Integration of Epigenetic Data in Bayesian Network Modeling of Gene Regulatory Network

Jie Zheng[1], Iti Chaturvedi[1], Jagath C. Rajapakse[1,2,3]

[1]Bioinformatics Research Centre, School of Computer Engineering, Nanyang Technological University, Singapore 639798
[2]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02142, USA
[3]Singapore-MIT Alliance, Singapore
{zhengjie,iti,asjagath}@ntu.edu.sg

Abstract. The reverse engineering of gene regulatory network (GRN) is an important problem in systems biology. While gene expression data provide a main source of insights, other types of data are needed to elucidate the structure and dynamics of gene regulation. Epigenetic data (e.g., histone modification) show promise to provide more insights into gene regulation and on epigenetic implication in biological pathways. In this paper, we investigate how epigenetic data are incorporated into reconstruction of GRN. We encode the histone modification data as prior for Bayesian network inference of GRN. Bayesian framework provides a natural and mathematically tractable way of integrating various data and knowledge through its prior. Applying to the gene expression data of yeast cell cycle, we demonstrate that integration of epigenetic data improves the accuracy of GRN inference significantly. Furthermore, fusion of gene expression and epigenetic data shed light on the interactions between genetic and epigenetic regulations of gene expression.

Keywords: Bayesian networks, gene regulatory network, epigenetics, histone modification, priors, gene expression, yeast cell cycle.

1 Introduction

Reconstruction of gene regulatory networks (GRN) is one of the most important problems in systems biology. Despite intense research, there remain many open problems in this area, partly due to the limited data available and the inherent noise and complexity of biological processes. On the other hand, thanks to the advances on data collection technologies such as next generation sequencing, new types of biological data are emerging, providing new insights and opportunities for GRN reconstruction.

Among the new data, epigenetic data are receiving more attention recently. Epigenetics is the study of changes in phenotypes (especially gene expression) caused by mechanisms other than the changes in DNA sequences (due to mechanisms of central dogma of molecular biology). Such data come from various

epigenetic processes, such as histone modification, DNA methylation, interferences by micro RNA, etc. It is believed that epigenetic control of gene expression represents an important layer of regulation beyond genes in DNA sequences. Analogous to genetic code for gene translation, it is hypothesized that there is an "epigenetic code" for controlling gene expression. The decoding of "epigenetic code" and an increased understanding of the mechanisms of epigenetic regulation of gene expression will bring about new breakthroughs in systems biology and translational medicine. For instance, epigenetic regulation plays an important role in the development of embryonic stem cells, as well as in reprogramming of induced pluripotent stem cells.

Nevertheless, the mechanisms of epigenetic regulation of gene expression remain poorly understood. To elucidate the interaction between genetic and epigenetic regulation of transcription, there is a need for incorporation of epigenetic information in the gene regulatory networks. This paper is motivated by our belief that by considering epigenetic information, the accuracy of GRN reconstruction can be improved. Furthermore, it will shed light on the interactions between genetic and epigenetic regulations.

Recently, there have been some attempts to infer causal relations between epigenetic features (especially histone modifications) and gene expression, and to elucidate the "epigenetic code". Yu et al. built a Bayesian network to model the combinatorial relationships among histone modifications and their effects on gene expression [1]. Cheng et al. gave a machine learning framework to predict gene expression from chromatin features [2]. They observed that chromatin features contribute a significant proportion of gene expression variation. The two papers both looked at the "global" effects of epigenetic features on gene expression while it is desirable to study the effects on individual genes. In particular, it would be highly interesting to examine what patterns occur in epigenetic features between a regulatory gene and its regulated gene. Ha et al. [3] used gene ontology analysis on plant genes to show that genes tend to have similar distribution patterns of histone modifications in the same functional classes but have different epigenetic patterns across different classes. This observation implies that genes involved in the same regulatory pathways have similar patterns of epigenetic features. Thus, the correlation of epigenetic feature distribution among genes can be used as additional information for the reconstruction of GRN.

There are mainly four types of approaches to GRN modeling, namely, information theory models, Boolean networks, differential equations, and Bayesian networks (see the review of [4] and references therein). Among the approaches, Bayesian networks have been the most mature framework for integration of heterogeneous data although analogous integration methods are being developed for other approaches as well. The Bayesian strategy of integration is realized by presenting additional information in the form of prior probability of network. This strategy has been developed to increase the accuracy of GRN reconstruction [5],[6]. The prior knowledge includes protein-protein interactions, transcription factor-DNA binding, sequence motifs, pathways, literature mining, etc. However, to our knowledge, no epigenetic features have been integrated in this framework.

In this paper, we integrate epigenetic features as prior knowledge in Bayesian network learning, based on the framework outlined in [6]. This approach is applied to gene expression data in [7] and histone modification (ChIP-Chip) data in [8] from the yeast genome. We first show that the histone modification profiles between regulators and target genes are more strongly correlated than a random pair of genes. Second, by comparing with benchmark regulatory networks identified from experiments [9], we demonstrated that the use of epigenetic prior can improve sensitivity by more than 10%. Interestingly, it is also observed that histone modification data alone can be used to infer GRN with lower false positive than using gene expression data. To the best of our knowledge, this is the first use of epigenetic information for the reverse engineering of GRN.

2 Methods

Bayesian network has long been used for GRN reconstruction. One of the strengths of Bayesian network is its ability to incorporate additional knowledge through the priors. In the following we will describe the construction of prior matrix B from epigenetic data of histone profiles obtained from [8].

The histone profiles of each gene i consist of a matrix $H_i = \{h^i_{l,h}\}_{t\times m}$ of positive float numbers, where m is the number of histone types and t is the number of loci assessed for the gene. Each row of H_i corresponds to a genomic locus near or within the gene (e.g. promoter, middle of transcribed region, etc.); each column of H_i corresponds to a type of histone modification (e.g. histone H3 lysine 9 acetylation (H3K9ac), histone H3K4 trimethylation (H3K4Me3), etc.). That is, $h^i_{l,h}$ represents the enrichment of the h^{th} histone modification at the l^{th} measured locus of gene i. Thus the genome wide histone dataset in [8] is represented as a 3-dimensional matrix $H = [H_t]^n_{t=1}$, where n is the number of genes.

To simplify analysis, for each gene we calculate the average enrichment of a histone type across t different loci of a gene, and obtain a vector of m float numbers each measuring the level of a certain type of histone modification. These vectors of histone features represent the epigenetic information of genes. Formally, the vector f_i of histone features is calculated for gene i as

$$f_i = \frac{1}{t}\left[\sum_{l=1}^{\tau} h^i_{l,l} \sum_{l=1}^{\tau} h^i_{l,2} \ldots \sum_{l=1}^{\tau} h^i_{l,m}\right] \tag{1}$$

Following [6], we define the biological prior knowledge matrix $B = \{b_{i,j}\}_{n\times n}$ as follows. Let matrix element $b_{i,j} \in [0.0, 1.0]$ represent the correlation between the histone modification patterns of gene i and gene j. If $b_{i,j} = 0.5$, we do not have any prior knowledge about the presence or absence of the edge (i, j); if $b_{i,j} < 0.5$, we have prior evidence of the absence; if $b_{i,j} > 0.5$, we have prior evidence of the presence of the edge. For more details, please see [6] and [5].

To estimate the epigenetic association between gene i and gene j, $b_{i,j}$ is defined by using the Pearson correlation coefficient ρ between the epigenetic

profiles of the two genes. While ρ is a number between -1 and 1, the Bayesian prior need to be between 0 and 1. Hence, we scale to between 0 and 1 linearly to define the Bayesian prior for epigenetic features as

$$b_{i,j} = \frac{1}{2}[\rho(f_i, f_j) + 1] \tag{2}$$

Then, we integrate the priors in B matrix into Bayesian network learning as follows. To construct a gene network G that fits the data the best, the posterior probability of the network is maximize

$$P(G|D) \propto P(D|G)P(G) \tag{3}$$

where $P(G)$ is the prior probability of network G. Let matrix $C = \{c_{i,j}\}_{n\times n}$ be the connectivity matrix of the GRN G, i.e. $c_{i,j} = 1$ if the edge (i, j) is present in G and $c_{i,j} = 0$ otherwise. Following [6], the energies associated with the presence and absence of edges are defined as

$$E(G) = \sum_{i,j=1}^{n} |b_{i,j} - c_{i,j}| \tag{4}$$

Then, the prior probability $P(G)$ is modeled by the Gibbs distribution

$$P(G) = \frac{1}{Z} e^{-\beta E(g)} \tag{5}$$

Here Z is a normalizing partition function defined as

$$Z = \sum_{G\in S} e^{-\beta E(g)} \tag{6}$$

where S is the set of all possible GRNs, and β is a positive number as hyper parameter. To find networks of high posterior probabilities, we search for edges that minimize the energy $E(G)$, thus taking into account the prior knowledge in matrix B.

There are many heuristic and stochastic algorithms for learning Bayesian networks: e.g., genetic algorithms, simulated annealing, Markov Chain Monte Carlo, etc. Since our goal is to integrate epigenetic prior knowledge to improve the accuracy of GRN reconstruction and the effect of a good prior approach should be independent of specific learning algorithms used, we choose a greedy learning algorithm for the present study. As a future work, other structure learning algorithms will be explored.

Starting from a random network (i.e. the entries in the connectivity matrix C are initialized to 0's and 1's at random), the algorithm searches for networks

with good fitness with the given data, through iterations of the following two basic steps: (1) make local change to an existing network by adding, deleting or replacing edges to generate a new network; (2) evaluate the posterior probability as the score of the proposed network. If the score is improved by local change in step (1), then the new network is accepted; otherwise it is rejected. The above two steps are iterated until some stopping criteria are met, e.g., the maximum running time or the maximum number of iterations are reached. The prior probabilities serve as weights in the evaluation of posterior probability of a proposed network. The result consists of a set of network structures (i.e. directed edges) and their corresponding posterior probability scores. In the end, the network with the highest score is output.

3 Experiments and Results

The above method was implemented as a Python program based on the open source library of Pebl [10]. Compared with other implementations of Bayesian network inference, a unique feature of Pebl is to allow easy implementation of soft prior constraints in the form of an energy matrix. However, Pebl is limited to static Bayesian networks and can handle only networks of small number of nodes. Here, we used the greedy learning algorithm implemented by Pebl.

First we use real data to show that the epigenetic profiles are more correlated between a regulatory pair of genes than a random pair. To this end, we downloaded a list of 87 confirmed regulator-gene interactions [11] and the histone modification profiles [8], both in the yeast genome (*S.cerevisiae*). There are totally 88 genes in the confirmed list, out of which 85 genes have histone profiles available, as shown in Figure 1. We did not use the edges in Figure 1 for verification because these are confirmed regulatory interactions but they do not represent a complete regulatory network (e.g. there is no feedback loop) .

The energy values of the 87 edges in the confirmed regulator-gene interactions as in Figure 1 are compared with the energy values of 87 random pairs of genes. As shown in Figure 2, the confirmed regulatory edges have significantly higher energy values (box on the right) than random edges (p-value < 0.01).

The result of Figure 2 shows that epigenetic information correlates with regulatory relationships, which supports our approach of using histone features as prior knowledge in Bayesian network inference. To further verify this approach, we compare Bayesian network algorithms with and without the energy matrix from histone profiles on an experimentally identified regulatory network (Figure 3) consisting of 9 genes related with yeast cell cycle [9].

The greedy algorithm as described in Section 2 for static Bayesian network inference is applied on the expression data of yeast cell cycle [7], with and without the prior of histone data. Moreover, we apply our method on histone data only. For each method, we compare the predicted edges and the established edges as in Figure 3, taking into account the edge directions. True positive (FP) is counted as the number of predicted edges matching the established edges (i.e. with the same end nodes and direction) and similarly for false positive (FP) and

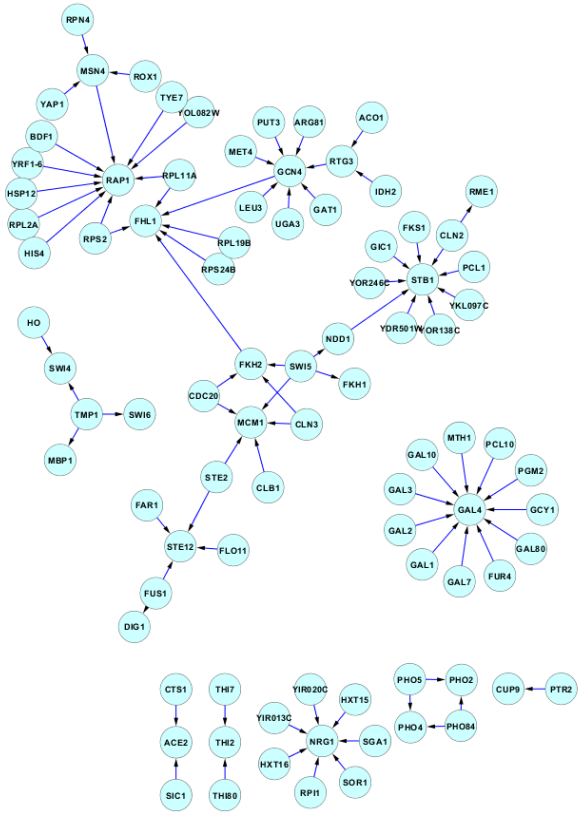

Fig. 1. Confirmed edges of regulator-gene interactions in yeast cell cycle.

false negative (FN). Since there are 9 nodes and edge directions are taken into account, the true negative (TN) should be 72 - (FP+FN+TP).

As shown in Table 1, in Experiment A, only expression data and no prior is used; in Experiment B, only histone profiles are used as input to Bayesian network inference; in Experiment C, both expression data and histone prior are used. The result of C has the highest sensitivity, and compared with result of A the use of prior in C improved the sensitivity by more than 10%. The specificity of C is slightly lower than A due to one more false positive. Interestingly, when only histone data are used, the specificity is the highest among the three experiments. The histone data, when used alone, can reduce FP to 23; however, when used as prior with gene expression in experiment C, both FP and TP increase.

Now, let us look at the predicted GRNs from the three experiments more carefully. As shown in Figure 4, a big fraction of false negative edges in experiment A are due to wrong predicted directions (the T-shaped lines), while in experiment B there are more missing edges (dashed lines). Note that several

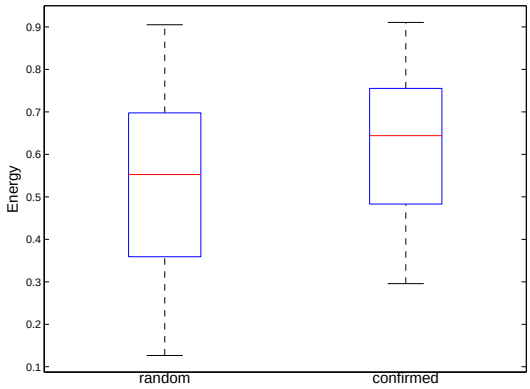

Fig. 2. Comparison between the energy values of random pairs of genes vs. confirmed pairs of genes in yeast (Wilcoxon test p-value < 0.01).

Table 1. Comparison of predicted networks with the benchmark network in three experiments. It shows the improvement of Bayesian network performance due to the use of epigenetic data as prior.

Experiment	TP	TN	FP	FN	Sensitivity(%)	Specificity(%)
A. Expression	5	26	29	12	29.41	47.27
B. Histone	4	32	23	13	23.53	58.18
C. Expression & histone	7	25	30	10	41.18	45.45

edges missed in experiment A have been detected in experiment B (e.g. FKH2 to SWI5, MCM1 to SWI5), and vice versa. This may explain why the result of experiment C, considering both gene expression and histone data, has a higher sensitivity than the other two experiments. To elucidate the effect of epigenetic data on the performance of GRN inference, however, larger datasets and networks will be needed.

4 Conclusion

In this paper, we proposed to integrate epigenetic data as prior knowledge of Bayesian network model for reconstruction of GRN. The approach has been applied to gene expressions of cell cycle and histone modification data of yeast genome. First, it was shown that the correlation of histone modification features between genes with experimentally confirmed regulatory-target gene pairs are stronger than the correlation between random pairs of genes. This suggests that histone features are associated with gene regulatory relation, and hence supports the rationale of our approach. Second, we demonstrated that histone data can also be used to reconstruct regulatory networks with performance comparable to gene expression data. Third, as demonstrated on an experimentally verified network from yeast cell cycle data, epigenetic prior improves the accuracy of

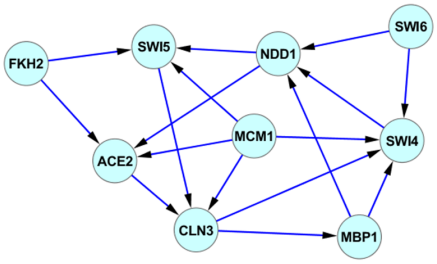

Fig. 3. The gene regulatory network of yeast cell cycle identified in [9].

Bayesian network inference GRN. As far as we know, this is the first paper to reconstruct GRN by incorporating histone modification data, which shows promise for pursuing further research.

However, as this is only a preliminary study in this direction, there remains much room for improvement. The major goal of this paper is to show that the fusion of gene expression with epigenetic data can improve the accuracy of GRN reconstruction. This goal has been achieved with the straightforward methods of GRN reconstruction and Bayesian integration [6]. It is desirable to develop data fusion approaches for more sophisticated GRN reconstruction methods under more realistic conditions in the future. For instance, one can implement a similar prior for dynamic Bayesian network (DBN) which is more compatible with the time-course gene expression data of yeast cell cycle than the static BN used here. Due to scalability of the greedy algorithm of Pebl, the accuracy is relatively low. More powerful learning algorithms with higher computational efficiency will be implemented. Then we can test the approach on larger expression and epigenetic datasets and benchmark networks. Despite the promising results in this paper, our model of epigenetic information is quite simplified, and we should model more realistic relations. For example, the two types of epigenetic data (i.e., histone acetylation and histone methylation) which we integrated here actually have different enrichment patterns along genes: acetylation tends to be enriched at the beginning of genes and methylation tends to be within transcribed regions. There are also combinatorial interactions among histone themselves, which has been modeled also using Bayesian network [1]. In this study we have performed the validation of our method only on a small network (Figure 3). This is mainly because of two reasons : (1) only a few experimentally verified GRNs are available so far due to the difficulty of GRN reconstruction and complexity of data; (2) we need to benchmark GRNs that have also high-throughput epigenetic information. It is still difficult to find a benchmark data that satisfy both criteria (i.e. experimentally verified and with epigenetic data). An important future work is to look for more such data, which can be used for either method evaluation or exploratory data analysis. Last but not least, as

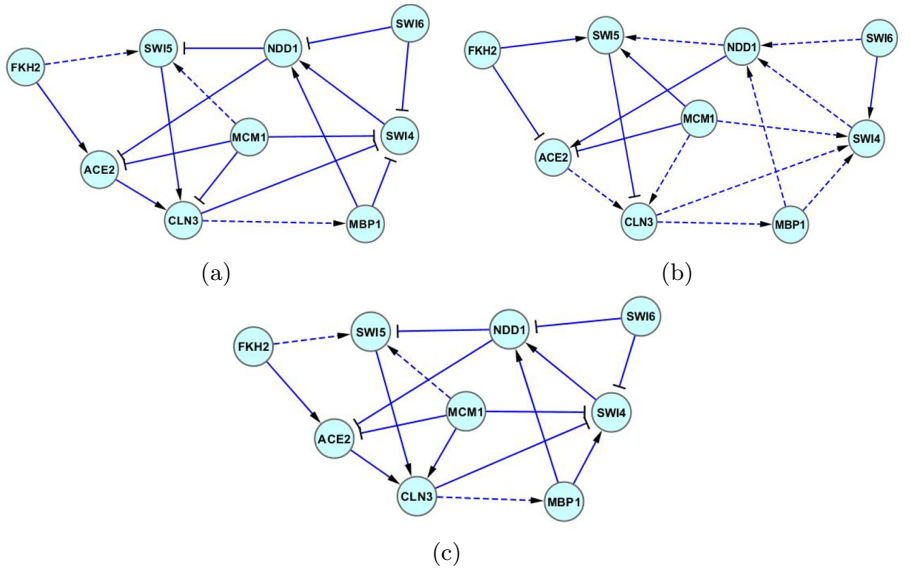

Fig. 4. BN modelling (a) Gene expression data only (b) Histone data only (c) Expression and histone data

there are only a few gold-standard GRNs available, one can experiment with synthetic networks and data taking into account epigenetic regulation.

References

1. H. Yu, S. Zhu, B. Zhou, H. Xue, J.-D. J. Han, Inferring causal relationships among different histone modifications and gene expression, Genome Research 18 (8) (2008) 1314–1324.
2. C. Cheng, K.-K. Yan, K. Yip, J. Rozowsky, R. Alexander, C. Shou, M. Gerstein, A statistical framework for modeling gene expression using chromatin features and application to modencode datasets, Genome Biology 12 (2) (2011) R15.
3. M. Ha, D. W.-K. Ng, W.-H. Li, Z. J. Chen, Coordinated histone modifications are associated with gene expression variation within and between species, Genome Research 21 (4) (2011) 590–598.
4. M. Hecker, S. Lambeck, S. Toepfer, E. van Someren, R. Guthke, Gene regulatory network inference: Data integration in dynamic models–a review, Biosystems 96 (1) (2009) 86–103.
5. S. Imoto, T. Higuchi, T. Goto, K. Tashiro, S. Kuhara, S. Miyano, Combining microarrays and biological knowledge for estimating gene networks via bayesian networks, Journal of Bioinformatics and Computational Biology 2 (1) (2004) 77–98.
6. D. Husmeier, A. V. Werhli, Bayesian integration of biological prior knowledge into the reconstruction of gene regulatory networks with bayesian networks, Computational systems bioinformatics 6 (2007) 85–95.

7. P. T. Spellman, G. Sherlock, M. Q. Zhang, V. R. Iyer, K. Anders, M. B. Eisen, P. O. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization, Mol Biol Cell 9 (12) (1998) 3273–97.
8. D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, R. A. Young, Genome-wide map of nucleosome acetylation and methylation in yeast, Cell 122 (4) (2005) 517–527.
9. I. Simon, J. Barnett, N. Hannett, C. T. Harbison, N. J. Rinaldi, T. L. Volkert, J. J. Wyrick, J. Zeitlinger, D. K. Gifford, T. S. Jaakkola, R. A. Young, Serial regulation of transcriptional regulators in the yeast cell cycle, Cell 106 (6) (2001) 697–708.
10. A. Shah, P. Woolf, Python environment for bayesian learning: Inferring the structure of bayesian networks from knowledge and data, Journal of Machine Learning Research 10 (2) (2009) 159–162.
11. T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L. Murray, D. B. Gordon, B. Ren, J. J. Wyrick, J.-B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, R. A. Young, Transcriptional regulatory networks in saccharomyces cerevisiae, Science 298 (5594) (2002) 799–804.