Contents lists available at ScienceDirect

Neurocomputing

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

Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method

You Zhou [a,b], Tao Huang [b], Guohua Huang [a], Ning Zhang [c,*], XiangYin Kong [b,*], Yu-Dong Cai [a,*]

[a] School of Life Science, Shanghai University, Shanghai, PR China
[b] Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences and Shanghai Jiao Tong University School of Medicine, Shanghai, PR China
[c] Department of Biomedical Engineering, Tianjin Key Lab of Biomedical Engineering Measurement, Tianjin University, Tianjin, PR China

ARTICLE INFO

Article history:
Received 7 August 2015
Received in revised form
4 October 2015
Accepted 10 October 2015

Keywords:
N-formylation
N-acetylation
Post-translational modification
Random forest
Incremental feature selection

ABSTRACT

Post-translational modifications play important roles in cell activities ranging from gene regulation to cytoplasmic mechanisms. Unfortunately, experimental methods investigating protein post-translational modifications such as high-resolution mass spectrometry are time consuming, labor-intensive and expensive. Therefore, there is a need to develop computational methods to facilitate fast and efficient identification. In this study, we developed a method to predict N-formylated methionines based on the Dagging method. Various features were incorporated, including PSSM conservation scores, amino acid factors, secondary structures, solvent accessibilities and disorder scores. An optimal feature set was selected containing 28 features using the mRMR (Maximum Relevance Minimum Redundancy) method and the IFS (Incremental Feature Selection) method. The prediction model constructed based on these features achieved an accuracy of 0.9074 and a MCC value of 0.7478. Analysis of these optimal features was performed, and several important factors and important sites were revealed to play important roles in N-formylation formation. We also compared N-formylation with N-acetylation, another type of important N-terminal modification of methionines. A total of top 34 MaxRel (most relevant) features were selected to discriminate between the two types of modifications, which may be candidates for studying the different mechanisms between N-formylation and N-acetylation. The results from our study further the understanding of these two types of modifications and provide guidance for related validation experiments.

1. Introduction

In eukaryotic cell nuclei, genomic DNA is packaged and ordered by histone proteins into chromatin, the machinery that drives processes such as replication, transcription and DNA repair [16]. Recently, there has been a growing appreciation in the field of chromatin research of the mapping of histone post-translational modifications (PTMs) that play significant roles in gene regulation and chromosome segregation [19,38,8]. In addition to core histones, other nuclear proteins such as the linker histone H1 and the high mobility group have also been reported to carry multiple PTMs that are important for many aspects of nuclear processes [13,53].

To date, considerable progress has been made in the identification of nuclear protein modifications, including acetylation, phosphorylation, methylation and ADP ribosylation [10,17,19,48]. In contrast to the above modifications that have been studied in detail, very little is known about a recently discovered chemical modification. The N-formylation of internal residues is primarily restricted to a few sites in H1 and reflects a PTM specific to chromatin proteins [16,27,54]. It was first recognized as a relatively abundant noncanonical secondary modification by transacylation with the highly reactive 3'-formylphosphate residues that arise from the 5'-oxidation of deoxyribose in DNA and appeared to be uniquely associated with histones and other nuclear proteins. Previous studies have also shown a sharply increased frequency of N-formylation of histones following oxidative stress [27]. Because histones and other chromosomal proteins have been reported to have a slow turnover rate [14,18], it has been speculated that N-formylation could accumulate with age, thereby contributing to the dysregulation of chromatin function. Moreover, N-formylation is stimulated by oxidative DNA damage, which may be involved in

[*] Corresponding authors.
E-mail addresses: zhouyou@sibs.ac.cn (Y. Zhou), tohuangtao@126.com (T. Huang), guohuahhn@126.com (G. Huang), zhni@tju.edu.cn (N. Zhang), xykong@sibs.ac.cn (X. Kong), cai_yud@126.com (Y.-D. Cai).

http://dx.doi.org/10.1016/j.neucom.2015.10.148
0925-2312/© 2016 Elsevier B.V. All rights reserved.

the development of diseases such as cancer.

Usually regarded as an endogenous mimic of N-acetylation and N-methylation due to their chemical similarities, N-formylation may also have functional consequences for histones in DNA binding and chromatin conformation. Furthermore, N-formylation has been observed at residues that are frequently N-acetylated and N-methylated, suggesting the existence of interference with the signaling functions of different PTMs [54].

Given the importance of elucidating the interference with epigenetic mechanisms by N-formylation, there is a serious need to map all of the formylated sites on proteins. Unfortunately, few experimental methods such as high-resolution mass spectrometry have been developed to investigate N-formylated residues [27,54]; moreover, these techniques are not only labor-intensive and expensive but also insufficient to identify all of the modified sites. Accumulating computational approaches have been applied for the identification of various PTMs such as acetylation, ubiquitination, methylation, phosphorylation and glycosylation [15,21,44,5,55,7] and demonstrate high efficiency and accuracy. Following these pioneering works, we developed a method to predict N-formylated methionine sites *in silico* for the first time and analyzed the factors that determined the N-formylation modification.

To better unravel the interplay between N-formylation and N-acetylation, there is a need to be able to discriminate between them in proteins. Herein, we compare the aforementioned N-formylation of the initiator methionine with the N-acetylation of the initiator methionine. This type of N-acetylation is one of two main acetylation forms and usually occurs during translation, thereby affecting protein stability, function and degradation [2,23,5]. Our N-terminal acetylated polypeptides contained a blocked initiator methionine denoted as the Ac-M-X type (Ac indicates an acetyl moiety, M indicates a methionine and X indicates an aspartate (D), glutamate (E) or asparagine (N) [24,5]. However, we did not develop a method for the prediction of N-acetylation sites because many such prediction methods have previously been described by studies such as [5]. We also utilized the same method applied for N-formylation site predictions to obtain an optimal feature set that identified differences in the properties between these two types of PTMs and provided further insights into them.

2. Methods

2.1. Dataset

All of the proteins used in this study were downloaded from the UniProt database (http://www.uniprot.org/, release 2013_07). The sequence-clustering program CD-HIT [35] was applied to remove proteins containing non-experimentally verified modified residues and proteins with sequence identities $>40\%$. That is to say, the protein sequences having pairwise sequence identity greater than 40% to one another were removed.

A total of 27 N-formylated proteins were used as positive samples. However, in nature, non-N-formylated proteins are much more than N-formylated ones. Therefore, in our study of the prediction of N-formylated sites, we randomly selected 81 non-N-formylated proteins as negative controls to make the ratio of positive:negative as 1:3. An additional 411 N-acetylated proteins were also retrieved for the discrimination between N-formylation and N-acetylation. All protein IDs and sequences are provided in Supplementary material I.

The sliding window strategy was utilized to extract positive and negative peptide samples [28,33,56,57]. Because the modified residues for both N-formylation and N-acetylation are the N-terminal residues in the protein sequences, the window was defined by extracting the first 11 residues from the N-terminus (including the modification site) of the protein sequences. Therefore, all sample peptides used in this study were 11 residues long.

For the prediction of N-terminal formylation, the 11-residue peptides with the first residue formylated were regarded as positive samples (27 samples) and the 11-residue peptides with the first residue non-formylated were regarded as negative samples (81 samples).

To discrimination between N-formylation and N-acetylation, the 11-residue peptides with the first residue formylated were regarded as positive samples (27 samples) and the 11-residue peptides retrieved from the N-acetylated proteins were regarded as negative samples (411 samples).

2.2. Feature extraction

We used the following features to encode all of the 11-residue peptides.

2.2.1. PSSM conservation scores

Evolutionary conservation is considered to be important for protein function [28]. This is especially true for conserved residues because they could be under stronger selective pressures. In this study, we computed the conservation status of every residue in a peptide using the Position Specific Iterative BLAST (PSI-BLAST) tool [1] to search the UniRef100 database (Release: 15.10) through 3 iterations with 0.0001 as the *E*-value cutoff. The computed value denoted the probability of the residue against its mutation to the 20 native types of amino acids. For an 11-residue peptide, all such 20-dimensional vectors for the 11 residues in the peptide composed a matrix called a position specific scoring matrix (PSSM). The $20^{*}11=220$ elements in the matrix were called PSSM conservation scores and were used in this study as one of the types of features to encode our peptides. This procedure was similar to that used in the previous study [56].

2.2.2. Amino acid factors

The 20 native amino acids have different physicochemical and biochemical properties, and therefore may affect protein structure and function in different ways. Atchley et al. [3] performed multivariate statistical analyses on the AAIndex [30] database and generated 5 different numerical scores for each amino acid to reflect their five properties: codon diversity, electrostatic charge, molecular volume, polarity and secondary structure. Herein, we used the 5 numerical scores for each residue in each 11-mer peptide (called the amino acid factor features) as another type of feature to encode our peptides. We also used the similar method in our previous study [28]. Note that in this study, because the N-terminal residue at position 1 in each peptide is always M, it was not necessary to incorporate the amino acid factors of that residue; thus, only the amino acid factors of the remaining 10 residues were used.

2.2.3. Structural features

It has been widely regarded that protein secondary structures could play important roles in residue modifications [46]. Therefore, the secondary structure state of every residue in a peptide was computed using SSpro4 [46]. The 3 different secondary structure states 'helix', 'strand', or 'other' for every residue were denoted as '100', '010' or '001', respectively, when encoding the peptides.

The solvent accessibilities of residues could be another factor exerting an effect on residue modifications [52]. Therefore, SSpro4 [46] was also used to compute the solvent accessibilities of every residue in each 11-residue peptide. The 'buried' or 'exposed' state of a residue was represented as '10' or '01', respectively.

Please cite this article as: Y. Zhou, et al., Prediction of protein N-formylation and comparison with N-acetylation based on a feature selection method, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2015.10.148