Computational reconstruction of transcriptional relationships from ChIP-Chip data

Ngoc Tu Le, Tu Bao Ho, and Bich Hai Ho

Abstract—Eukaryotic gene transcription is a complex process, which requires the orchestrated recruitment of a large number of proteins, such as sequence-specific DNA binding factors, chromatin remodelers and modifiers, and general transcription machinery, to regulatory regions. Previous works have shown that these regulatory proteins favor specific organizational theme along promoters. Details about how they cooperatively regulate transcriptional process, however, remain unclear.

We developed a method to reconstruct a Bayesian network model representing functional relationships among various transcriptional components. The positive/negative influence between these components was measured from protein binding and nucleosome occupancy data and embedded into the model. Application on *S.cerevisiae* ChIP-Chip data showed that the proposed method can recover confirmed relationships, such as Isw1-Pol II, TFIIH-Pol II, TFIIB-TBP, Pol II-H3K36Me3, H3K4Me3-H3K14Ac, etc. Moreover, it can distinguish co-locating components from functionally related ones. Novel relationships, e.g., ones between Mediator and chromatin remodeling complexes (CRCs), and the combinatorial regulation of Pol II recruitment and activity by CRCs and general transcription factors (GTFs), were also suggested.

Conclusion: Protein binding events during transcription positively influence each other. Among contributing components, GTFs and CRCs play pivotal roles in transcriptional regulation. These findings provide insights into the regulatory mechanism.

Index Terms—Transcriptional relationship, Bayesian network, ChIP-Chip data, histone modification, nucleosome positioning, chromatin remodeling complex

1 Introduction

Transcription in the context of chromatin is a complex process with the purpose of activating a set of genes in response to environmental stimuli. Nucleosome, the fundamental unit of chromatin formed by wrapping 147bp of DNA around an octamer of histone proteins [1], is known to function as the barrier preventing the access of transcriptional components to DNA sequences. To facilitate transcription in such context, in addition to general transcription machinery, e.g., activators, Mediator, GTFs, and Pol II, the cell must resort to other factors, such as chromatin remodelers and modifiers. The main function of these factors is to alter the interactions of histone-DNA or histone-histone, by which transcription machinery can gain access to *cis*-regulatory elements to initiate the process [2], [3]. Elucidating the interplay of various components involved in transcription is therefore a critical step toward understanding gene regulation.

Technological advances for studying protein-DNA interactions on large scale, e.g., the combinations of chromatin immunoprecipication (ChIP) with high-throughput technologies including DNA microarray (ChIP-Chip) or massively parallel sequencing (ChIP-Seq), have made it possible to produce genome-wide maps of various transcription-related components, such as general transcription machinery and its regulators [4], nucleosomes [5], [6], transcription factors (TFs) [7], [8], histone modifications [9], [10], and other chromatin components [11]. The availability of such data offers unprecedented opportunities to computationally uncover their genome-wide relationships. For example, Wang *et al.* reconstructed a whole-genome map of transcriptional cooperativity among TFs from mapping data [12]; Dai *et al.* identified statistically significant interactions between CRCs and post-translational modifications of histone proteins in *S.cerevisiae* [13]; Stenseel *et al.* reconstructed targeting interactions among 43 chromatin components in *Drosophila* cells [14]. To our knowledge, however, most of the previous works on transcriptional network reconstruction only concentrated on investigating the interactions of either TFs-TFs, TFs-genes, or genes-genes [15]. Consequently, the role of each component as well as how they cooperatively regulate transcriptional process remain elusive. In other words, the whole picture of functional relationships among transcriptional components is far from complete.

We propose a novel computational method to address the above-mentioned problem by reconstructing a Bayesian network-based model representing functional relationships among various transcriptional components. Firstly proposed by Friedman *et al.* to

• *The authors are with the School of Knowledge Science, Japan Advanced Institute of Science and Technology, Nomi, Ishikawa 923-1292, Japan E-mail: {ngoctule, bao, haihb}@jaist.ac.jp*

reveal gene interactions from expression data [17], Bayesian networks (BNs) have been widely applied to model many kinds of biological interactions, such as protein-protein interactions [18], [14], protein signaling networks [19], and interactions of histone modifications [20]. Employing the search-and-score approach, we develop a method to infer network structure in the context where no positive training data is available. Moreover, based on the observation that transcriptional components all work on chromatin substrate and cause change in nucleosome organization, the positive/negative influence between any two functionally related components was measured from binding and nucleosome positioning data and integrated into the network model.

When applied on genome-wide ChIP-Chip data of 36 transcriptional components in *S.cerevisiae*, including general transcription machinery, chromatin remodeling complexes, and histone modifications, our method not only recovers previously confirmed relationships, e.g., Rsc9-Rfx1, Isw1-Pol II, TFIIH-Pol II, Pol II-H3K36Me3, TFIIH-H3K4Me3, H3K4Me3-H3K14Ac, but also suggests several new ones, e.g., Mediator-CRCs, (CRCs,GTFs)-Pol II, which may provide insights into transcriptional regulation. Resulting network model showed that transcriptional components positively influence each other. Among them, GTFs, Mediator, and CRCs play critical roles in regulating the outcome of the whole process. Our method can also be extended to reconstruct more accurate model as data on other aspects of transcription become available.

2 METHODS

2.1 Materials

2.1.1 Histone modification

Genome-wide ChIP-Chip data of 8 histone modifications, including H3 K9Ac/K14Ac, H4Ac, H3K4Me1/2/3, H3K36Me3, and H3K79Me3, were taken from Pokholok *et al.* [21]. Modification levels at each promoter, defined as the region from 500bp upstream to 100bp downstream around Transcription Start Site (TSS), were measured as the averages of overlapping probes or the nearest one, in case of no overlap.

2.1.2 Protein binding

Binding data of diverse representative components of the gene regulatory machinery in *S.cerevisiae*, including 8 sequence-specific transcription factors (TFs), 8 chromatin remodelers, 6 GTFs and TATA-binding protein (TBP), 2 components of Pol II, and 3 components of Mediator, were taken from Venters and Pugh [4]. Binding profile at the promoter of these components was assigned as the average of binding levels of corresponding TSS and Upstream Activating Sequence (UAS). After removing promoters lacking binding information and combining with histone modification data above, we received a contingency table of 36 columns and 4498 rows for further analysis.

2.1.3 Nucleosome occupancy

Genome-wide nucleosome occupancy data at 4bp resolution were taken from Lee *et al.* [22]. Occupancy profile for each promoter was derived as follows. At first, the promoter was divided into 4 bins of 150bp long. Then, bin occupancy value was calculated as the sum of the occupancy levels of all probes belonging to the bin. Each occupancy profile was finally represented by a 4-dimensional vector $(Occ_1, Occ_2, Occ_3, Occ_4)$, where Occ_i $(i = 1, \ldots, 4)$ was the nucleosome occupancy level at bin i.

2.1.4 Genomic annotations

Genomic annotations of *S.cerevisiae* (SGD/sacCer2 assembly) were extracted from the tables provided by the UCSC Genome Browser [23].

2.2 Bayesian network

2.2.1 Definition

A Bayesian network (BN) for a set of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is a probabilistic model consisting of two components [24], [25]:

- A network structure S, which is a directed acyclic graph (DAG), representing conditional independence relationships among variables in $\mathbf{X}$.
- A set P of local probability distributions associated with each variable.

Markov condition guarantees that these two components, (S, P), encode a joint probability distribution on $\mathbf{X}$, given by:

$$p(\mathbf{x}) = \prod_{i=1}^{n} p\left(x_i | \mathbf{Pa}_i\right)$$

in which the terms of the product on the right hand side correspond to the local probability distributions P, and $\mathbf{Pa}_i$ are the parents of x_i.

2.2.2 BN structure learning

As mentioned above, a BN contains two components. Thus, there are two steps in learning a BN model from data: parameter learning, which specifies the local probability distributions P, and structure learning, which identifies the structure S. The main target of our work is to uncover the dependencies among transcriptional components, hence we focus on the latter learning problem.

Score-based search method was employed to learn a BN structure representing the functional relationships among transcriptional components. The aim of this method is to identify network structures that "best" describe the data by some measure. A search procedure, starting from an initial structure, explores the

space of possible network structures step-by-step. At each step, it scores the corresponding structure to identify the network with maximum score. Because exhaustive search in the structure space is infeasible [26], greedy hill-climbing can be used as search strategy. Simulated annealing approach may be applied to escape from local maxima.

To score a candidate network, we used a Bayesian scoring metric, which was originated from [27], and further developed by [24] as Bayesian metric with Dirichlet prior and equivalence (BDe) metric:

$$p(S|D) \propto$$
$$p(S)\prod_{i=1}^{n}\prod_{j=1}^{q_i}\left[\frac{\Gamma(N_{ij}^{'})}{\Gamma(N_{ij}^{'}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N_{ijk}^{'}+N_{ijk})}{\Gamma(N_{ijk}^{'})}\right]$$

where n is the number of variables, N_{ijk} is the number of instances in the data set D having variable x_i in state k with its parents in the j-th instantiation in current structure S, $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$ and $\Gamma(.)$ is Gamma function. $N_{ijk}^{'}$ and $N_{ij}^{'}$ have the same meaning but correspond to prior knowledge for the parameters. When no prior knowledge is available, they can be estimated as $N_{ijk}^{'} = N/(r_i q i)$ with N is the equivalence sample size, r_i is the number of states of variable x_i and q_i is the number of instantiations of the parents of variable x_i. Finally, $p(S)$ is the prior probability of the structure. In our work, we assumed the uniform distribution on the structure S.

2.3 Partial correlation

Assume that X and Y are two random variables and $\mathbf{Z} = (Z_1, Z_2, ..., Z_p)$ is a set of controlled random variables. Partial correlation is employed to measure the relationship between X and Y after eliminating the influence of $\mathbf{Z}$. The relationship between X and $\mathbf{Z}$ can be estimated via a linear regression model $X = \alpha_X + \mathbf{Z}\beta_X + \mu_X$, similarly for that between Y and $\mathbf{Z}$. The partial correlation between X and Y while controlling $\mathbf{Z}$ is measured as the Pearson correlation between μ_X and μ_Y.

3 Results and Discussion

3.1 Reconstruction of the network model

3.1.1 Discretization.

The contingency table described in *Materials* was input to network inference algorithm. Because the network model described in section 2.2 only accepts discrete variables, the next step is to transform data into discrete values. The discretization of histone modification and protein binding data is biologically relevant because biological outcomes have been shown to be related with different states of histone modifications and/or transcriptional machinery bindings at regulatory regions [21], [9], [4]. Basically, the discretization can be done in various ways, and this may affect the resulting model [30], [29]. Previous investigations, however, have shown that 2- or 3-category schemes could be suitable choices [20], [14]. Also, the structure learning algorithm employed here has been shown to perform best with 3-category discretized data [29]. Thus, we chose to discretize each feature into 3 values: "low" (0), "medium" (1), and "high" (2), using 3 discretization schemes, as suggested in [14]. The ranges were either 0th-10th, 11th-90th, 91th-100th percentile (Scheme 1), or 0th-20th, 21th-80th, 81th-100th percentile (Scheme 2), or 0th-33th, 34th-67th, 68th-100th percentile (Scheme 3), for each feature.

3.1.2 Setting for BN Inference.

The structures of static BNs were inferred with Banjo (http://www.cs.duke.edu/∼amink/software/banjo/), which supports the network model described in *Methods*. Empirical running showed that, with more than $1,300,000$ search iterations the network score was not significantly improved, so each search was set to finish at this number of iterations.

3.1.3 Bootstrapping and selection of the threshold for confidence scores

As the search-and-score method may output a different network on each run, the highest scored graph, bootstrapping method proposed in [17] was employed to estimate the confidence scores of edges in the resulting network. Given a dataset D of N instances, a new dataset $D^{'}$ was created by resampling with replacement N times from D. Then a BN was inferred on $D^{'}$. These two steps of resampling and inferring were repeated m times, generating m different BNs. The confidence score of each edge was estimated as the proportion of networks containing that edge. We also employed data permutation method from [17] to assess the credibility of confidence estimates. By which 10 randomly permuted datasets were created from original data, for each a network model with corresponding confidence scores was derived by the same procedure. Figure 1 illustrates confidence score distributions of permuted (average of 10) and original data corresponding to 3 above-mentioned discretization schemes. It shows that random data could not generate relationships with confidence score higher than 0.2. Thus relationships with confidence scores greater than 0.2 can be considered significant.

In order to select a threshold for confidence scores (in other words, to decide whether an edge is included in the resulting model or not), Steensel *et al.* used a known physical and genomic interaction database of proteins (BioGRID) as the surrogate reference list for resulting interactions [14]. Because no such database exists for transcriptional relationships, we proposed the following cross-validation based method to select

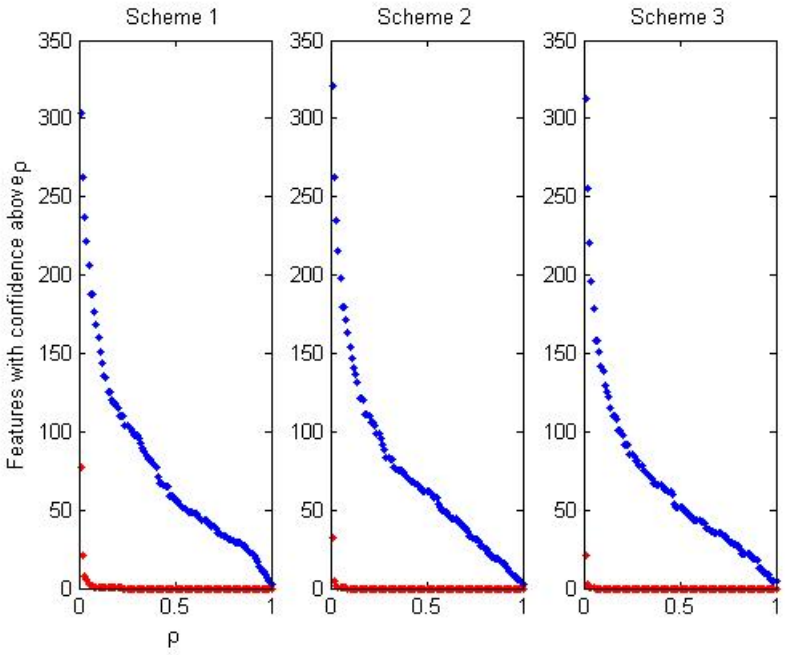

Fig. 1. Confidence score distributions of original (blue) and randomly permuted (red) data

the threshold, denoted by τ, for confidence scores. Our method is motivated by the work of Yu *et al.*, which also employed cross-validation based approach to deal with the problem of lacking positive training data when inferring causal relationships of histone modifications [20]. At first, input data D was split into two equal parts, D_1 and D_2, T times. Each time, 3 bootstrapped BNs corresponding to D, D_1 and D_2, named $globalBN$, $partialBN_1$, and $partialBN_2$, respectively, were learned as above. Then, we defined a measure, acc_i:

$$acc_i = \frac{\sharp(partialBN_i \cap globalBN)}{\sharp(partialBN_i)}, i = \{1, 2\}$$

where the denominator is the number of edges of $partialBN_i$ and the numerator is the number of edges that appear in both $partialBN_i$ and $globalBN$.
The selection criteria was chosen as:

$$Stability = \sum_{i=1}^{2} SNR(acc_i)$$

where $SNR(acc_i)$ is the signal-to-noise ratio of acc_i after T times of data splitting and network learning steps. It is easy to see that acc_i ($i = \{1, 2\}$), thus $Stability$, are the functions of τ. We then chose τ that maximized $Stability$ as the cut-off threshold. Intuitively, acc_i ($i = \{1, 2\}$) measure how well 2 cross-validated networks agreed with the network derived from complete data. A reasonable choice of τ and discretization scheme would therefore make this agreement most stable no matter how cross-validated data were derived.

To assess the effectiveness of the proposed method, we created 9 random discrete BNs with following structures: (8 nodes, 10 edges), (9 nodes, 15 edges), and (10 nodes, 14 edges). Each node represented either 2-, 3-, or 4-category discrete variable. From each of these BNs, 3 datasets of 5000 instances were randomly simulated. All simulation steps were done using Tetrad (http://www.phil.cmu.edu/projects/tetrad). We then applied the proposed method to reconstruct network structures from these 27 simulated datasets. The number of cross-validation T was set from 5 to 15, and the number of bootstrapping N to 100 for computational reason. Threshold τ was searched in the range of $[0.5; 0.9]$ with step of 0.05. We employed the 2 measures defined in [29], namely Recall (Rec) and Imprecision (Impre), to evaluate reconstruction performance. Table 1 presents average values of Rec, Impre, and *Stability* corresponding to 9 simulated BNs. It shows that the proposed method gave the best performance on 3-category data, which also gave $Stability$ the highest values, no matter how complicated network structures were. The performance was even higher if we do not care about the directions of the edges: with 3-category data, average Rec and Impre values for above 3 structures were (80,0), (93.33,0), and (95.24,0) (in % scale), respectively. This means the proposed method could correctly recover many edges while no false possitive was included in the resulting models.

The experiment also showed that the performance on 22 of 27 simulated datasets did not change with $T \geq 10$ (data not shown). Thus, to derive BN model for transcriptional relationships, we randomly split input data into two equal datasets 10 times, resulting

TABLE 1
Performance of the proposed method on simulated data

Network structure	2-category			3-category			4-category		
	Rec(%)	Impre(%)	*Stability*	Rec(%)	Impre(%)	*Stability*	Rec(%)	Impre(%)	*Stability*
8 nodes, 10 edges	56.67	24.07	3.53	80	0	1.3E+15	56.67	0	36.17
9 nodes, 15 edges	42.22	0	42.68	80	14.29	128	75.55	15.02	50.19
10 nodes, 14 edges	52.38	20.2	31.81	88.1	8.87	175	50	7.41	38.35

in 20 datasets. For each, bootstrap procedure was run 100 times to generate a corresponding consensus BN. Threshold τ was searched in the range of $[0.5; 0.9]$ with step of 0.05. Table 2 shows the values of $Stability$ with corresponding values of τ for each discretization scheme. The value of τ and the discretization scheme that gave the highest value to $Stability$, $\tau = 0.6$, $Scheme\ 2$, and $Stability = 72.2043$, were chosen. To infer the network structure, bootstrap procedure was run 1000 times on the whole data to identify confidence score for each edge. After setting $\tau = 0.6$, we received a BN of 48 edges (Fig. 2), representing functional relationships among transcriptional components in question. Confidence scores of the network are given in the supplementary material (graphWeights.xls).

3.2 Transcriptional components show positive effects on each other

One disadvantage of BN in modeling biological interactions is that it does not contain information about positive/negative influence between the variables. For example, it is hard to know whether the binding of one protein would promote or inhibit the binding of another protein from the corresponding network model [14]. To enhance the semantic of BN models, one can impose constraint on the sign (+ or -) of each edge and learns it directly from model parameters [28], [29]. This approach, however, increases the cost of model reconstruction. Furthermore, because the signs are deduced from the highest scored graph, it is not appropriate in our context, where the resulting model is inferred from the consensus graph. Hence, we propose an alternative way to derive the signs of the relationships among transcriptional components. From the observation that these components all work on chromatin substrate and they either directly or indirectly alter nucleosome organization [31], the influence between two components was measured as the sign (+ or -) of their partial correlation value considering nucleosome profile as controlling variables. Concretely, partial correlation was computed for the two components based on their binding profiles with nucleosome profile, represented by a 4-dimensional vector described in *Methods*, as 4 controlling variables.

The result, available in the supplementary material (graphSigns.xls), shows that almost all relationships were marked as positive (+), i.e., regulatory components may positively influence the activity of each other. This is consistent with the observation that during transcription, especially at initiation stage, regulatory components are cooperatively recruited and the bindings of some components may promote the bindings of the others to facilitate the process [34], [35]. Notably, we found two relationships marked as negative, $H3K4Me3 \rightarrow H3K4Me1$ and $PolII(Rpo21) \rightarrow TFIIA(Toa2)$. The former is supported by the finding of Morillon *et al.* that the appearance of H3K4Me3 was coincident with the drop of H3K4Me1 at the promoter of MET16 gene [32]. Though, we did not find any direct evidence for the latter one. It is possibly due to the removal of GTFs, including TFIIA (Toa2), after the full PIC is assembled and move to elongation step. Another relationship also reported by Morillon *et al.* is the negative effect of Isw1 on the activity of Pol II. In [32], Isw1 is assumed to be recruited to the promoter of MET16 to prevent the moving of Pol II and the early onset of transcription. The link $Isw1 \rightarrow PolII(Rpo21)$ in our model, however, was assigned with positive mark. This contradiction may be explained by the combinatorial effects of other co-locating CRCs, such as Ino80, Rsc9, and Swi3, on the activity of Isw1 [33].

3.3 Resulting network model confirms previously reported relationships

Comparing the occupancy of transcription machinery and chromatin regulators at promoter regions, Venters *et al.* [4] found that they were clustered into 6 groups whose members occupied a common set of genes, including {Isw2,Ioc2,Ioc3,Rsc9}, {Tfg1,Tfa1,Swr1,Taf1}, {Swi3,Ino80,Isw1,Spt3}, {Ssl1,Sua7,TBP}, {Nut1,Srb5,Rgr1}, and {Rpo21,Toa2}. Their further investigation of the "location-linkage" between sequence-specific TFs and CRCs showed that Rap1, Ifh1, and Cin5 did not exhibit significant co-occupancy with any tested CRCs, while Rfx1, Xbp1, and Yap6 showed strong co-occupancy with all tested CRCs except SWI/SNF (Swi3). Among the latter three, Xbp1 and Rfx1 also displayed location-linkage with co-occupying CRCs. These co-occupancy and location-linkage raised the question that whether they reflect the functional relationships among factors or just are the indirect consequence of interactions between those factors with other linked proteins [35]. Our network model provides support for both hypotheses. For example, in the latter four among 6 groups above, all of their members have at

TABLE 2
The value of *Stability* with corresponding τ and discretization scheme

τ	*Stability (Scheme 1)*	*Stability (**Scheme 2**)*	*Stability (Scheme 3)*
0.5	36.8696	54.858	62.9467
0.55	46.7434	49.9313	48.9445
0.6	54.2189	**72.2043**	49.7512
0.65	43.4048	66.1212	36.767
0.7	38.7191	70.4882	37.2507
0.75	40.7891	27.0812	37.0849
0.8	32.5331	29.6679	45.0156
0.85	26.2679	15.8474	37.3976
0.9	31.011	25.5555	37.463

TABLE 3
The top dominant factors with corresponding dominance scores ($dScore$) given by our method.

Factor	$dScore$ ($k \geq 2$)
$Xbp1$	4.0648
$Srb5$	3.9916
$Sua7$	2.9605
$H3K4Me3$	2.5182
$Ssl1$	2.4526
$Toa2$	2.3758
$Ino80$	2.1942
$H4Ac$	2.0803
$Rsc9$	1.7098
$Rpo21$	1.1832

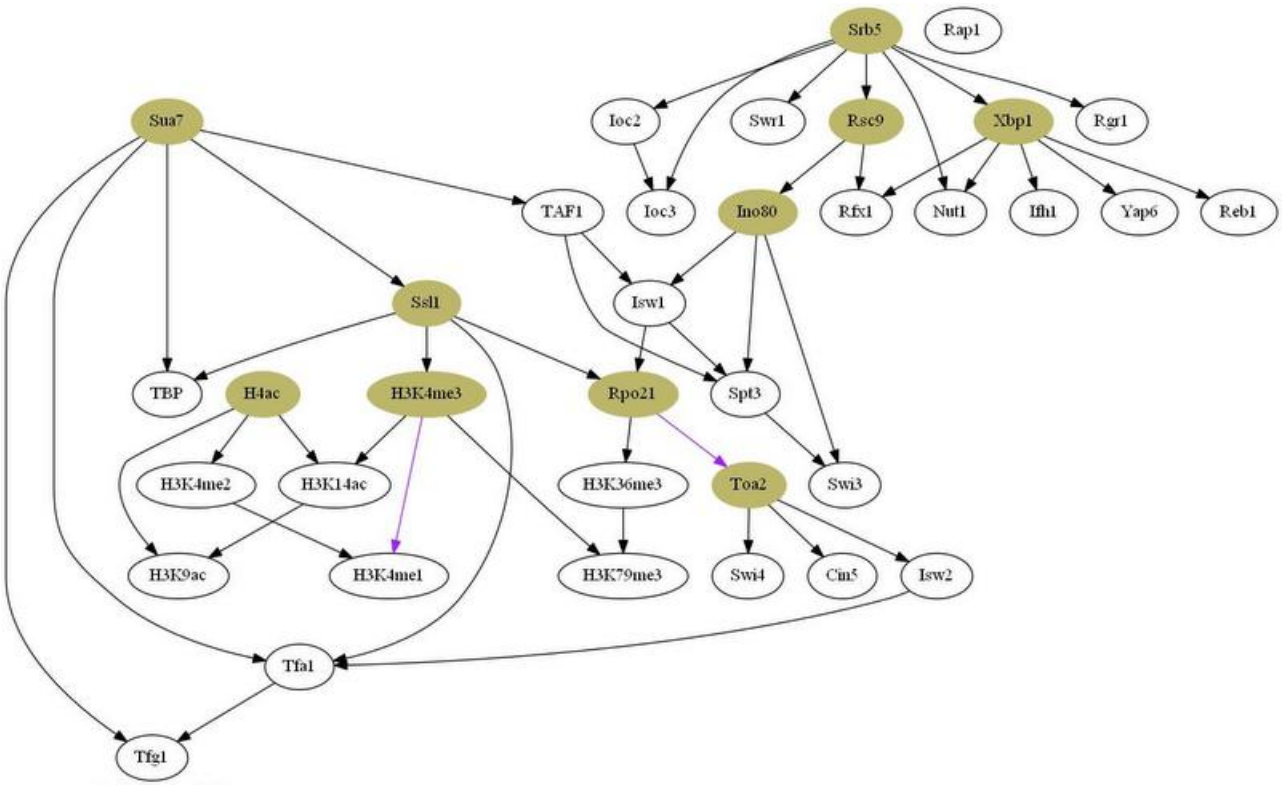

Fig. 2. Bayesian network model of transcriptional relationships. 10 most dominant factors are illustrated by filled nodes, and purple edges represent negative influence between two respective factors.

least one direct functional linkage to the others in the same group, e.g., $\{Ino80 \rightarrow Isw1, Ino80 \rightarrow Spt3, Ino80 \rightarrow Swi3, Isw1 \rightarrow Spt3, Spt3 \rightarrow Swi3\}$, $\{Sua7 \rightarrow Ssl1, Ssl1 \rightarrow TBP, Sua7 \rightarrow TBP\}$, $\{Rpo21 \rightarrow Toa2\}$, $\{Srb5 \rightarrow Nut1, Srb5 \rightarrow Rgr1\}$. In each of the other two groups, we found only one direct functional linkage, e.g., $\{Ioc2 \rightarrow Ioc3\}$, $\{Tfa1 \rightarrow Tfg1\}$. In comparison with the reported co-occupancy and location-linkage between TFs and CRCs, we found only one strong direct functional linkage $Rsc9 \rightarrow Rfx1$ (with confidence score of 0.969), suggesting that the remaining ones might be the consequence of indirect interactions between those factors with linked proteins (possibly the Mediator).

Among 8 histone modifications in the model, two had direct relationships with transcription machinery, H3K4Me3 and H3K36Me3. H3K4Me3 is a modification related to active chromatin in many eukaryotic organisms [21], [9]. In yeast, tri-methylation of H3K4 is catalyzed by Set1 methylase, which has been recruited to active regions in chromatin by TFIIH-associated kinase Kin28, a substance phosphorylating the PolII C-terminal domain (CTD) and mediating the transition between initiation and elongation [36]. These relationships were confirmed by our model by the links $TFIIH(Ssl1) \rightarrow H3K4Me3$ and $TFIIH(Ssl1) \rightarrow PolII(Rpo21)$. H3K36Me3 is reported to coincidently appear at MET16 promoter with H3K4Me3 after induction but change quickly after the onset of transcription [32]. Consistent with this observation, the link $PolII(Rpo21) \rightarrow H3K36Me3$ suggests that tri-methylation of H3K36 may be controlled by different mechanism that is also linked with the recruitment and elongation of Pol II.

During transcription, there may be crosstalks among histone modifications [37]. For example, the crosstalk between H3K4Me3 and H3K14Ac in yeast is known to create positive feedback loops in which H3K4Me3 and H3K14Ac may reinforce each other [35]. However, because BN model is limited to the DAG class, our network only presents the link $H3K4Me3 \rightarrow H3K14Ac$. Also, the link $H3K14Ac \rightarrow H3K9Ac$ suggests that H3K14 acetylation by Gcn5 acetyltransferase may cause the acetylation of neibouring domain (H3K9), consistent with the role of Gcn5 in catalyzing these two acetylation events [38].

3.4 Pivotal roles of Mediator, GTFs, and CRCs in trancsriptional regulation

We further investigated biological characteristics of the resulting model by identifying its *dominant factors* and their related relationships following the method in [17]. By which, dominance score of a factor (node) X is calculated as $dScore(X) = \sum C_0(X,Y)^k$, where $C_0(X,Y)$ denotes the confidence score of X being an ancestor of Y, k is the constant to reward high confidence features. Table 3 shows 10 most dominant factors with corresponding dominance scores given by our method ($k \geq 2$. With $k = 1$ only their orders were changed). It shows consistency with previous reports about the important roles of the Mediator (Srb5), GTFs (Sua7, Ssl1), and CRCs (Rsc9, Ino80) in transcription regulation [4], [34], [39], [40].

The network shows that Mediator may impose its effect on transcription through either direct or indirect interactions with remodeling complexes, e.g. RSC (Rsc9), INO80 (Ino80), and ISW1 (Isw1), suggesting the critical roles of these complexes in the process. So far, CRCs are known to translocate or even remove nucleosomes from DNA sequences, by which facilitating the recruitment of Pol II and transcription machinery to the promoter [2], [3]. Their importance in transcription is demonstrated by the fact that some activators are dispensable for maintainance of transcription when nucleosomes are unable to reassemble at a gene promoter [41], [42]. Recent observations on chromatin remodeling activity induced by heat shock at several genes lead to the hypothesis that, instead of acting individually several CRCs may function cooperatively [43], [44], [33]. Taken together with the result from [4], our model provides support for this hypothesis and further suggests that the cooperation of CRCs may be a critical step in genome-wide transcriptional regulation.

One advantage of BN model is that it permits the identification of a group of variables (parents) that combinatorially regulate another one (the common child). In our model, there are two groups of factors cooperatively regulating the recruitment and activity of Pol II (Rpo21), one is GTFs ($Sua7 \rightarrow Ssl \rightarrow Rpo21$), and the other is CRCs ($Srb5 \rightarrow Rsc9 \rightarrow Ino80 \rightarrow Isw1 \rightarrow Rpo21$). Although the importance of GTFs and CRCs in transcriptional regulation have been reported in literature [4], [3], several observations suggested that there may be redundancy in the functions of individual components. For example, the activation of several genes may not need the presence of such critical factors as TFIIA and CTD of Pol II [45], [46]. Our model confirms that, even such redundancy exists, the activities of both GTFs and CRCs are important to transcription, from initiation stage to the onset of elongation. This conclusion is supported by the observation that, even partial PIC had been formed at some genes, PIC is only fully assembled when the -1 nucleosome is removed [47], [4].

4 Conclusion

Transcription in eukaryotic organisms is a complex process requiring the involvement of a large number of proteins, e.g., GTFs, TFs, chromatin remodelers and

modifiers. However, their detailed roles as well as how they function together to regulate transcription are still unclear. Using genome-wide mapping data of the transcription machinery and histone modifications from *S.cerevisiae*, we proposed a computational method to reconstruct a Bayesian network model representing functional relationships among various transcriptional components. Our network model showed high consistency with previous knowledge about their interactions during transcription. A number of novel functional relationships was also suggested, which may bring insights into transcriptional regulation.

ACKNOWLEDGMENTS

We would like to gratefully thank Mr. Hatermink and his colleagues for kindly sharing Banjo codes and documents. Also, we would like to thank anonymous reviewers for their helpful comments that suggested significant improvement on the manuscript. The first and the third authors have been supported by Japanese Government Scholarship (Monbukagakusho) to study in Japan.

REFERENCES

[1] K. Luger, A. W. Mader, A. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 a resolution," *Nature*, vol. 389, pp. 251–260, 1997.

[2] B. Li, M. Carey, and J. L. Workman, "The role of chromatin during transcription," *Cell*, vol. 128, no. 4, pp. 707–719, 2007.

[3] S. Henikoff, "Nucleosome destabilization in the epigenetic regulation of gene expression," *Nature Reviews Genetics*, vol. 9, no. 1, pp. 15–26, 2008.

[4] B. J. Venters and B. F. Pugh, "A canonical promoter organization of the transcription machinery and its regulators in the saccharomyces genome," *Genome Res.*, vol. 19, no. 3, pp. 360–71, 2009.

[5] I. Albert, T. N. Mavrich, L. P. Tomsho, J. Qi, S. J. Zanton, S. C. Schuster, and B. F. Pugh, "Translational and rotational settings of h2a.z nucleosomes across the saccharomyces cerevisiae genome," *Nature*, vol. 446, no. 7135, pp. 572–576, 2007.

[6] T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, IstvanAlbert, and B. F. Pugh, "Nucleosome organization in the drosophila genome," *Nature*, vol. 453, pp. 358–362, 2008.

[7] C. T. Harbison, D. B. Gordon, T. I. Lee, N. J. Rinaldi, K. D. Macisaac, T. W. Danford, N. M. Hannett, J. B. Tagne, D. B. Reynolds, J. Yoo, E. G. Jennings, J. Zeitlinger, D. K. Pokholok, M. Kellis, P. A. Rolfe, K. T. Takusagawa, E. S. Lander, D. K. Gifford, E. Fraenkel, and R. A. Young, "Transcriptional regulatory code of a eukaryotic genome," *Nature*, vol. 431, no. 7004, pp. 99–104, 2004.

[8] R. Jothi, S. Cuddapah, A. Barski, K. Cui, and K. Zhao, "Genome-wide identification of in vivo protein-dna binding sites from chip-seq data," *Nucleic Acids Res.*, vol. 36, no. 16, pp. 5221–31, 2008.

[9] 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, no. 4, pp. 823–837, 2007.

[10] 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, no. 7, pp. 897–903, 2008.

[11] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, and B. van Steensel, "Systematic protein location mapping reveals five principal chromatin types in drosophila cells," *Cell*, vol. 143, no. 2, pp. 212–24, 2010.

[12] Y. Wang, X. S. Zhang, and Y. Xia, "Predicting eukaryotic transcriptional cooperativity by bayesian network integration of genome-wide data," *Nucleic Acids Res.*, vol. 37, no. 18, pp. 5943–58, 2009.

[13] Z. Dai, X. Dai, Q. Xiang, J. Feng, J. Wang, Y. Deng, and C. He, "Genome-wide analysis of interactions between atp-dependent chromatin remodeling and histone modifications," *BMC Genomics*, vol. 10, no. 304, 2009.

[14] B. van Steensel, U. Braunschweig, G. J. Filion, M. Chen, J. G. van Bemmel, and T. Ideker, "Bayesian network analysis of targeting interactions in chromatin," *Genome Res.*, vol. 20, no. 2, pp. 190–200, 2009.

[15] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. di Bernardo, "How to infer gene networks from expression profiles," *Mol Syst Biol.*, vol. 3, no. 78, 2007.

[16] N. T. Le and T. B. Ho, "Reconstruction of histone modification network from next-generation sequencing data," *Bioinformatics and Bioengineering (BIBE)*, Oct 2011.

[17] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using bayesian networks to analyze expression data," *J Comput Biol.*, vol. 7, no. 3-4, pp. 601–20, 2000.

[18] R. Jansen, H. Yu, D. Greenbaum, Y. Kluger, N. J. Krogan, S. Chung, A. Emili, M. Snyder, J. F. Greenblatt, and M. Gerstein, "A bayesian networks approach for predicting protein-protein interactions from genomic data," *Science*, vol. 302, no. 6544, pp. 449–53, 2003.

[19] K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan, "Causal protein-signaling networks derived from multiparameter single-cell data," *Science*, vol. 308, no. 5127, pp. 523–9, 2005.

[20] 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, no. 8, pp. 1314–1324, 2008.

[21] 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, and R. A. Young, "Genome-wide map of nucleosome acetylation and methylation in yeast," *Cell*, vol. 122, no. 4, pp. 517–527, 2005.

[22] W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow, "A high-resolution atlas of nucleosome occupancy in yeast," *Nature Genetics*, vol. 39, no. 10, pp. 1235–1244, 2007.

[23] D. Karolchik, A. S. Hinrichs, T. S. Furey, K. M. Roskin, C. W. Sugnet, D. Haussler, and W. J. Kent, "The ucsc table browser data retrieval tool," *Nucleic Acids Research*, vol. 32, pp. D493–D496, 2004.

[24] D. Heckerman, D. Geiger, and D. M. Chickering, "Learning bayesian networks: The combination of knowledge and statistical data," *Machine Learning*, vol. 20, pp. 197–243, 1995.

[25] F. V. Jensen and T. D. Nielsen, *Bayesian Networks and Decision Graphs*, 2nd ed. New York: Springer-Verlag, 2001.

[26] D. M. Chickering, "Learning bayesian networks is np-hard," *Technical report*, 1994.

[27] G. Cooper and E. Herskovits, "A bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, pp. 309–47, 1992.

[28] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young, "Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks," *Pac Symp Biocomput.*, pp. 422–33, 2001.

[29] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, pp. 3594–603, 2004.

[30] V. A. Smith, E. D. Jarvis, and A. J. Hartemink, "Evaluating functional network inference using simulations of complex biological systems," *Bioinformatics*, vol. 18, no. Suppl 1, pp. S216–24, 2002.

[31] A. Jansen and K. J. Verstrepen, "Nucleosome positioning in saccharomyces cerevisiae," *Microbiol Mol Biol Rev.*, vol. 75, no. 2, pp. 301–320, 2011.

[32] A. Morillon, N. Karabetsou, A. Nair, and J. Mellor, "Dynamic lysine methylation on histone h3 defines the regulatory phase of gene transcription," *Mol Cell*, vol. 18, no. 6, pp. 723–734, 2005.
[33] T. Y. Erkina, Y. Zou, S. Freeling, V. I. Vorobyev, and A. M. Erkine, "Functional interplay between chromatin remodeling complexes rsc, swi/snf and iswi in regulation of yeast heat shock genes," *Nucleic Acids Res.*, vol. 38, no. 5, pp. 251–260, 2009.
[34] B. Li, M. Carey, and J. L. Workman, "The Role of Chromatin during Transcription," *Cell*, vol. 128, no. 4, pp. 707–719, 2007.
[35] B. J. Venters and B. F. Pugh, "How eukaryotic genes are transcribed," *Crit Rev Biochem Mol Biol.*, vol. 44, no. 2-3, pp. 117–41, 2009.
[36] H. H. Ng, F. Robert, R. A. Young, and K. Struhl, "Targeted recruitment of set1 histone methylase by elongating pol ii provides a localized mark and memory of recent transcriptional activity," *Mol Cell*, vol. 11, no. 3, pp. 709–719, 2003.
[37] K. Luger, A. W. Mader, A. K. Richmond, D. F. Sargent, and T. J. Richmond, "The language of covalent histone modifications," *Nature*, vol. 403, no. 6765, pp. 41–5, 2000.
[38] W. Zhang, J. R. Bone, D. G. Edmondson, B. M. Turner, and S. Y. Roth, "Essential and redundant functions of histone acetylation revealed by mutation of target lysines and loss of the gcn5p acetyltransferase," *EMBO J.*, vol. 17, no. 11, pp. 3155–67, 1998.
[39] R. Biddick and E. T. Young, "Yeast mediator and its role in transcriptional regulation," *C R Biol.*, vol. 328, no. 9, pp. 773–82, 2005.
[40] R. D. Kornberg, "Mediator and the mechanism of transcriptional activation," *Trends Biochem Sci.*, vol. 30, no. 5, pp. 235–9, 2005.
[41] M. W. Adkins and J. K. Tyler, "Transcriptional activators are dispensable for transcription in the absence of spt6-mediated chromatin reassembly of promoter regions," *Mol Cell*, vol. 21, no. 3, pp. 405–16, 2006.
[42] H. Zhang and J. C. Reese, "Exposing the core promoter is sufficient to activate transcription and alter coactivator requirement at rnr3," *Proc Natl Acad Sci USA*, vol. 104, no. 21, pp. 8833–8, 2007.
[43] B. Xella, C. Goding, E. Agricola, E. di Mauro, and M. Caserta, "The iswi and chd1 chromatin remodelling activities influence adh2 expression and chromatin organization," *Mol Microbiol.*, vol. 59, no. 5, pp. 1531–41, 2006.
[44] K. C. Lindstrom, J. C. Vary, M. R. Parthun, J. Delrow, and T. Tsukiyama, "Isw1 functions in parallel with the nua4 and swr1 complexes in stress-induced gene repression," *Mol Cell Biol.*, vol. 26, no. 16, pp. 6117–29, 2006.
[45] S. Chou, S. Chatterjee, M. Lee, and K. Struhl, "Transcriptional activation in yeast cells lacking transcription factor iia," *Genetics*, vol. 153, no. 4, pp. 1573–81, 1999.
[46] S. W. Hong, S. M. Hong, J. M. Yoo, Y. C. L. Y. S. Kim, J. T. Lis, and D. K. Lee, "Phosphorylation of the rna polymerase ii c-terminal domain by tfiih kinase is not essential for transcription of saccharomyces cerevisiae genome," *Proc Natl Acad Sci USA*, vol. 106, no. 34, pp. 14 276–80, 2009.
[47] S. J. Zanton and B. F. Pugh, "Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock," *Genes Dev.*, vol. 20, no. 16, pp. 2250–65, 2006.

Ngoc Tu Le received his MSc degree in Knowledge Science from Japan Advanced Institute of Science and Technology (JAIST), Japan. He has been a PhD candidate in JAIST from 2010. His research interests include data mining and machine learning techniques, and their applications to elucidating epigenetic gene regulation.

Tu Bao Ho is currently a professor of School of Knowledge Science, Japan Advanced Institute of Science and Technology. He received a BT in applied mathematics from Hanoi University of Technology (1978), MS and PhD in computer science from Pierre and Marie Curie University, Paris (1984, 1987). His research interests include knowledge-based systems, machine learning, knowledge discovery and data mining, and computational biomedicine.

Bich Hai Ho got her MSc degree in Computing science in University of East Anglia, UK. She has been a PhD candidate in JAIST from 2008. Her research interests include applications of data mining techniques to understanding nucleosome dynamics.