Deducing Causal Relationships among Different Histone Modifications, DNA Methylation and Gene Expression

Yunfeng Qi[1], Yan Zhang[1], Jie Lv[1], Hongbo Liu[1], Jiang Zhu[1] and Jianzhong Su[1]
[1]*College of Bioinformatics Science and Technology, Harbin Medical University, Harbin, 150081, China*
yanyou1225@yahoo.com.cn

Abstract

Histone modifications and DNA methylation are two major epigenetic factors regulating gene expression. However, the mechanism in which DNA methylation and histone modifications co-regulate gene expression was little studied. In our study, classifications of DNA methylaion and gene expression showed the complicated relationship between gene expression and epigenetic factors. A Bayesian network was constructed by using the high-resolution maps of histone modifications, DNA methylation and gene expression in human CD4+ T cells to deduce causal and combinatorial relationships among them. PolII was found as the only direct regulator to gene expression, which was not found in prior studies. Our Bayesian network showed that epigenetic factors such as H3K4me3, H3K27me3 and DNA methylation are key regulators of gene expression, though indirectly. However they were considered to combinatorially stablize the state and structure of chromatin.

1. Introduction

Histone modifications and DNA methylation are two major epigenetic factors which have distinct regulatory role in gene expression. DNA methylation is vital in biological functions and plays critical roles in biological processes including gene expression regulation, chromosomal stability, genomic imprinting, and X-inactivation [1]. Histones are modified at its N-terminal tails, including methylation and acetylation. These histone marks affect gene transcription [3]. Epigenetic factors regulate gene transcription in a complicated way and they are coordinated in an uncovered manner to regulate gene expression levels [4-5].

Several studies focused on the activities of epigenetic factors in gene expression regulation. DNA methylation and gene expression have been widely studied in Rakyan et al's work. They found that the regulation of several promoters with a wide range of CpG densities is regulated by DNA methylation mechanism and these genes are associated with tissue-specific transcriptional programs. In addition, they found that only part of the non-promoter CGIs are likely to be regulatory elements as promoter-CGIs are thought to be [2]. Recently, Wang et al. published maps of 39 histone modifications in human CD4+ T cells. They showed that a large number of patterns of histone modifications were associated with promoters and enhancers [4]. Additionally, Yu et al. constructed a Bayesian network to detect the coordinated regulation of histone modifications. They found a lot of modification marks having strong correlation or anti-correlation with gene expression [5]. Now, high-through-put technologies were employed to produce histone modifications and DNA methylation maps in human genome. But what inferred from the data is an essential study, including the elucidation of the coordinated regulation of DNA methylation, histone modification marks to gene expression.

A more comprehensive epigenetic regulation network was built to explain the causal relationships among different epigenetic factors and gene expression with the Bayesian network theory. In this network, 39 histone modifications, PolII, CTCF, H2A.Z, DNA methylation and gene expression were incorporated. DNA methylation and histone modifications in a complex regulatory network seem to be regulating chromatin structure and genome function in a combinatorial pattern, but it is not discovered yet [7]. The major regulatory relationships in Yu et al. are included in our network, but ours seem to be a more comprehensive network. The Bayesian network provided a novel vision in epigenetic regulation. Epigenetic factors such as H3K4me3, H3K27me3 and DNA methylation are not direct regulators of gene expression. They possibly affect the state and structure of chromatin, which in turn influences the gene expression as a consequence.

2. Materials and Methods

2.1. Datasets

978-0-7695-3736-8/09 $25.00 © 2009 IEEE
DOI 10.1109/ICNC.2009.606

The DNA methylation data was derived from Rakyan et al. They reported a novel resource for human genome-wide tissue-specific DNA methylation profiles in 13 normal somatic tissues, in addition to the placenta, sperm, and an immortalized cell line [2]. We only selected human CD4+ T cell data in this study.

The histone modification data was taken from Barski et al. They used ChIP-seq experiment to sequence histone modifications in human CD4+ T cells including 39 histone modifications, variant H2A.Z, RNA polymerase II, and the insulator binding protein CTCF [3, 4]. Additionally, the gene expression data in human CD4+ T cells was also taken from their study.

The annotation of Refseq genes was downloaded from UCSC (NCBI build 36). If missing probes in expression data（GSE10437）, the associated genes were filtered out. As a result, 21,503 genes passed the filtering procedure.

2.2. Associating a gene profile with methylation and histone marks

Refseq genes with methylation annotation were got from the study of Rakyan et al. The probes in their study were aligned to the regions around gene Transcription Start Sites (TSSs) with 10kb upstream and 1kb downstream. Redundant probes were averaged for genes by methylation status. The histone marks were also mapped to TSS-proximal regions but with different cutoffs as (-10kb, 1kb), (-5kb, 1kb), (-2kb, 1kb) and (-1kb, 1kb). Finally, a gene profile associating with methylation and histone marks was built. It is the basis for further analysis.

2.3. Classification study

Initially, two detailed studies were performed focusing on classification based on different labels.

Firstly, a classifying analysis with DNA methylation as the class labels was performed. The cutoffs of low and high methylation level as 0.24 and 0.45 were sat, respectively. The two class labels are associated with nearly equal numbers of genes, with 4141 and 4150 genes in two classes, respectively. We used Naive Bayes and Logistic regression classifiers to do this analysis with 10-fold Cross-Validation procedures. Evaluation values of Accuracy (ACC), precision and AUC under ROC curve were estimated by Weka utilities.

Secondly, another classifying analysis with gene expression as the class labels was also performed. Here, high and low expression gene numbers are 10,561 and 9,360, respectively. Classification and evaluation procedures were the same as the last section.

2.4. Bayesian network analysis

A Bayesian network is constructed based on conditional probability and joint conditional probability distributions to derive feature dependency. Finally, the result is a directed acyclic graph (a graph without loops), where a source edge connected two points, that is, the occurrence of the target node depends on that of the source node.

Here, the WinMine package was utilized to build a Bayesian network. 21,503 genes were used to build the Bayesian network. All the features were discretized including DNA methylation and histone modifications features. k = 3 in k-means clustering was chosen for data discretization, as Yu et al. considered that, the network was the most robust at k = 3, as compared with k = 2 or 4 levels [5]. k = 2 or 4 were also used for data discretization, and effects was worse than k = 3 (data not shown). DNA methylation and gene expression. Pearson correlation coefficient (PCC) in the Bayesian network was computed using R.

3. Result

3.1. DNA methylation status is correlated with other features in the gene profile

Naive Bayes and Logistic regression were used to classify discriminative DNA methylation status of 8291 gene by Weka. 10-fold Cross-Validation procedure was performed to evaluate the gene profiles. The evaluation values in Table 1-4 incorporate ACC, precision and AUC under ROC curve.

The results of evaluation of DNA methylation are shown in Table 1-2 and Fig. 1. The classification efficiency improved, when upstream region was narrowing. In Fig.1, the AUC for (-1kb, 1kb) is also larger than the AUC for (-10kb, 1kb). The core histone marks contribute the DNA methylation most, compared with gene-distal regions. The highest ACC is 89.44% for (-1kb, 1kb) around TSSs, while precision and AUC under ROC is also the same. The results further indicated that histone methylation marks play important roles in predicting the methylation status of CpG islands [6]. Although different upstream cutoffs and different classification methods were selected, the results seemed to be good. In a word, the evaluation proved that DNA methylation state was inferable from histone modifications data.

Table 1. Evaluation results for DNA methylaion by Naive Bayes classifier.

	Accuracy	Precision	ROC
-10k, 1k	74.89%	0.765	0.845
-5k, 1k	76.57%	0.775	0.858
-2k, 1k	78.35%	0.795	0.876
-1k, 1k	80.21%	0.815	0.884

Table 2. Evaluation results for DNA methylaion by Logistic regression classifier.

	Accuracy	Precision	ROC
-10k, 1k	86.71%	0.868	0.941
-5k, 1k	88.43%	0.885	0.949
-2k, 1k	89.42%	0.894	0.954
-1k, 1k	89.44%	0.895	0.954

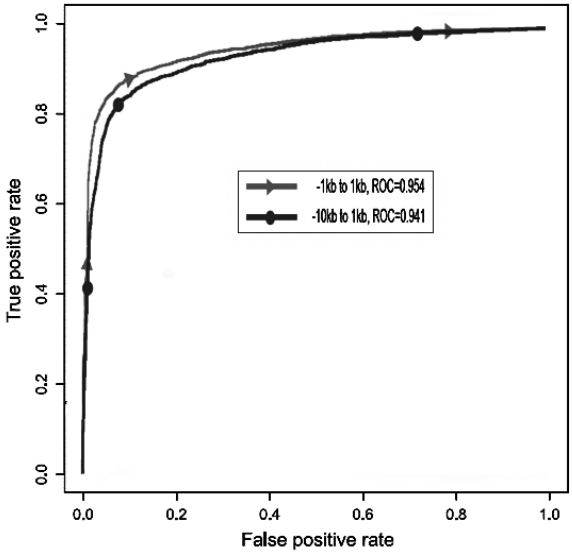

Figure 1. AUC under ROC curves of evaluating on DNA methylation using Logistic regression at -10k, 1k and -1k, 1k gene profile. The result showed that (-1k, 1k) gene profile has a better classification efficiency for DNA methylation.

3.2. Gene expression is correlated with other features in the gene profile

The classification study was performed as the same as last section accept that in this section, the class labels were changed to gene expression level. The results of classification of gene expression are shown in table 3-4. The classification efficiency improved, when upstream region was narrowing. The highest ACC achieves 86.02% for (-1kb, 1kb) around TSSs. The ROC curves for the analysis are not shown.

We supposed that the core histone marks contributes the DNA methylation most, compared with gene-distal regions. The results showed that the gene expression level can be inferred from other data sources, such as DNA methylation and various histone modification marks. Though RNA polymerase II is the most influential factor in regulating gene expression, the contribution of various epigenetic factors to gene expression can not be neglected. Our result further supported the well-known standpoint that DNA methylation and histone modifications have causal relationships with gene expression [3-5]. As the evaluation results seem to be good, in the next section, the Bayesian network was constructed to explore the causal relationship between regulatory factors using cutoff (-1k, 1k).

Table 3. Evaluation results of gene expression by Naive Bayes

	Accuracy	Precision	ROC
-10k, 1k	72.91%	0.745	0.856
-5k, 1k	75.33%	0.768	0.875
-2k, 1k	77.63%	0.789	0.889
-1k, 1k	79.75%	0.807	0.897

Table 4. Evaluation results for gene expression by Logistic regression classifier

	Accuracy	Precision	ROC
-10k, 1k	82.98%	0.831	0.888
-5k, 1k	84.62%	0.847	0.899
-2k, 1k	85.53%	0.856	0.908
-1k, 1k	86.02%	0.862	0.913

3.3. Bayesian network

The high-resolution maps of histone modifications and DNA methylation in human CD4+ T cell were used to build a Bayesian network, which deduced causal and combinatorial relationships among histone modifications, DNA methylation and gene expression. The Bayesian network is shown in Fig. 2.

3.3.1 Causal Relationships of DNA methylation. Li et al. made a high-resolution mapping of epigenetic modifications in the rice genome, uncovering interplay between DNA methylation, histone methylation, and gene expression. They found that heterochromatin had less H3K4me2 and H3K4me3 and more methylated DNA than euchromatin. However, centromeres had a different epigenetic composition. Highly methylated DNA but no H3K4 methylation is prevailing at most transposable element. But there are methylated DNA and di- and/or trimethylated H3K4 at more than half of protein-coding genes [8]. They showed that DNA

methylation and H3K4me3 mark were mutually expelling. Some researchers proposed a standpoint that DNA methylation dictates histone H3K4 methylation [9]. But our results indicated that H3K4 methylation dictates DNA methylation.

The relationship between H3K27me2 and DNA methylation is not clear yet. The relationship between H3K27 methylation and Hox gene silencing has been reported [10]. H3K27 methylation may mediate DNA methylation and Hox gene silencing.

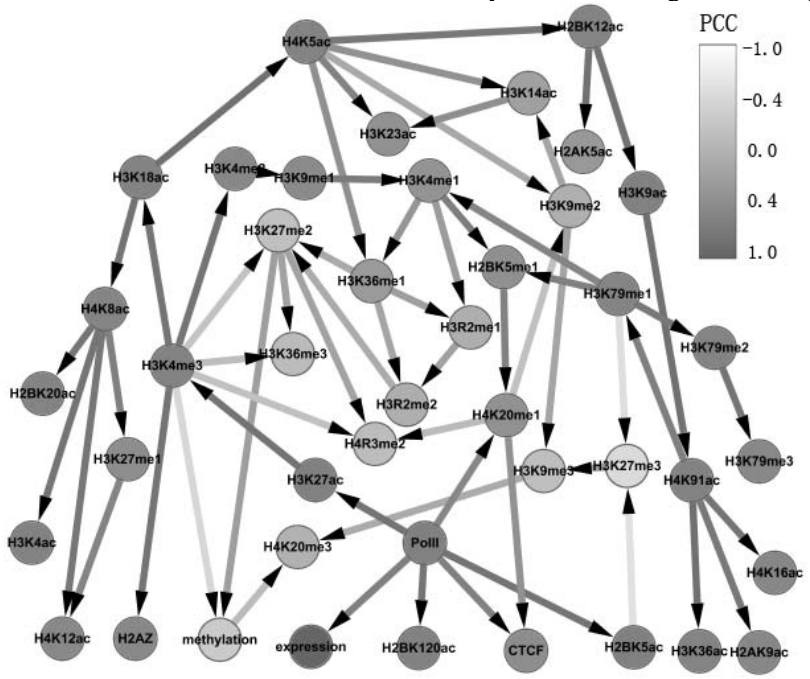

Figure 2. A Bayesian network incorporating 44 codes and 62 edges was constructed. The construction used sequence counts within (-1k, 1k) for mapping histone modification data and (-10k, 10k) for mapping DNA methylation data. The edges of our Bayesian network are colored according to PCC between the two modifications level for the 21,503 genes. The nodes are colored by corresponding PCC between each feature and the gene expression.

3.3.2. Causal Relationships of histone methylation. Causal relationships of Histone methylation have also been studied by Yu et al. Some discrepancies were found between our work and theirs.

H3K79me1→H3K79me2→H3K79me3 was a regulation chain discovered only by our Bayesian network. The inferred regulatory relationship was suggested by Frederiks et al.'s study. They found Dot1 can generate different methylation states of histone H3 Lysine 79 (H3K79) [11]. Our result demonstrated that Dot1 may have different catalysis to mono-, di- and trimethylation states of lysines 79 in histone H3.

H3K4me3→ H3K4me2→ H3K9me1→ H3K4me1 is also a regulation chain. They are supported by a study in female embryonic stem cells. In their study, O'Neill et al. found that H3K4me3, H3K4me2 and H3K4me1 showed the synchronization depletion during X-inactivation in female embryonic stem cells [12]. H3K4me3 regulates H3K4me2 so loss of H3K4me2 occurs 1-2 days later than loss of H3K4me3. And H3K4me1 is not directly regulated by H3K4me3 or H3K4me2. H3K4me1 shows different depleted state, for it has a complex relation with H3K9me1 and H3K79me1.

3.3.3. Causal Relationships of gene expression and PolII. There was no direct regulator for gene expression except PolII in our bayesian network. PolII was the only node which had out-degrees and no in-degree in our bayesian network. In addition, some parts of our bayesian network were not identical with Yu et al. In their study, H3K27me3 was a direct inhibitory regulator to PolII and gene expression. H3K4me3 is a direct active regulator to PolII [5].But these causal relationships had not been discovered in our study. Although lots of epigenetic modifications such as H3K27me3 and H3K4me3 are strongly correlated with gene expression, our model inferred that they do not directly influence on gene expression.

Relationship of PolII→ H4K20me1 and PolII→ CTCF are found in both ours and Yu et al.'s Bayesian networks [5]. Our network shows that PolII not only receives regulation, but also regulate some epigenetic factors.

3.3.4. Causal Relationships of histone acetylations. Histone acetylation marks all had positive correlation with gene expression. Furthermore, they were connected relative closely. Various histone acetylation marks have similar positive influences upon gene expression.

4. Discussion

Bayesian network was built to deduce causal relationships among different histone modifications, DNA methylation and gene expression. But several discrepancies were found between our and Yu et al.'s study, due partly to different genes, different gene expression data and histone acetylation data analyzed. Their Bayesian network is more credible and robust, but ours is more comprehensive. However some causal relationships exist in both Bayesian networks. E.g. PolII → H4K20me1, H3K27me3 → H3K9me3, H3R2me1 → H3R2me2 and so on. These causal relationships, we suppose, are more robust. The result indicates that the two studies have similarity in detail. In addition, the network might be similar in diverse tissues or under different conditions.

Causal relationships extracted here are too hard to be experimentally validated entirely. The landscape of them showed in Fig.2 need more biological evidence to support in future. Methyltransferases and demethylases of DNA and histone should be added to the Bayesian network, and then the epigenetic regulation could be explained in a more systematical manner. Thus, extended epigenetic regulation networks incorporating more factors are needed to give us a broader understanding of the gene expression regulation.

5. Acknowledgements

This work was supported in part by the Natural Science Founda-tion of Heilongjiang Province (Grant Nos.D2007-35), the Heilong-jiang Province Department of Education Outstanding Overseas Scientist grant (Grant Nos. 1152hq28), and the Innovation and Technology special Fund for researchers of Harbin (Grant Nos. RC2007LX003004).

6. References

[1] Bird, A, "DNA methylation patterns and epigenetic memory," *Genes Dev,* vol. 16, pp. 6-21, Jan 1 2002.

[2] V. K. Rakyan, T. A. Down, et al. , "An integrated resource for genome-wide identification and analysis of human tissue-specific differentially methylated regions (tDMRs)," *Genome Res,* vol. 18, pp. 1518-29, Sep 2008.

[3] A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-resolution profiling of histone methylations in the human genome," *Cell*, vol. 129, pp. 823-37, May 18 2007.

[4] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao, "Combinatorial patterns of histone acetylations and methylations in the human genome," *Nat Genet*, vol. 40, pp. 897-903, Jul 2008.

[5] H. Yu, S. Zhu, B. Zhou, H. Xue, and J. D. Han, "Inferring causal relationships among different histone modifications and gene expression," *Genome Res*, vol. 18, pp. 1314-24, Aug 2008.

[6] S. Fan, M. Q. Zhang, and X. Zhang, "Histone methylation marks play important roles in predicting the methylation status of CpG islands," *Biochem Biophys Res Commun*, vol. 374, pp. 559-64, Sep 26 2008.

[7] B. E. Bernstein, A. Meissner, and E. S. Lander, "The mammalian epigenome," *Cell*, vol. 128, pp. 669-81, Feb 23 2007.

[8] X. Li, X. Wang, K. He, Y. Ma, N. Su, H. He, V. Stolc, W. Tongprasit, W. Jin, J. Jiang, W. Terzaghi, S. Li, and X. W. Deng, "High-resolution mapping of epigenetic modifications of the rice genome uncovers interplay between DNA methylation, histone methylation, and gene expression," *Plant Cell*, vol. 20, pp. 259-76, Feb 2008.

[9] C. Y. Okitsu and C. L. Hsieh, "DNA methylation dictates histone H3K4 methylation," *Mol Cell Biol*, vol. 27, pp. 2746-57, Apr 2007.

[10] R. Cao, H. Wang, J. He, H. Erdjument-Bromage, P. Tempst, and Y. Zhang, "Role of hPHF1 in H3K27 methylation and Hox gene silencing," *Mol Cell Biol*, vol. 28, pp. 1862-72, Mar 2008.

[11] F. Frederiks, M. Tzouros, G. Oudgenoeg, T. van Welsem, M. Fornerod, J. Krijgsveld, and F. van Leeuwen, "Nonprocessive methylation by Dot1 leads to functional redundancy of histone H3K79 methylation states," *Nat Struct Mol Biol*, vol. 15, pp. 550-7, Jun 2008.

[12] L. P. O'Neill, H. T. Spotswood, M. Fernando, and B. M. Turner, "Differential loss of histone H3 isoforms mono-, di- and tri-methylated at lysine 4 during X-inactivation in female embryonic stem cells," *Biol Chem*, vol. 389, pp. 365-70, Apr 2008.