICLE INFO

Article history:
Received 27 August 2012
Received in revised form
21 November 2012
Accepted 25 November 2012
Available online 24 October 2013

Keywords:
Pupylated protein
Pupylation sites
Pseudo-amino acid composition
Extreme learning machine
Bioinformatics
PupS

ABSTRACT

Pupylation is one of the most important post-translational modifications of prokaryotic proteins playing a key role in regulating a wild range of biological processes. Prokaryotic ubiquitin-like protein can attach to specific lysine residues of substrate proteins by forming isopeptide bonds for the selective degradation of proteins in *Mycobacterium tuberculosis*. In order to comprehensively understand these pupylation-related biological processes, identification of pupylation sites in the substrate protein sequence is the first step. The traditional wet-lab experimental approaches are both laborious and time-consuming. To timely and effectively discover pupylation sites when facing with the avalanche of new protein sequences emerging during the post-genomic Era, a novel computational predictor called PupS (pupylation site predictor) is proposed. PupS is constructed on the pseudo-amino acid composition and trained with extreme learning machine. The jackknife cross-validation results on the training dataset show that the area under an ROC Curve (AUC) value is 0.6483 by PupS, and an AUC of 0.6779 is obtained on the independent set. Our results also demonstrate that ELM is complementary to other algorithms and that constructing an ensemble classifier will generate better results. PupS software package is available at http://www.csbio.sjtu.edu.cn/bioinf/PupS/.

1. Introduction

Prokaryotic ubiquitin-like protein (Pup) is the first identified post-translational small protein modifier in prokaryotes. Pup can attach to specific lysine (K) residues of substrate proteins by forming isopeptide bonds for the selective degradation of proteins in *Mycobacterium tuberculosis* (Mtb), which is similar to ubiquitin (Ub) mediated proteolysis in eukaryotes [1]. Pupylation in prokaryotes plays critical roles in numerous regulatory functions such as protein degradation and signal transduction [2]. Although pupylation and ubiquitylation in eukaryotes are similar in functional roles, the enzymology of pupylation and ubiquitylation is different. In contrast with the three-step reaction of eukaryotic ubiquitylation with E1, E2, and E3 ligases [3], prokaryotic pupylation requires only two steps, where two enzymes thus are involved in the pupylation. Firstly, the C-terminal glutamine of Pup is deamidated to glutamic acid by deamidase of Pup (Dop) [4]. Secondly, the deamidated Pup is attached to specific lysine (K) residues of pupylated substrate proteins by proteasome accessory factor A (PafA) [5].

The identification of pupylated substrate proteins along with pupylation sites can provide valuable insights into the substrate specificity and functions of pupylation. With the application of the large-scale proteomics technology, such as tandem mass spectrometry, the number of identified pupylated proteins continues growing [6–9]. The proteome-wide analyses already revealed that there are at least several hundreds of potential pupylated proteins in the model organism *M. smegmatis*, in which the selective degradation of pupylation-mediated protein was proposed to be highly dynamic and dependent on the culture conditions [8,9]. Although much progress has been achieved in this regard, the number of experimentally verified pupylated proteins and pupylation sites are still relatively small as the traditional experimental approaches are laborious and time-consuming. Alternative bioinformatics method is highly desired that can quickly predict the potential true pupylation sites throughout the entire protein sequences, which will provide timely and helpful information for further experimental verifications.

To the best of our knowledge, only one predictor GPS-PUP [10] is available specifically for predicting pupylation sites currently. GPS-PUP is mainly based on the no interval alignment scoring method and a training set including 127 experimentally identified pupylation sites in 109 prokaryotic pupylated proteins. In this article, to speed up the progress in this field, we developed a new

* Corresponding author. Tel.: +86 21 34205320; fax: +86 21 34204022.
E-mail address: hbshen@sjtu.edu.cn (H.-B. Shen).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.11.058

sequenced based computational method called PupS for predicting pupylation sites. We used an advanced machine learning method, extreme learning machine (ELM) [11], to establish the prediction model. ELM has also been found powerful in dealing with complex biological data. For example, ELM was applied for the protein sequence and microarray gene expression cancer diagnosis classifications [18,19]. Recently, ELM was also incorporated with the N-to-1 neural networks to detect the transmembrane β-barrel (TMBB) proteins [20]. Pseudo-amino acid composition (PseAAC) was used to encode the amino acid sequences into vectors, which can not only represent amino acid composition information, but also can capture the correlations between amino acid residues [12] and has been widely used for protein attribute prediction [13–16] (detailed review in [17]).

Several features distinguish current study from other relative works. Firstly, a larger new non-redundant dataset is constructed in this study that is important for training a solid statistic machine learning algorithm. Secondly, the pseudo-amino acid composition (PseAAC) feature applied in this paper is able to capture the information of sequence-order correlations. Thirdly, the applied prediction model of ELM is much more faster than other algorithm, such as SVM, which enables current model capable of dealing with large-scale datasets.

2. Materials and methods

2.1. Datasets

A freely accessible database named PupDB [21] has been established which integrates information of both pupylated proteins and pupylation sites. In this study, for the purpose of training a solid predictor, we constructed a new non-redundant dataset according to the following steps:

(1) All pupylated proteins and pupylation sites with experimental evidences were extracted from the latest PupDB database version [21]. This initial dataset contains 182 pupylated proteins including 215 pupylation sites, where 2 of these pupylation sites that are not lysine (K) amino acid were not considered.
(2) The remaining 180 pupylated proteins including 213 pupylation sites were further inputted to the CD-HIT [22] method to remove redundancy with pairwise sequence identity cut-off 30%. We then obtained the final non-redundant dataset consisting of 145 pupylated protein sequences including 174 pupylation sites.
(3) Ten pupylated proteins were randomly selected from the original non-redundant dataset to construct an independent test set, and then the remaining 135 pupylated proteins act as the training set.
(4) We then used a peptide centered with the residue lysine (K) to encode the target that needed to be predicted as positive (pupylation site) or negative (non-pupylation site). By testing several different sizes and according to the two sample logo (see Fig. 1), a window size of 25 residues was adopted in this study. Each residue lysine (K) can then be represented with a peptide segment consisting of 12 residues upstream and 12 residues downstream of the lysine (K). Since the number of pupylation sites and putative non-pupylation sites were imbalanced (less than 1:12), a relatively balanced procedure is applied, and three times negative samples were selected to match the positive ones in the training set. But in the independent test set, we retained all the positive and negative samples in order to simulate the real situation.
(5) Finally, the 135 training pupylated proteins including 158 positive peptide segments (pupylation sites) and 1928 negative peptides (putative non-pupylation sites) constitute the training set. In the meantime, 10 testing pupylated proteins including 16 positive samples (pupylation sites) and 279 negative samples (putative non-pupylation sites) constitute the testing set.

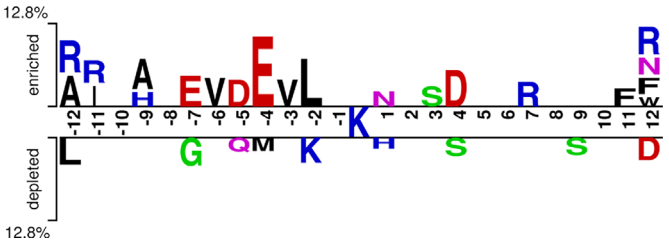

Fig. 1. The two sample logo of the position-specific residue composition in the vicinity of the 174 pupylation sites and 2207 non-pupylation sites in a window of 25 residues, illustrating compositional differences between pupylation sites and putative non-pupylation sites. Only amino acid residues significantly enriched and depleted ($P < 0.05$; t-test) are shown.

2.2. Methods

2.2.1. Sequence analysis of the position-specific attributes

To determine whether pupylation and putative non-pupylation sites have distinct sequence properties, we calculated statistically significant differences in the distribution of amino acid residues in the vicinity of 174 (158+16) pupylation peptides and 2207 (1928+279) putative non-pupylation segments. The two sample logo of the position-specific residue composition in the vicinity of the pupylation sites and putative non-pupylation sites in a window with length 25 was created as shown in Fig. 1 [23]. Polar amino acids (G, S, T, Y, C) show as green, acidic (Q, N) purple, basic (K, R, H) blue, positive (D, E) red, and hydrophobic (A, V, L, I, P, W, F, M) amino acids as black in the two sample logo.

The two sample logo showed compositional differences between pupylation sites and putative non-pupylation sites. The most distinct feature of pupylation sites is the enrichment of positive amino acid (E) at positions −4 and −7, positive amino acid (D) at positions −5 and 4, hydrophobic amino acids (A, V, and L) at positions −12, −9, −6, −3 and −2, and positively charged amino acid (R) at positions −12, −11, 7 and 12. On the contrary, the depletion of lysine (K) at position −2, polar amino acids (G and S) at positions −7, 4 and 9, positive amino acid (D) at position 12, and hydrophobic amino acid (L) at position -12 are observed around pupylation sites. These statistics show that there are strong correlations between the residues around the pupylation sites, requiring proper encoding methods.

2.2.2. Encoding protein sequence with pseudo-amino acid composition

According to the definition of amino acid composition (AAC), AAC of a protein sequence can be represented by 20 discrete numbers with each denoting the occurrence frequency of one of the 20 native amino acids in the protein. But, if using the 20 dimensional AAC to represent a protein sequence, all its sequence order information would be lost. For instance, AAC cannot catch the strong residue correlations around the pupylation sites as shown in Fig. 1. In view of this, instead of the conventional AAC, we adopt PseAAC to represent a protein sample in a $(20+\lambda)$ dimensional vector. The first 20 elements in PseAAC reflect the traditional global amino acid composition in the sequence and the later λ elements represent the local correlations among residues [12]. In our application, each sample is presented with a peptide