Genomics xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Genomics

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

iKcr-PseEns: Identify lysine crotonylation sites in histone proteins with pseudo components and ensemble classifier

Wang-Ren Qiu[a,b,c], Bi-Qian Sun[a], Xuan Xiao[a,b,*], Zhao-Chun Xu[a], Jian-Hua Jia[a,b], Kuo-Chen Chou[b,c,d,**]

[a] Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333403, China
[b] Gordon Life Science Institute, Boston, MA 02478, United States
[c] Center for Informational Biology, University of Electronic Science and Technology of China, Chengdu 610054, China
[d] Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Keywords:
PTMs
Lysine crotonylation
5-Tier coupling
Chou's general PseAAC
Ensemble classifier
Chou's intuitive metrics

ABSTRACT

Lysine crotonylation (Kcr) is an evolution-conserved histone posttranslational modification (PTM), occurring in both human somatic and mouse male germ cell genomes. It is important for male germ cell differentiation. Information of Kcr sites in proteins is very useful for both basic research and drug development. But it is time-consuming and expensive to determine them by experiments alone. Here, we report a novel predictor called iKcr-PseEns that is established by incorporating five tiers of amino acid pairwise couplings into the general pseudo amino acid composition. It has been observed via rigorous cross-validations that the new predictor's sensitivity (Sn), specificity (Sp), accuracy (Acc), and stability (MCC) are 90.53%, 95.27%, 94.49%, and 0.826, respectively. For the convenience of most experimental scientists, a user-friendly web-server for **iKcr-PseEns** has been established at http://www.jci-bioinfo.cn/iKcr-PseEns, by which users can easily obtain their desired results without the need to go through the complicated mathematical equations involved.

1. Introduction

Mounting evidences have suggested that histone PTMs (post-translational modifications) play a crucial role in various biological processes, including cell differentiation and organismal development. Meanwhile, the aberrant modification of histones may cause various diseases such as cancers [1,2]. Lysine crotonylation (Kcr) is an evolutionarily conserved PTM in histone proteins. Therefore, knowledge of Kcr sites in proteins is very important for in-depth understanding the physiological roles of crotonylation [3] and drug development as well. Given a histone protein sequence that contains many lysine (K) residues, can we identify which one can be crotonylated and which one cannot?

Although the information of crotonyllysine can be determined by means of mass spectrometry-based proteomics approach [4], it is time-consuming and expensive. Therefore, it is highly desired to develop computational methods to deal with this problem. Actually, many studies have been carried out for identifying the sites of various types of PTMs in protein and DNA/RNA sequences (see, e.g., [5–28]). All these methods have provided their web-servers that are very useful for most experimental scientists, particularly for those working in the field of drug development [11,29]. But to our best knowledge, no web-server predictor whatsoever available for identifying the crotonylation sites in histone proteins. The present study was initiated in an attempt to fill such an empty area.

As elaborated in a series of recent publications (see, e.g., [30–40]), in order to develop a really useful sequence-based statistical predictor for a biological system, one should observe the following 5-step rules [41]: (1) construct or select a valid benchmark dataset to train and test the predictor; (2) formulate the biological sequence samples with an effective mathematical expression that can truly reflect their intrinsic correlation with the target to be predicted; (3) introduce or develop a powerful algorithm (or engine) to operate the prediction; (4) properly perform cross-validation tests to objectively evaluate the anticipated accuracy of the predictor; (5) establish a user-friendly web-server for the predictor that is accessible to the public. Below, we are to describe how to deal with these steps one-by-one.

* Correspondence to: X. Xiao, Computer Department, Jingdezhen Ceramic Institute, Jingdezhen 333403, China.
** Correspondence to: K.-C. Chou, Gordon Life Science Institute, Boston, MA 02478, United States.
E-mail addresses: wqiu@gordonlifescience.org (W.-R. Qiu), xxiao@gordonlifescience.org (X. Xiao), kcchou@gordonlifescience.org (K.-C. Chou).

http://dx.doi.org/10.1016/j.ygeno.2017.10.008
Received 10 October 2017; Received in revised form 23 October 2017; Accepted 25 October 2017

2. Materials and method

2.1. Benchmark dataset

In this study, the benchmark dataset was constructed as follows. By searching UnitProt database [42] at http://www.uniprot.org/, we obtained 55 histone proteins that contain experiment-confirmed crotonylation sites as given in Supporting Information S1 and 46 histone proteins that do not contain any experiment-confirmed crotonylation sites as given in Supporting Information S2.

For facilitating description later, the Chou's peptide formulation [43–45] was adopted. It has been widely used in many different areas of computational biology (see, e.g., [7–9,12,15,17,19–21,36,46,47]).

According to Chou's scheme [43], a potential Kcr site-containing peptide sample can be generally expressed by

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

where the double-line character $\mathbb{K}$ is used to emphasize the importance of amino acid code K in this study, the subscript ξ is an integer, $\mathrm{R}_{-\xi}$ represents the ξ-th upstream amino acid residue from the center, the $\mathrm{R}_{+\xi}$ the ξ-th downstream amino acid residue, and so forth. The $(2\xi+1)$-tuple peptide sample $\mathbf{P}_{\xi}(\mathbb{K})$ can be further classified into the following two categories:

$$\mathbf{P}_{\xi}(\mathbb{K}) \in \begin{cases} \mathbf{P}^{+}_{\xi}(\mathbb{K}), & \text{if its center is a true Kcr site} \\ \mathbf{P}^{-}_{\xi}(\mathbb{K}), & \text{otherwise} \end{cases} \tag{2}$$

where $\mathbf{P}^{+}_{\xi}(\mathbb{K})$ denotes a true Kcr segment with K at its center, $\mathbf{P}^{-}_{\xi}(\mathbb{K})$ denotes a corresponding false Kcr segment, and the symbol $\in$ means "a member of" in the set theory.

In literature the benchmark dataset usually consists of a training dataset and a testing dataset: the former is for training a model, while the latter for testing it. But as elucidated in a comprehensive review [48], there is no need at all to artificially separate a benchmark dataset into such two parts if the prediction model is tested by the jackknife or subsampling (K-fold) cross-validation because the outcome thus obtained is actually from a combination of many different independent dataset tests. Thus, the benchmark dataset for the current study can be formulated as

$$\mathbb{S}_{\xi} = \mathbb{S}^{+}_{\xi} \cup \mathbb{S}^{-}_{\xi} \tag{3}$$

where the positive and negative subsets, $\mathbb{S}^{+}_{\xi}$ and $\mathbb{S}^{-}_{\xi}$, only contain the true and false Kcr samples, $\mathbf{P}^{+}_{\xi}(\mathbb{K})$ and $\mathbf{P}^{-}_{\xi}(\mathbb{K})$, respectively (see Eq. (2)); while $\bigcup$ denotes the symbol of "union" in the set theory [48].

The concrete procedures to construct the benchmark dataset are given below. (1) By sliding the $(2\xi+1)$-tuple peptide window of Eq. (1) along each of the 55 crotonylated protein sequences in Supporting Information S1 and 46 non-crotonylated proteins in Supporting Information S2, collected were only those peptide segments with $\mathbb{K}$=K at the center. (2) If the segment thus picked around the two ends of the protein is less than $(2\xi+1)$, the lacking code was filled with the same code of its nearest neighbor. (3) The peptide segment sample thus obtained was put into the positive subset $\mathbb{S}^{+}_{\xi}$ if its center was true Kcr site, namely experimentally annotated as "can be crotonylated"; otherwise, into the negative subset $\mathbb{S}^{-}_{\xi}$. (4) To reduce redundancy and homology bias, none of included peptide segments had pairwise sequence identity with any other in a same subset. By strictly observing the above procedures, we obtained an array of benchmark datasets with different ξ values, and hence different lengths of peptide samples as well (see Eq. (1)), as illustrated below

$$\text{Sample length in } \mathbb{S}_{\xi} = \begin{cases} \vdots \\ 13 \text{ residues, when } \xi = 6 \\ 15 \text{ residues, when } \xi = 7 \\ \vdots \\ 35 \text{ residues, when } \xi = 17 \\ 37 \text{ residues, when } \xi = 18 \\ \vdots \end{cases} \tag{4}$$

Besides, the numbers of samples thus obtained would also depend on the value of ξ [15]; i.e.,

$$N(\xi) = N^{+}(\xi) + N^{-}(\xi) \tag{5}$$

where $N^{+}(\xi)$ denotes the number of samples in the positive benchmark dataset $\mathbb{S}^{+}_{\xi}$, while $N^{-}(\xi)$ the number of samples in the negative benchmark dataset $\mathbb{S}^{-}_{\xi}$.

But the preliminary tests had indicated that when $\xi = 17$, i.e., the window's width was of 35 residues, the outcomes were the most promising. Accordingly, hereafter, we would focus on the case of $\xi = 17$; thus Eqs. (1) and (3) can be, respectively, reduced to

$$\mathbf{P}(\mathbb{K}) = \mathrm{R}_{-17}\mathrm{R}_{-16}\cdots\mathrm{R}_{-2}\mathrm{R}_{-1}\mathbb{K}\mathrm{R}_{+1}\mathrm{R}_{+2}\cdots\mathrm{R}_{+16}\mathrm{R}_{+17} \tag{6}$$

and

$$\mathbb{S} = \mathbb{S}^{+} \bigcup \mathbb{S}^{-} \tag{7}$$

where the positive subset $\mathbb{S}^{+}$ contains $N^{+} = 169$ true Kcr samples, while the negative subset $\mathbb{S}^{-}$ contains $N^{-} = 866$ false Kcr samples. The detailed sequences of these samples along with their positions in the proteins as well as the proteins' codes are given in Supporting Information S3.

2.2. Sample representation or formulation

Now let us consider the 2nd step of the Chou's 5-step rule [41]; i.e., how to formulate the peptide sequences concerned with an effective mathematical expression that can truly reflect their essential correlation with the target investigated. For simplifying the description, without losing generality we can convert Eq. (6) into

$$\mathbf{P}(\mathbb{K}) = \mathrm{R}_{1}\mathrm{R}_{2}\cdots\mathrm{R}_{16}\mathrm{R}_{17}\mathrm{R}_{18}\mathrm{R}_{19}\mathrm{R}_{20}\cdots\mathrm{R}_{34}\mathrm{R}_{35} \tag{8}$$

where R_{1}represents the 1st residue of the peptide sample investigated, R_{2} the 2nd residue, R_{3} the 3rd residue, and so forth. Note that for the current study, R_{18} is always fixed at K as shown in Eq. (6) and Supporting Information S3.

Since all the existing machine-learning algorithms, such as SVM (Support Vector Machine) [49,50], KNN (K-Nearest Neighbor) [51], PCA (Principal Component Analysis) [52], and RF (Random Forest) [15,27], can only handle vectors [11], we have to convert the sequential expression of Eq. (8) into a vector. But a vector defined in a discrete model might completely leave out all the sequence-order information. To deal with this problem, the PseAAC (Pseudo Amino Acid Composition) was introduced [53,54]. Ever since the concept of PseAAC was introduced, it has swiftly penetrated into nearly all the areas of computational proteomics (see, e.g., [5,8,36,38,40,55–71] and a long list of references cited in two review papers [29,72]). Encouraged by the successes of using PseAAC to deal with protein/peptide sequences, its idea and approach have been extended to deal with DNA/RNA sequences [13,23,30–32,39,73] in computational genomics/genetics via PseKNC (Pseudo K-tuple Nucleotide Composition) [74–77]. Recently, a very powerful web-server called "Pse-in-One" [76] and its updated version "Pse-in-One 2.0" [77] were developed, by which users can generate any pseudo components for both protein/peptide and DNA/RNA sequences as they wish or define.

According to the concept of the Chou's pseudo components [41,75], the peptide samples of Eq. (8) can be generally formulated as

$$\mathbf{P}(\mathbb{K}) = [\Psi_1 \quad \Psi_2 \quad \cdots \quad \Psi_u \quad \cdots \quad \Psi_\Omega]^{\mathrm{T}} \tag{9}$$