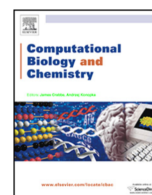
Computational Biology and Chemistry xxx (2014) xxx–xxx

Contents lists available at ScienceDirect

Computational Biology and Chemistry

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



Research Article

Identification and characterization of lysine-methylated sites on histones and non-histone proteins

Tzong-Yi Lee [a], Cheng-Wei Chang [b], Cheng-Tzung Lu [a], Tzu-Hsiu Cheng [a], Tzu-Hao Chang [c,*]

[a] Department of Computer Science and Engineering, Yuan Ze University, Chung-Li, Taiwan
[b] Department of Information Management, Hsing Wu University, New Taipei City, Taiwan
[c] Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan

ARTICLE INFO

Article history:
Accepted 23 December 2013

Keywords:
Post-translational modification
PTM
Lysine
Methylation
Histone
Non-histone
Support vector machine
SVM

ABSTRACT

Protein methylation is a kind of post-translational modification (PTM), and typically takes place on lysine and arginine amino acid residues. Protein methylation is involved in many important biological processes, and most recent studies focused on lysine methylation of histones due to its critical roles in regulating transcriptional repression and activation. Histones possess highly conserved sequences and are homologous in most species. However, there is much less sequence conservation among non-histone proteins. Therefore, mechanisms for identifying lysine-methylated sites may greatly differ between histones and non-histone proteins. Nevertheless, this point of view was not considered in previous studies. Here we constructed two support vector machine (SVM) models by using lysine-methylated data from histones and non-histone proteins for predictions of lysine-methylated sites. Numerous features, such as the amino acid composition (AAC) and accessible surface area (ASA), were used in the SVM models, and the predictive performance was evaluated using five-fold cross-validations. For histones, the predictive sensitivity was 85.62% and specificity was 80.32%. For non-histone proteins, the predictive sensitivity was 69.1% and specificity was 88.72%. Results showed that our model significantly improved the predictive accuracy of histones compared to previous approaches. In addition, features of the flanking region of lysine-methylated sites on histones and non-histone proteins were also characterized and are discussed. A gene ontology functional analysis of lysine-methylated proteins and correlations of lysine-methylated sites with other PTMs in histones were also analyzed in detail. Finally, a web server, MethyK, was constructed to identify lysine-methylated sites. MethK now is available at http://csb.cse.yzu.edu.tw/MethK/.

© 2014 Elsevier Ltd. All rights reserved.

1. Introduction

Post-translational modifications (PTMs) are chemical modifications that occur after a protein has been translated from RNA. It is one of the latter steps in the biosynthesis of many proteins. The cellular regulation of PTMs plays a critical role in many biological processes, which influence the structural and functional diversity of proteomes and determine cellular plasticity and dynamics. Protein methylation is one type of reversible PTM (Paik and Kim, 1967), and it has not been as widely studied as other types of PTM, such as phosphorylation, acetylation, and ubiquitination. Studies over the last few years identified protein methylation as being involved in many important biological processes including transcription, protein–protein interactions, signal transduction, and regulation of gene expressions (Lee et al., 2005; Springer et al., 1979; Stallcup, 2001). Protein methylation reactions can occur either on nitrogen atoms in the N-terminus or on side-chains of lysine (K), arginine (R), histidine (H), proline (P), alanine (A), and glutamine (Q) residues (Clarke, 1993). Among these, lysine and arginine are the most frequently observed amino acids (AAs) methylated. Lysine methylation can be catalyzed by many kinds of transferase enzymes, which are called protein lysine methyltransferases (PKMTs). Lysine residues can attach one, two, or three methyl groups onto an AA's ε-amine group by replacing one, two, or three hydrogen atoms, to respectively yield mono-, di-, or tri-methyllysine (Zhang and Bruice, 2008). Most recent research focused on histones because lysine methylation of histone regulates critical roles in many biological processes. The methylated histones, H3K4 and H3K36, have positive control of gene activity (Pokholok et al., 2005), and H3K27 and H4K20 are associated with repressive modifications in *Saccharomyces cerevisiae* (Cao et al., 2002; Schotta et al., 2004). Additionally, several lysine methylation

* Corresponding author.
E-mail address: kevinchang@tmu.edu.tw (T.-H. Chang).

1476-9271/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2014.01.009

sites on non-histone proteins were discovered to impact protein stability, activity, and functions. For example, p53 was found at K370, K372, and K382 to activate or repress p53 activity (Huang and Berger, 2008). A transcription factor, TAF10, at lysine 189 (K189) is involved in transcriptional regulation of the TAF10 target gene (Kouskouti et al., 2004).

Conventional biological experimental methods for identifying methylation sites on proteins are time-consuming and labor-intensive. Therefore, in silico computational methods are widely used to obtain potential methylation sites for further analysis. As shown in Table 1, different approaches were proposed to identify protein methylation sites. Daily et al. (2005) developed an approach for predicting methylated arginine and lysine based on the biological hypothesis that PTMs preferentially occur in intrinsically disordered regions using a support vector machine (SVM). Sequence are encoded by a set of features including AA frequencies, aromatic content, a flexibility scalar, the net charge, hydrophobic moment, beta entropy, disorder information, and PSI-BLAST profiles. Chen et al. (2006) provided a web server, MeMo, to predict protein methylation of lysine and arginine using an SVM with features from AA information. Experimentally verified methylated sites from UniProtKB/SwissProt (vers. 48) and literature surveys were used as the positive dataset. Shien et al. (2009) presented a method named MASA to identify lysine-, arginine-, glutamate-, and asparagine-methylated sites using combined structural characteristics, including AAs, the accessible surface area (ASA), and a positional weighted matrix. Experimentally confirmed methylation sites were taken from MeMo and UniProtKB/Swiss-Prot (vers. 53), and an SVM was adopted as a learning classifier. Shiu et al. (2009) constructed a prediction server called BPB-PPMS that combines the SVM with feature extraction through Bi-profile Bayes to identify protein methylation sites of arginine and lysine. Experiment data were collected from UniProtKB/Swiss-Prot (vers. 56.1). Using this method, feature vectors were encoded in a bi-profile manner containing attributes from positive and negative position-specific profiles. These profiles were generated in order to calculate frequencies of occurrence of each AA at each position of the peptide sequence in the positive and negative datasets. Most recently, Shi et al. (2012) studied lysine residues with two major modifications, methylation and acetylation. They presented a method called PLMLA that incorporates protein sequence information, secondary structures, and AA properties to predict methylation and acetylation sites on lysine residues. The training and testing dataset were extracted from PhosphoSitePlus (vers. 2011_09) and UniProtKB/Swiss-Prot (vers. 2011_09).

Lysine methylation datasets used in previous studies contained two kinds of proteins, histones and non-histone proteins. Histones possess highly conserved sequences, but there is much less sequence conservation among non-histone proteins. Therefore, mechanisms to identify lysine-methylated sites may greatly differ between histones and non-histone proteins. However, this point of view was not considered in previous studies. Herein, we present an approach to identify lysine-methylated sties on histones and non-histone proteins using SVM models with numerous features, such as the AA composition (AAC), ASA, etc. Features of the flanking region of lysine-methylated sites on histones and non-histone protein were also characterized and discussed. In addition, a gene ontology (GO) functional analysis of lysine-methylated proteins and correlations of lysine-methylated sites with other PTMs in histones were also analyzed in detail.

2. Materials

The UniProtKB/Swiss-Prot Database (Magrane and Consortium, 2011) is a high-quality manually annotated and reviewed section of the UniProt Knowledgebase (UniProtKB). It is a non-redundant protein sequence database with information extracted from the literature and a curator-evaluated computational analysis. Each entry contains the AA sequence, protein name or description, taxonomic data, and citation information. Release 2012_01 of UniProtKB/Swiss-Prot contained 534,242 sequence entries, comprising 189,454,791 AAs abstracted from 206,707 references. PhosphoSitePlus (Hornbeck et al., 2012) is an online systems biology resource that provides comprehensive information and tools for studying experimentally observed protein PTMs and is comprised of over 130,000 non-redundant modification sites that are primarily of human and mouse proteins. The Human Protein Reference Database (HPRD) (Keshava Prasad et al., 2009) is a resource for experimentally derived information about the human proteome including protein–protein interactions, PTMs, enzyme–substrate relationships, and disease associations.

The positive training data were collected from PhosphoSitePlus (release 2012_04), which contains 2756 proteins with 6535 experimentally confirmed methylation sites. We filtered out methylation sites which were not on lysine (K). Since mono-, di-, and tri-methyllysine sites sometimes occur alternately at the same position, we only collected a site once in our dataset. Finally, 44 histones containing 133 lysine-methylated sites and 878 non-histone proteins containing 1306 lysine-methylated sites were collected as respective positive training data for histone and non-histone model training.

Additionally, methylated lysine sites were also collected from the HPRD (release 9) and UniProtKB/Swiss-Prot (release 2010_11), and identical data in training dataset were filtered out from the independent dataset. In UniProtKB/Swiss-Prot, 3023 methylation sites within 1150 proteins were annotated as "methylated lysine" or "methyllysine". Only experimentally verified data were collected, and therefore, lysine-methylated sites with the annotation of "similarity", "potential", or "probable" were removed. The procedure of data collection from the HPRD was the same as that for PhosphoSitePlus. Ultimately, 49 histones containing 153 lysine-methylated sites and 118 non-histone proteins containing 194 lysine-methylated sites were respectively used as the independent data for histone and non-histone model testing.

Non-methylated lysine sites were also collected from the databases mentioned above as the negative dataset. Since the negative training data were much larger than the positive training data, this unbalanced characteristic usually leads to a skewed

Table 1
Comparison of features of previous approaches to identify protein-methylated sites.

Tool	Focus site	Method	Main features	Dataset
(Daily et al., 2005)	Arginine, lysine	SVM	Amino acids, intrinsic disorder	UniProtKB/Swiss-Prot v45
MeMo (Chen et al., 2006)	Arginine, lysine	SVM	Amino acids	UniProtKB/Swiss-Prot v48, PubMed literatures
MASA (Shien et al., 2009)	Arginine, lysine, glutamate, asparagine	SVM	Amino acids, accessible surface area	UniProtKB/Swiss-Prot v53, MeMo dataset
BPB-PPMS (Shiu et al., 2009)	Arginine, lysine	SVM	Bi-profile Bayes	UniProtKB/Swiss-Prot v56
PLMLA (Shi et al., 2012)	Lysine	SVM	Secondary-structure, grouped-weight, and position-weight amino acid composition	UniProtKB/Swiss-Prot v2 011_09, PhosphoSitePlus V2011_09

SVM, support vector machine.

Please cite this article in press as: Lee, T.-Y., et al., Identification and characterization of lysine-methylated sites on histones and non-histone proteins. Comput. Biol. Chem. (2014), http://dx.doi.org/10.1016/j.compbiolchem.2014.01.009

Download English Version:

https://daneshyari.com/en/article/6487115

Download Persian Version:

https://daneshyari.com/article/6487115

Daneshyari.com