Artificial Intelligence in Medicine xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Identify and analysis crotonylation sites in histone by using support vector machines

Wang-Ren Qiu[a,b], Bi-Qian Sun[a], Hua Tang[c], Jian Huang[b,*], Hao Lin[b,*]

[a] Computer Department, Jingdezhen Ceramic Institute, Jingdezhen, 333403, China
[b] Key Laboratory for NeuroInformation of Ministry of Education, School of Life Science and Technology and Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu, China
[c] Department of Pathophysiology, Southwest Medical University, Luzhou, 646000, China

ARTICLE INFO

Article history:
Received 27 December 2016
Received in revised form 25 January 2017

Keywords:
PTMs
Crotonyllysine
Sequence information
Support Vector machine

ABSTRACT

Objective: Lysine crotonylation (Kcr) is a newly discovered histone posttranslational modification, which is specifically enriched at active gene promoters and potential enhancers in mammalian cell genomes. Although lysine crotonylation sites can be correctly identified with high-resolution mass spectrometry, the experimental methods are time-consuming and expensive. Therefore, it is necessary to develop computational methods to deal with this problem.

Methods: We proposed a new encoding scheme named position weight amino acid composition to extract sequence information of histone around crotonylation sites. We chose protein data from Uniprot database. A series of steps were used to construct a strict and objective benchmark dataset for training and testing the proposed method. All samples were characterized by a significant number of features derived from position weight amino acid composition. The support vector machine was used to perform classification.

Results: Based on a series of experiments, we found that the sensitivity (Sn), specificity (Sp), accuracy (Acc), and Matthew's correlation coefficient (MCC) were respectively 71.69%, 98.7%, 94.43%, and 0.778 in jackknife cross-validation. Comparison results demonstrated that our proposed model was better than random forest algorithm. We also performed the feature analysis on samples.

Conclusion: Identification of the Kcr sites in histone is an indispensable step for decoding protein function. Therefore, the method can promote the deep understanding of the physiological roles of crotonylation and provide useful information for developing drugs to treat various diseases associated with crotonylation.

1. Introduction

Histone posttranslational modifications (PTMs) play a crucial role in regulating a wide range of biological processes and can extend the chemical repertoire of the 20 standard amino acids by introducing new functional groups such as phosphate, acetate, amide groups, or methyl group. Lysine crotonylation (Kcr) is a newly discovered histone PTM, which is specifically enriched at active gene promoters and potential enhancers in mammalian cell genomes [1]. The haploid male germ cell differentiation program controls essential steps of male gametogenesis and relies partly on a significant number of sex chromosome-linked genes. These genes need to escape chromosome-wide transcriptional repression of sex chromosomes, which occurs during meiosis and is largely maintained in post-meiotic cells. A newly discovered histone lysine modification, crotonylation (Kcr), marks X/Y-linked genes that are active in post-meiotic male germ cells. Histone Kcr conferring the resistance to transcriptional repressors may be a dominant element in making these genes active in the globally repressive environment of haploid cell sex chromosomes. Furthermore, the same mark is associated with post-meiotically activated genes on autosomes. Histone Kcr therefore appears to be an indicator of the male haploid cell gene expression program and a notable element of genome programming in the post-meiotic phases of spermatogenesis.

Post-translational modification of proteins can be experimentally detected by a variety of techniques, including mass spectrometry [2], Eastern blotting [3], and Western blotting [4,5]. Tan and co-authors combined a mass spectrometry-based approach with the histone peptide analysis method and identi-

* Corresponding authors.
E-mail addresses: qiuone@163.com (W.-R. Qiu), Sunbiq@126.com (B.-Q. Sun), Tanghua771211@aliyun.com (H. Tang), hj@uestc.edu.cn (J. Huang), hlin@uestc.edu.cn (H. Lin).

http://dx.doi.org/10.1016/j.artmed.2017.02.007
0933-3657/

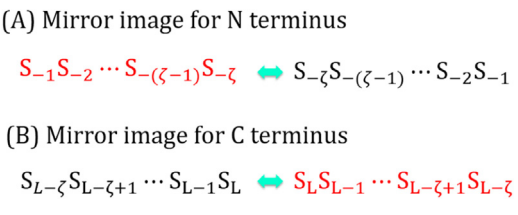

Fig. 1. Schematic illustration of the mirror images of the ξ residues for (A) the C-terminus, and (B) the N-terminus. The red symbol ⇔ represents a mirror. The real peptide segment is colored in black, whereas its mirror image is in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

fied crotonyllysine as a new type of PTM for the first time [1]. Tan et al. also reported that Kcr existed in the eukaryotic cells from yeast to human and that Kcr was different from lysine acetylation in genomic distribution and regulation [6,7]. Bao et al. [8] identified some 'eraser' enzymes for lysine-crotonylated histone marks with chemical proteomics approach. However, the experimental approach to identify crotonyllysine sites is time-consuming and expensive. In a pioneer work, Huang et al. [9] used a discrete hidden markov model to detect histone crotonyllysine sites. However, because the used samples were improper, an overfitting dataset was generated for the training process [9].

In this work, we presented a new approach to predict crotonylation sites based on support vector machine (SVM). Position weight amino acid compositions were utilized to extract sequence features of histones. We analyzed the features, investigated the effect of window length, solved the problem of binary classification with SVM, and evaluated the performance of SVM classifier through jackknife test.

2. Materials and methods

2.1. Benchmark dataset

The benchmark dataset was extracted from the Uniprot database (http://www.ebi.ac.uk/uniprot/). Firstly, with the keywords of histone; human and mouse; we obtained 101 proteins with 343 experimental-confirmed crotonylation sites from the Uniprot database. Secondly; the sliding window strategy was used to extract positive and negative samples from protein sequences; which were represented by peptide sequences with lysine (K) symmetrically surrounded by flanking residues. If the candidate crotonylation sites were near the N- or C-terminus; the missing amino acid was filled with its mirror image (Fig. 1). If their centers had been experimentally annotated as the crotonylation sites; the obtained peptide segment samples were put into the positive subset S_{ζ}^{+}. Otherwise; the obtained peptide segment samples were put into the negative subset S_{ζ}^{-}. Moreover; the redundancy reducing process was also carried out with training datasets. For example; for two crotoylation peptide sequences with 100% identity; when the crotonylation sites in the two histones were in the same positions; only one peptide sequence was kept. After strictly completing the above procedures; we attained 169 positive sites and 847 negative sites for peptides with crotonylation sites. All models in this study are trained and tested on the benchmark dataset. The data can be found from http://lin.uestc.edu.cn/database/Kcr/data.docx. In fact; the performance of our model can objectively and strictly examined by an independent dataset. However; in this study we did not use such a stringent criterion because the currently available data in references do not allow us to do so. Otherwise; the number of sample for some subsets would be too few to have statistical significance. Thus; in the future; we will collect enough data to examine our model.

Table 1
The results of different window sizes on the benchmark dataset by using jackknife test.

ζ	MCC	Acc (%)	Sn (%)	Sp (%)
17	0.567	89.57	56.21	95.61
16	0.528	89.02	49.11	96.25
15	0.778	94.43	71.69	98.70
14	0.698	92.64	58.49	99.05
13	0.652	91.65	54.71	98.58
12	0.654	91.65	66.03	98.22
11	0.694	92.44	64.15	97.75
10	0.672	92.04	58.49	98.34
9	0.515	88.96	38.99	98.34
8	0.502	88.76	36.47	98.58

To make the descriptions more rigorously and clearly, we adopted the Chou's scheme [10] to formulate the peptide samples, according to the method adopted in previous studies on carbonylation sites [11], hydroxyproline and hydroxylysine sites [12,13], human phosphorylatation sites [14], the nitrotyrosine sites [15], multiple lysine PTM sites [16] and protein–protein binding sites [17]. According to Chou's scheme, a potential crotonyllysine site-contained sample can be generally expressed as

$$\mathbf{P}_{\zeta}(K) = S_{-\zeta}S_{-(\zeta-1)}\cdots S_{-2}S_{-1}KS_{+1}S_{+2}\cdots S_{+(\zeta-1)}S_{+\zeta}, \quad (1)$$

where the center K represents "lysine"; the subscript ζ is an integer; $S_{-\zeta}$ represents the ζ-th upstream amino acid residue from the center; the $S_{+\zeta}$ represents the ζ-th downstream amino acid residue, and so forth. The $(2\zeta+1)$-tuple peptide sample $\mathbf{P}_{\zeta}(K)$ can be further classified into the following two categories:

$$\mathbf{P}_{\zeta}(K) \in \begin{cases} \mathbf{P}_{\zeta}^{+}(K), \text{ if its center is a crotonyllysine site} \\ \mathbf{P}_{\zeta}^{-}(K), \text{ otherwise} \end{cases}, \quad (2)$$

where $\mathbf{P}_{\zeta}^{+}(K)$ denotes a true crotonyllysine segment with lysine at its center; $\mathbf{P}_{\zeta}^{-}(K)$ denotes a false crotonyllysine segment with lysine at its center; the symbol " ∈ " means "a member of" in the set theory.

Benchmark datasets generally consist of a set of training data and testing data set. The former is used to train model and the latter is used to test the model. However, as indicated in a comprehensive review [18], if the prediction model is examined by the jackknife test or subsampling (K-fold) cross-validation, the obtained result is actually from a combination of different independent dataset tests and it is not required to artificially partition a benchmark dataset into two parts. Therefore, the benchmark dataset S_{ζ} for the current study can be formulated as

$$S_{\zeta} = S_{\zeta}^{+} \cup S_{\zeta}^{-}, \quad (3)$$

where the positive subset S_{ζ}^{+} only contains the samples of true crotonyllysine segments $\mathbf{P}_{\zeta}^{+}(K)$; the negative subset S_{ζ}^{-} only contains the samples of false crotonyllysine segments $\mathbf{P}_{\zeta}^{-}(K)$ (see Eq. (2)); "⋃" represents the symbol for "union" in the set theory.

According to Eq. (1), the benchmark dataset with different ζ values contains peptide segments with different amounts of amino acid residues:

$$S_{\zeta} \text{ contains the peptides of } \begin{cases} 17 \text{ residuse}, & \text{when } \zeta = 8 \\ 19 \text{ residuse}, & \text{when } \zeta = 9 \\ 21 \text{ residuse}, & \text{when } \zeta = 10 \\ 23 \text{ residuse}, & \text{when } \zeta = 11 \\ 25 \text{ residuse}, & \text{when } \zeta = 12 \\ \vdots & \vdots \end{cases}. \quad (4)$$

However, as shown in Table 1, many preliminary tests indicated that the best outcomes were obtained when ζ = 15 (the sample's

Please cite this article in press as: Qiu W-R, et al. Identify and analysis crotonylation sites in histone by using support vector machines. Artif Intell Med (2017), http://dx.doi.org/10.1016/j.artmed.2017.02.007