Gene 576 (2016) 99–104

Contents lists available at ScienceDirect

Gene

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

Research paper

Prediction of sumoylation sites in proteins using linear discriminant analysis

Yan Xu [a], Ya-Xin Ding [a], Nai-Yang Deng [b], Li-Ming Liu [c,*]

[a] Department of Information and Computer Science, University of Science and Technology Beijing, Beijing 100083, China
[b] College of Science, China Agricultural University, Beijing 100083, China
[c] School of Statistics, Capital University of Economics and Business, Beijing, 100070, China

ARTICLE INFO

Article history:
Received 9 April 2015
Received in revised form 24 August 2015
Accepted 28 September 2015
Available online 9 November 2015

Keywords:
F-score
Post-translational modification
Pseudo amino acid

ABSTRACT

Sumoylation is a multifunctional post-translation modification (PTM) in proteins by the small ubiquitin-related modifiers (SUMOs), which have relations to ubiquitin in molecular structure. Sumoylation has been found to be involved in some cellular processes. It is very significant to identify the exact sumoylation sites in proteins for not only basic researches but also drug developments. Comparing with time exhausting experiment methods, it is highly desired to develop computational methods for prediction of sumoylation sites as a complement to experiment in the post-genomic age. In this work, three feature constructions (AAIndex, position-specific amino acid propensity and modification of composition of *k*-space amino acid pairs) and five different combinations of them were used to construct features. At last, 178 features were selected as the optimal features according to the Mathew's correlation coefficient values in 10-fold cross validation based on linear discriminant analysis. In 10-fold cross-validation on the benchmark dataset, the accuracy and Mathew's correlation coefficient were 86.92% and 0.6845. Comparing with those existing predictors, SUMO_LDA showed its better performance.

1. Introduction

Protein sumoylation is an essential post-translational modification (PTM) in proteins by the small ubiquitin-related modifiers (SUMOs). It plays an important role in protein activities, including subcellular transport, transcription, DNA repair and signal transduction (Hay, 2005; Kroetz, 2005; Seeler and Dejean, 2003). It has been found that CTD SUMOylation promotes protein binding and Claspin is one of the SUMOylation-dependent binding proteins. Claspin localizes to the mitotic centromeres depending on mitotic SUMOylation (Ryu et al., 2015). Sumoylation is also discovered to be involved in various diseases and disorders (Dorval and Fraser, 2007; Seeler et al., 2007; Li et al., 2005), especially neural diseases (Shinbo et al., 2006; Dorval and Fraser, 2006), such as Alzheimer's disease and Parkinson's disease. Hence, identifying sumoylation sites in proteins is significant for not only basic researches but also drug developments.

SUMO proteins are highly conserved across eukaryote (Weissman, 2001) including budding yeast, nematodes and vertebrate cells. Identification of sumoylation sites with experimental approaches is significantly limited, labor-intensive and time-consuming for its reversibility and instability. As a complement to experimental methods it is highly desired to develop computational methods to predict potential sumoylation sites.

Some computational methods have been proposed. For instance, Xue and his co-works designed convenient online tools SUMOsp 1.0 (Xue et al., 2006), SUMOsp 2.0 (Ren et al., 2009) and GPS-SUMO (Zhao et al., 2014) based on group-based phosphorylation scoring algorithm. SUMOpre (Xu et al., 2008) based on multiple linear regression and SUMOhydro (Chen et al., 2012) based on the support vector machine were developed by Xu et al. and Chen et al., respectively. Each predictor has its own merits and supplied contributions to the identification of sumoylation sites. In this work, we employed linear discriminant analysis in sumoylation site prediction, which was more rapid and efficient. Features were constructed through integrating physicochemical properties and sequence conservation. The optimal 178 features were selected according to the Mathew's correlation coefficient value in 10-fold cross-validation.

To develop a predictor based on the sequence information, some basic procedures summarized in Chou (2011) should be considered. (i) Construct or select a benchmark dataset to train and test the predictor. (ii) Formulate the protein sequence samples with feature vectors that can truly reflect the correlation with the target to be prediction. (iii) Introduce or develop a useful algorithm (or engine) to operate the prediction. (iv) Perform cross-validation tests to evaluate the performance of the predictor.

Abbreviations: PTM, post-translational modification; LDA, linear discriminant analysis; PSSM, position specific scoring matrix; PSPM, position-specific propensity matrices; CKSAAP, the composition of *k*-space amino acid pair; *F*-score, feature score.

* Corresponding author.
E-mail address: llm5609@163.com (L.-M. Liu).

http://dx.doi.org/10.1016/j.gene.2015.09.072
0378-1119/

2. Materials and methods

2.1. Benchmark dataset

The training data used in this work were derived from Jian Ren's article (Zhao et al., 2014). There were 912 sumoylation sites from 510 proteins. The complete sequences of these proteins were derived from the UniProt (release 2015_07, http://www.uniprot.org/), a database including abundant information of protein biological functions from articles. For every lysine (K) amino acid, the corresponding peptide fragments were generally formulated by

$$\mathbf{P} = \mathrm{R}_{-\xi}\mathrm{R}_{-(\xi-1)}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbf{K}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+(\eta-1)}\mathrm{R}_{+\eta} \quad (1)$$

where the subscripts ξ and η were integers, $\mathrm{R}_{-\xi}$was the ξ-th upstream amino acid residue from lysine (K), while the R_{η} the η-th downstream amino acid residue, and so forth. Hereafter, a peptide was defined as SUMO peptide or non-SUMO peptide if its center K was a sumoylation or non-sumoylation site, that is

$$\mathbf{P}\in\begin{cases}\text{SUMO}, & \text{K was sumoylation site}\\ \text{non-SUMO}, & \text{otherwise}\end{cases} \quad (2)$$

Consequently, $\xi=\eta=10$ were chosen through some trials. When the upstream or downstream in a protein was less than 10, the lacking residues would be filled with the dummy code X. For convenient description, we rewrote Eq. (1) as

$$\mathbf{P} = \mathrm{R}_1\mathrm{R}_2\cdots\mathrm{R}_{10}\mathrm{R}_{11}\mathrm{R}_{12}\cdots\mathrm{R}_{20}\mathrm{R}_{21} \quad (3)$$

where the R_{11} was the center K, $\mathrm{R}_i(i=1,2,\cdots,21,i\neq11)$ was any amino acid at the i-th position. To avoid homology bias, we got 753 peptides in which none had ≥40% pairwise sequence identity to any other. A total of 753 experimentally verified sumoylation and 4518 non-sumoylation sites were derived. Based on large numbers of experiments, the performance was the best when the non-SUMO peptide number was 4518, which was six times of the SUMO ones. The benchmark dataset $\mathbb{S}$ was constituted of

$$\mathbb{S} = \mathbb{S}^+ + \mathbb{S}^- \quad (4)$$

where the positive dataset $\mathbb{S}^+$ contained $N^+=753$ SUMO peptides and the negative dataset $\mathbb{S}^-$ contained $N^-=4518$ non-SUMO peptides, respectively.

2.2. Sample formulation or feature vector

One of the keys to develop a sequence-based computational predictor is to effectively represent its sequences as mathematical expressions (feature construction), which can reflect the intrinsic correlation with the attribute to be predicted (Chou, 2009). There were a variety of feature constructions such as BLOSUM62 matrix (Henikoff and Henikoff, 1992), PSSM (position specific scoring matrix) (Guo et al., 2004; Ding et al., 2014), PSPM (position-specific propensity matrices) (Xu et al., 2014), CKSAAP (the composition of k-space amino acid pair) (Chen et al., 2007, 2008) and so on, which were widely used and showed their effective performance. In this work, three feature construction approaches and five different combinations of them were utilized to represent peptide fragments into mathematical expressions.

2.2.1. AAIndex

Each amino acid has its own specific physicochemical and biologic properties which have direct or indirect effects on protein properties. Different combinations of those properties can also influence structures and functions of proteins. AAIndex (Kawashima et al., 2008) is a database which contains various physicochemical and biologic properties of amino acids. Several combinations of physicochemical properties have been adopted to transform sequence fragments into mathematical expressions, which have shown efficient effects in (Zhao et al., 2013). 14 properties were selected from AAIndex database, including hydrophobicity, polarity, polarizability, solvent, accessible, net charge index of side chains, molecular weight, *PK-N*, *PK-C*, melting point, optical rotation, entropy of formation, heat capacity and absolute entropy. For the dummy amino acid X, it was defined 0 as its physicochemical property value. Therefore, each amino acid was constructed into 14 features through AAIndex database. For a peptide fragment, a 294-D($14\times21=294$) feature vector was obtained through AAIndex database.

2.2.2. PSAAP (position-specific amino acid propensity)

The PSAAP (position-specific amino acid propensity) (Tang et al., 2007) has shown its good performance in Xu et al. (2013). The main idea of PSAAP was to indicate the occurrence frequency of each amino acid appeared on each position.

We used the numerical code 1, 2, 3,…, 20 to represent the 20 native amino acids according to the alphabetic order of their single letter code, and use 21 to represent the dummy amino acid X. We calculated the following 21×21 Position Specific Amino Acid Propensity (PSAAP) matrix

$$\mathbf{Z} = \begin{bmatrix} z_{1,1} & z_{1,2} & \cdots & z_{1,21} \\ z_{2,1} & z_{2,2} & \cdots & z_{2,21} \\ \vdots & \vdots & \ddots & \vdots \\ z_{21,1} & z_{21,2} & \cdots & z_{21,21} \end{bmatrix} \quad (5)$$

where the column corresponded to the position in Eq. (3) and the row corresponded to the amino acid index, respectively. The elements were calculated by

$$z_{i,j} = \mathrm{score}^+(i,j) - \mathrm{score}^-(i,j) \qquad (i=1,2,\cdots,21, j=1,2,\cdots,21) \quad (6)$$

where $\mathrm{score}^+(i,j)$ was the occurrence frequency of the i-th amino acid ($i=1,2,\cdots,21$) at the j-th column ($j=1,2,\cdots,21$) which was derived from the positive benchmark dataset $\mathbb{S}^+$, $\mathrm{score}^-(i,j)$ was the occurrence frequency of the i-th amino acid ($i=1,2,\cdots,21$) at the j-th column ($j=1,2,\cdots,21$) which was derived from the negative benchmark dataset $\mathbb{S}^-$. By the propensity matrix **Z**, the feature vector corresponding to Eq. (3) was formulated as

$$\mathbf{P} = [\psi_1\ \psi_2\cdots\psi_u\cdots\psi_{21}]^{\mathrm{T}} \quad (7)$$

where $\psi_u(u=1,2,\cdots,21)$ was uniquely defined by

$$\psi_u = \begin{cases} z_{1,u} & \text{when } \mathrm{R}_u=\mathrm{A} \\ z_{2,u} & \text{when } \mathrm{R}_u=\mathrm{C} \\ \vdots & \vdots \\ z_{20,u} & \text{when } \mathrm{R}_u=\mathrm{Y} \\ z_{21,u} & \text{when } \mathrm{R}_u=\mathrm{X} \end{cases} \qquad (u=1,2,\cdots,21) \quad (8)$$

where R_u was any amino acid at the u-th position in Eq. (3).

2.2.3. MCKSAAP (modification of composition of k-space amino acid pair)

The composition of k-space amino acid pair (CKSAAP) has been successfully used in predicting mucin type O-glycosylation sites in mammalian (Chen et al., 2007, 2008) and palmitoylation sites (Wang et al., 2009). It can be easily seen that feature vectors formulated by CKSAAP were high dimension and spare. Inspired by PSAAP (position-specific amino acid propensity), modification of composition of k-space amino acid pair (MCKSAAP) was constructed which took the position specificity into consideration to reduce the dimensions. The main idea of MCKSAAP was to indicate the occurrence frequencies of amino acids pairs at different positions, which was inspired by PSAAP method. There were 21 amino acids including the dummy amino acid X and the number of the possible dipeptides was $21\times21=441$.