Computational Biology and Chemistry 71 (2017) 98–103

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Predicting lysine glycation sites using bi-profile bayes feature extraction

Zhe Ju*, Juhe Sun, Yanjie Li, Li Wang

College of Science, Shenyang Aerospace University, 110136, People's Republic of China

ARTICLE INFO

Article history:
Received 31 May 2017
Received in revised form 14 September 2017
Accepted 7 October 2017
Available online 12 October 2017

Keywords:
Post-translational modification
Glycation
Bi-profile bayes
Support vector machine

ABSTRACT

Glycation is a nonenzymatic post-translational modification which has been found to be involved in various biological processes and closely associated with many metabolic diseases. The accurate identification of glycation sites is important to understand the underlying molecular mechanisms of glycation. As the traditional experimental methods are often labor-intensive and time-consuming, it is desired to develop computational methods to predict glycation sites. In this study, a novel predictor named BPB_GlySite is proposed to predict lysine glycation sites by using bi-profile bayes feature extraction and support vector machine algorithm. As illustrated by 10-fold cross-validation, BPB_GlySite achieves a satisfactory performance with a Sensitivity of 63.68%, a Specificity of 72.60%, an Accuracy of 69.63% and a Matthew's correlation coefficient of 0.3499. Experimental results also indicate that BPB_GlySite significantly outperforms three existing glycation sites predictors: NetGlycate, PreGly and Gly-PseAAC. Therefore, BPB_GlySite can be a useful bioinformatics tool for the prediction of glycation sites. A user-friendly web-server for BPB_GlySite is established at 123.206.31.171/BPB_GlySite/.

1. Introduction

As one of the common and important post-translational modifications (PTMs), lysine glycation can potentially affect various biological processes, such as conformation, efficacy and immunogenicity (Miller et al., 2011). Glycation is the process of the typically covalent bonding of a sugar molecule (such as fructose or glucose) to a protein or lipid molecule. In contrast to glycosylation requiring the controlling action of enzymes, glycation is a nonenzymatic modification process. The unstable Schiff base firstly rearranges to form a more stable Amadori product. Subsequently, the Amadori product can react further to form the advanced glycation end products (AGEs) which are irreversible cross-linked products (Cho et al., 2007; Lapolla et al., 2001). Lysine glycation can occur in both intracellular and extracellular proteins (Garlick and Mazer, 1983; Shilton and Walton, 1991). In general, intracellular glycation is more complex than extracellular glycation due to the multiple potential sources in cytoplasm can also react to form AGEs. Kinetic analysis of glycation reaction has shown that the amount of glycation at steady state is proportional to the glucose concentration, to protein half-life and to the rate of glycation (Schleicher and Wieland, 1986). Previous studies have demonstrated that the glycation is closely related to the occurrence and development of various human diseases, such as diabetes and its vascular complications (Ahmed et al., 2005), renal failure (Agalou et al., 2005), Parkinson's disease and Alzheimer's disease (Ling et al., 1998). Therefore, deciphering the underlying molecular mechanisms and the biological function of glycation might be beneficial in the treatment of the above-mentioned diseases. However, the molecular mechanism of glycation remains largely unknown.

To better understand the molecular mechanisms of glycation, the fundamental step is to identify glycated substrates and their corresponding glycation sites with high accuracy. Several large-scale proteomics methods such as mass spectrometry (Zhang et al., 2009; Thornalley and Naila, 2014) have been applied to detect glycation sites. However, as we know, the experimental approaches are often labor-intensive and time-consuming. The computational studies of protein glycation are gaining increasing attention. Up to now, several computational methods have been developed to predict glycation sites from protein sequences. With the ensemble artificial neural network algorithm, Johansen et al. (2006) proposed the first predictor NetGlycate for the prediction of lysine glycation sites. Liu et al. (2015) proposed a computational method PreGly for predicting glycation sites, which used amino acid factors, amino acid occurrence frequency and the composition

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

https://doi.org/10.1016/j.compbiolchem.2017.10.004
1476-9271/

of k-spaced amino acid pairs feature extraction based on maximum relevance minimum redundancy (mRMR) feature selection algorithm. Recently, Xu et al. (2017) developed a predictor named Gly-PseAAC to predict glycation sites by using position-specific amino acid propensity and support vector machine (SVM) algorithm. However, the predicted performance of Gly-PseAAC obtained the Matthew's correlation coefficient 0.3166 is not satisfactory, and there is still room for improvement.

To improve the prediction performance of glycation sites predictor, it is important to find an effective feature extraction method to distinguish between the glycation sites and non-glycation sites. In this study, a commonly used feature extraction technique, called Bi-Profile Bayes (BPB) (Shao et al., 2009) was used to encode every training peptide. Based on many aspects of assessments, we found the BPB was more suitable for encoding the protein sequence around the glycation sites than other feature extraction methods including amino acid composition (AAC), pseudo amino acid composition (PseAAC), amino acid factors (AAF), binary encoding (BE) and composition of *k*-spaced amino acid pairs (CKSAAP). Furthermore, by combining BPB feature extraction with SVM algorithm, a novel predictor named BPB_GlySite was constructed to predict glycation sites from protein sequences. As illustrated by 10-flod cross-validation test, the performance of BPB_GlySite outperformed three existing predictors significantly for predicting lysine glycation sites. Finally, we analyzed the importance of the positions around glycation sites based on bi-profile bayes features. Feature analysis showed that the residues in some positions around glycation sites might play the most important role in the prediction of glycation sites. These analytical and predictive results might offer some useful clues for studying the mechanisms of glycation and related experimental validations.

As demonstrated by previous publications (Chou, 2011; Jia et al., 2015; Xu et al., 2014), to establish an accurate glycation sites prediction system, we should carry out the following procedures: (a) construct a valid and reliable training dataset to train and test the prediction model; (b) extract effective features from peptide samples to distinguish between the glycation sites and non-glycation sites; (c) develop a robust and powerful algorithm to operate the prediction; (d) perform proper cross-validation tests to objectively assess the performance of the predictor; (e) establish a user-friendly and accessible web-server for the proposed predictor. Next, we will describe above steps one-by-one.

2. Materials and methods

2.1. Dataset

Xu's training set (Xu et al., 2017) was used to train and test our model. Xu's training set was retrieved from protein lysine modifications database CPLM (Liu et al., 2014), and it consisted of 223 experimentally annotated glycation lysine sites and 446 non-glycation lysine sites. The sliding window was used to represent every lysine residue K of dataset. According to Xu's work (Xu et al., 2017) and our preliminary trials, the window size was set to 15. Thus, every training sample was represented as a peptide segment of length with 7 residues downstream and 7 residues upstream of lysine residue K. To unify the length of each peptide, the added residue 'X' was used to fill the positions without sufficient residues. The glycated peptides were used as positive training samples, while the non-glycated peptides were used as negative training samples.

2.2. Feature construction

As an effective feature extraction technique, Bi-Profile Bayes (BPB) encoding has been successfully applied to various biology problems, including the prediction of protein methylation sites (Shao et al., 2009), O-GlcNAcylation sites (Jia et al., 2013), mitochondrial proteins of malaria (Jia et al., 2011), caspase cleavage sites (Song et al., 2010), type III secreted effectors (Wang et al., 2011). In this study, BPB was used to encode training peptides.

Given a sequence fragment $S = s_1 s_2 ... s_n$, where s_j (j = 1,2, . . . ,n) stands for one amino acid and n denotes the length of the sequence fragment. S belongs to one of two categories, C_1 or C_{-1}, where C_1 and C_{-1} represent glycation sites and non-glycation sites, respectively. According to Bayes' rule, assume that s_j (j = 1,2, . . . ,n) are mutually independent, the posterior probability of S for the two categories can be given by:

$$P(c_1|S) = P(S|c_1)P(c_1)/P(S) = \prod_{j=1}^{n} P(s_j|c_1)P(c_1)/P(S) \quad (1)$$

$$P(c_{-1}|S) = P(S|c_{-1})P(c_{-1})/P(S) = \prod_{j=1}^{n} P(s_j|c_{-1})P(c_{-1})/P(S) \quad (2)$$

Formulas (1) and (2) can be reformulated as:

$$\log(P(c_1|S)) = \sum_{j=1}^{n} \log(P(s_j|c_1)) - \log(P(S)) + \log(P(c_1)) \quad (3)$$

$$\log(P(c_{-1}|S)) = \sum_{j=1}^{n} \log(P(s_j|c_{-1})) - \log(P(S)) + \log(P(c_{-1})) \quad (4)$$

Assume that prior distribution of category is uniform, i.e. $P(c_1) = P(c_{-1})$, the decision function can be represented by Formula (5):

$$\begin{aligned} f(S) &= \mathrm{sgn}(\log(P(c_1|S)) - \log(P(c_{-1}|S))) \\ &= \mathrm{sgn}(\sum_{j=1}^{n} \log(P(s_j|c_1)) - \sum_{j=1}^{n} \log(P(s_j|c_{-1}))) \end{aligned} \quad (5)$$

According to the literature (Shao et al., 2009), Formula (5) can further be written as:

$$f(S) = \mathrm{sgn}(\vec{w} \bullet \vec{p}) \quad (6)$$

where $\mathrm{sgn}(x) = \begin{cases} 1, \mathrm{if} x \geq 0 \\ -1, \mathrm{if} x < 0 \end{cases}$; $\vec{w} = (w_1, w_2, ..., w_n, w_{n+1}, ..., w_{2n})$ is weigh vector; $\vec{p} = (p_1, p_2, ..., p_n, p_{n+1}, ..., p_{2n})$ is the posterior probability vector, p_1, p_2, ..., p_n represent the posterior probability of each amino acid at each position in the positive training peptides (category $C1$) and p_{n+1}, p_{n+2}, ..., p_{2n} represent the posterior probability of each amino acid at each position in the negative training peptides (category $C2$), which is the so-called Bi-profile. Here, the posterior probability was calculated by the occurrence of each amino acid at each position in training peptides. Therefore, every training peptide was encoded as 30-dimensional vectors by BPB encoding scheme. For example, for a given training peptide 'AVTALWGKVNVDEVG', it was encoded as $(p_A^+, p_V^+, ..., p_V^+, p_G^+, p_A^-, p_V^-, ..., p_V^-, p_G^-)$ by BPB encoding. Where '$p_A^+, p_V^+, ..., p_V^+, p_G^+$' mean the occurrence of A,V, . . . ,V, G at each position in positive training peptides, respectively; '$p_A^-, p_V^-, ..., p_V^-, p_G^-$' mean the occurrence of A,V, . . . ,V, G at each position in negative training peptides, respectively.