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

Deciphering histone code of transcriptional regulation in malaria parasites by large-scale data mining

Haifen Chen[a], Stefano Lonardi[b], Jie Zheng[a,c,*]

[a] School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore
[b] Department of Computer Science and Engineering, University of California Riverside, 900 University Avenue, Riverside, CA 92521, USA
[c] Genome Institute of Singapore, A*STAR (Agency for Science, Technology, and Research), Biopolis, Singapore 138672, Singapore

ARTICLE INFO

Article history:
Accepted 23 December 2013

Keywords:
Histone code
Histone modification
Plasmodium falciparum
Association rule mining
Gene expression prediction

ABSTRACT

Histone modifications play a major role in the regulation of gene expression. Accumulated evidence has shown that histone modifications mediate biological processes such as transcription cooperatively. This has led to the hypothesis of 'histone code' which suggests that combinations of different histone modifications correspond to unique chromatin states and have distinct functions. In this paper, we propose a framework based on association rule mining to discover the potential regulatory relations between histone modifications and gene expression in *Plasmodium falciparum*. Our approach can output rules with statistical significance. Some of the discovered rules are supported by literature of experimental results. Moreover, we have also discovered *de novo* rules which can guide further research in epigenetic regulation of transcription. Based on our association rules we build a model to predict gene expression, which outperforms a published Bayesian network model for gene expression prediction by histone modifications.

The results of our study reveal mechanisms for histone modifications to regulate transcription in large-scale. Among our findings, the cooperation among histone modifications provides new evidence for the hypothesis of histone code. Furthermore, the rules output by our method can be used to predict the change of gene expression.

1. Introduction

1.1. Background

Understanding the epigenetic regulation of transcription by histone modifications is a central problem in molecular biology. The basic unit of chromatin is nucleosome, which consists of an octamer of four core histone proteins around which 147 base pairs of DNA is wrapped. The core histones are subject to many covalent modifications including methylation and acetylation. These histone modifications play essential roles in many biological processes such as transcriptional regulation (Li et al., 2007), X-chromosome inactivation (Brinkman et al., 2006), meiotic recombination hotspots (Wu et al., 2012), and cancer (Esteller, 2007). The activation and repression of gene expression are mediated by histone modifications through direct structural changes of chromatin or recruiting effector proteins (Berger, 2007). In 2001, the hypothesis of 'histone code' was proposed that the combinatorial patterns of histone modifications of different types, positions and times dictate the dynamics of RNA transcription (Jenuwein and Allis, 2001). In the following decade, the histone code has been under intense research (Cheng et al., 2011; Fischle et al., 2003; Kimura et al., 2004; Margueron et al., 2005; Xu et al., 2010). Nevertheless, the existence of the histone code remains an open question. As related high-throughput data (e.g. microarray, RNA-seq, ChIP-chip, and ChIP-seq) become available, it is a good opportunity for computational biologists to crack the histone code by large-scale data analysis and modelling.

In this paper, we aim to study the transcriptional regulation of the most deadly malaria parasite, *Plasmodium falciparum*, which claims about a million human deaths worldwide each year (Nayyar et al., 2012). Despite intense research for decades (Duraisingh et al., 2005; Gupta et al., 2013; van Noort and Huynen, 2006), the mechanism of gene regulation of this species remains a mystery. *Plasmodium* genomes are known for AT-richness, lack of transcription factors, and its unique cyclic patterns of gene expression along its life cycle (Gardner et al., 2002). As there is no evidence of DNA methylation in *P. falciparum*, its histone modifications are likely to be more critical for transcriptional regulation than other species. Recently, there have been efforts to uncover the histone code for *P. falciparum*, which implicate some histone-based mechanisms conserved with other species such as yeast and human (Cabral

[*] Corresponding author at: School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore.
E-mail addresses: hchen009@e.ntu.edu.sg (H. Chen), stelo@cs.ucr.edu (S. Lonardi), zhengjie@ntu.edu.sg (J. Zheng).

1476-9271/$ – see front matter © 2014 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.compbiolchem.2014.01.002

et al., 2012; Cui and Miao, 2010). However, most of them are experimental works focused on only a few gene families crucial for virulence (e.g. *var*, *rif* genes). There is a need for Bioinformatics approaches to systematically elucidating the histone code for the malaria parasite in large scale.

Here, we apply association rule mining to reveal the regulatory relations between histone modifications and gene expression of *P. falciparum*. As a well-studied technique of data mining, association rule mining was firstly proposed to discover the regularities between products in large transactional database of supermarkets (Agrawal et al., 1993). It has also been applied to problems in Bioinformatics (Chen and Chen, 2006; Morgan et al., 2007; Lopez et al., 2008; Wang et al., 2010). Results from Chen's research (Chen and Chen, 2006) show that association rule mining is a promising technique in biological data analysis, which inspires our work. In Wang et al. (2010), association rule mining was applied on histone modifications data of yeast to identify histone modification patterns that might have effects on transcriptional states. However, the rules they generated were among histone modifications themselves instead of between histone modifications and gene expression. By contrast, in our paper the histone modifications and gene expression have been synchronized for each specific gene and time point, and we identify the association rules directly between histone modifications and expression levels of the same genes. The rules discovered as such are more likely to directly uncover the histone code for gene regulation.

We proposed here a framework based on association rule mining to systematically discover rules for histone code in *P. falciparum*. Our method can generate rules which explicitly show how combinatorial histone modifications dictate gene expression levels. Some of the rules have already been reported in literature. Moreover, our experiments showed that, the prediction of gene expression based on our rules is more consistent with observed gene expression than Bayesian network. Our method is robust to changing distribution in data items. To our knowledge, this is the first computational method tailored for the malaria parasite of *P. falciparum*, which takes into account the dynamic change of both gene expression and histone modifications over time points. Our results are encouraging to decipher the histone code and offer valuable insights into the study of epigenetic regulatory mechanisms in transcription.

1.2. *Related works*

Bioinformatics approaches have been proposed to study the relationships between histone modifications and gene expression. Those approaches could be classified into three categories: network-based methods, clustering, and regression models.

In 2008, Yu et al. (2008) built a Bayesian network to infer causal and combinatorial relationships between gene expression and histone modifications. Yu's method can generate tree-like rules between histone modifications and gene expression, which can be viewed as the primary shape of histone code. However, they did not predict gene expression using those rules. In 2012, a correlation-based network of histone modifications, DNA methylation and gene expression was constructed by Su et al. (2012). Although this network can explain some aspects of epigenetic regulation of gene expression, their study can only serve as preliminary results for histone code.

A semi-supervised biclustering method was proposed by Teng and Tan (2012) to find the combinatorial chromatin modifications which are correlated with gene expression in human enhancers. Similar application of clustering methods can also be found in Ha et al. (2011). Although the clustering methods are good at capturing "big pictures", they need further analysis to provide detailed information for histone code.

Regression models are used frequently in studying the relationships between histone modifications and gene expression. Karlić et al. (2010) utilized linear regression techniques to demonstrate systematically that histone modifications levels can predict gene expression accurately. Similar applications of linear regression models can also be found in McLeay et al. (2012), do Rego et al. (2012), Xu et al. (2010). Other regression methods based on machine-learning models have also been proposed in this field, e.g., support vector machine (SVM) (Cheng et al., 2011; Cheng and Gerstein, 2012), Random Forests (RF) and so on (Dong et al., 2012). Although these regression methods can show how well histone modifications can predict gene expression, they are unable to demonstrate how the combinations of histone modifications regulate gene expression in an explicit way. Moreover, these model-based methods produce the relationships between histone modifications and gene expression with a "black box" model which makes it difficult for human to understand and interpret the underlying physical processes and biological meanings.

Overall, most existing methods for the study of histone code are either without prediction of gene expression or lack of interpretability. In the following section, we will present our approach which can generate explicit rules describing how combinatorial histone modifications correspond to particular states of gene expression. Our method provides a straightforward way to decipher the histone code without any assumption of models and the tuning of parameters. Furthermore, our rules can predict the trend of gene expression reasonably well.

2. Methods

2.1. *Data of P. falciparum*

The input data to our analysis are the transcriptional levels and histone modification (HM) enrichment profiles across the *P. falciparum* genome from Rovira-Graells et al. (2012) and Gupta et al. (2013) (with GEO accession number GSE39238). The transcriptional levels were obtained from long oligonucleotides based microarray experiments and the histone modification profiles were from ChIP-on-chip (also known as ChIP-chip) experiments. The dataset includes 14 variables: gene expression (also referred as RNA or cDNA) and 13 histone modification marks, i.e. H4K5ac, H4K8ac, H4K12ac, H4K16ac, H4ac4, H3K9ac, H3K14ac, H3K56ac, H4K20me1, H4K20me3, H3K4me3, H3K79me3, and H4R3me2. The data are time-series, with six time points for each variable, which show the variation of transcription and histone modification

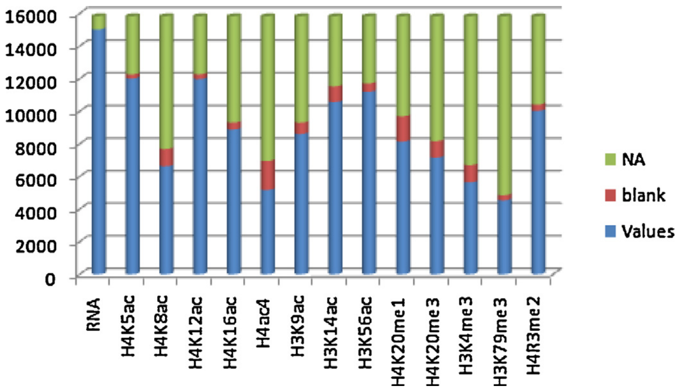

Fig. 1. Distributions of missing data in RNA and different HMs. The vertical axis is the number of probes (for microarray analysis of gene expression and histone modifications), 'NA' means the values are missing for all six time points, 'blank' means there are 2∼5 missing values. Here 'NA' and 'blank' are treated equally as missing data.