Reconstruction of histone modification network from next-generation sequencing data

Ngoc Tu Le
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
email: ngoctule@jaist.ac.jp

Tu Bao Ho
Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
email: bao@jaist.ac.jp

***Abstract*—Post-translational modifications (PTMs) of histone proteins play critical roles in establishing functionally separated domains on chromatin and regulating important biological processes, such as transcription. These modifications often act in cooperative manner, forming complicated "histone codes". Elucidation of functional relationships among them will, therefore, significantly increase our understanding of cell differentiation, development, and cancer pathogenesis. Biological evidence has shown that nucleosome positioning can provide invaluable information about interactive effects of PTMs. However, to our knowledge, none of previous works has exploited this information in the reconstruction of histone modification networks.**

We propose a computational approach based on Bayesian network to reconstruct a network representing functional relationships of histone modifications. Our approach employed the search-and-score method to infer the network structure using interactive information of histone modifications, which is measured by the correlation between each modification with nucleosome positioning. When applied on human CD4+ T cell ChIP-Seq dataset, containing 38 different histone modifications and binding information of three other proteins, H2A.Z, PolII and CTCF, our method not only outperformed previous approaches in recovering known relationships but also suggested many new ones, confirming its validity and efficiency. Our unbiased method for inferring the network structure can also be applied to reconstruct interaction networks of other epigenetic factors.

***Keywords*-histone modification; Bayesian network; ChIP-Seq data; histone code**

I. Introduction

Eukaryotic genomes are packaged into chromatin, a highly condensed structure with fundamental repeating units, the nucleosomes. Each nucleosome is formed by wrapping 147bp of DNA around a histone core, an octamer of proteins that contains a central $(H3-H4)_2$ tetramer flanked on both side by two $H2A - H2B$ dimers [1]. The histone core is subject to various covalent modifications occurring mostly on its N-terminal residues, such as acetylation, methylation, phosphorylation and ubiquitination. Biological evidence has shown that histone modifications play important roles in many cellular processes, such as transcription, replication and DNA repair [2]. They are also implicated in the cell fate determination and tumorigenesis [3], [4]. It is observed that different combinations of histone modifications can result in distinct downstream events; and that they together help to stabilize chromatin states and properly propagate such states in cell division by forming broad domains on chromatin [5], [6], led to the hypothesis of "histone codes" [7]. One of such observations was reported by Wang *et al.*, where the authors discovered a "backbone" of 17 modifications associated with 3286 human promoters [8]. Therefore, elucidation of genome-wide histone modification patterns and their functional implications will significantly increase our knowledge of cell development process.

Technological advances for studying histone modifications on a genomic scale, such as 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 generate genome-wide maps of various histone modifications [8], [9]. This leads to the development of a plethora of computational methods to analyze such data, ranging from methods for discovering concentrated combinatorial modification patterns [10], [11] to the ones for identifying large, dispersed epigenomic domains [12], [13]. For example, an unsupervised method proposed by Hon *et al.*, named ChromaSig, can identify genome-wide histone modification "motifs" and use these "motifs" to characterize novel functional genomic elements [10].

Among the number of approaches for analyzing PTM data, Bayesian network-based methods offer a promising way to identify not only co-occurence patterns but also the (in)dependence relationships of different histone modifications. Bayesian network (BN) is a class of probabilistic graphical models, which has been widely applied to reconstruct many kinds of cellular networks, such as gene regulatory networks and protein interaction networks [14], [15]. However, its application in analyzing histone modification data is still limited. The first attempt applying BN to reveal causal relationships of histone modifications was proposed by Yu *et al.* [16]. Although shown to be useful, their approach suffers from an important drawback. The algorithm employed in their work to identify compelled edges, which may represent causal relationships, requires the data satisfying Causal Sufficiency assumption. In other words, it assumes that no hidden confounder should exist in the data [17], [18]. This assumption, unfortunately, is not guaranteed in the context of histone modification data because there are many other PTMs along with their biological functions yet to be known [5]. Lv *et al.*, using the same framework, proposed an alternative way to reconstruct histone modification networks by inferring a BN on the Closeness Measure (CM), which was proposed to capture interactive information of histone modifications by their correlations with DNA methylation data [19]. Their model, however, is difficult to interpret from biological perspective.

Our approach is based on the observation that interactions of histone modifications can cause change in nucleosome distribution and, through this change, exert their regulatory effects on cellular processes. For example, histone modifying enzymes are known to affect histone-DNA interactions and regulate nucleosome stability thereof [20], making the sites occupied by nucleosomes more accessible to cellular machineries, as suggested in the "regulated nucleosome mobility" model [21]. The aim of our work, therefore, is twofold: first, to describe a method to capture interactive information of histone modifications from nucleosome positioning data; second, to propose an unbiased method to infer a BN representing (in)dependence relationships of histone modifications. When applied on human CD4+ T cell data, containing 38 histone modifications and binding information of three proteins, H2A.Z, PolII, CTCF, our method outperformed previous approaches in recovering known relationships. It also revealed a number of unreported dependencies among those modifications, which may suggest novel functional insights into the regulatory mechanism by PTMs.

II. Materials and Methods

A. Data Preparation

- *Chromatin modification.* Experimental ChIP-Seq data of 20 histone methylations, 18 acetylations and 3 other proteins, H2A.Z, PolII and CTCF, in human CD4+ T cell were obtained from [8], [9].
- *Nucleosome positioning.* Nucleosome positioning data in resting human CD4+ T cell were obtained from [22]
- *The gene set.* Gene expression data for resting human CD4+ T cell were obtained from [22]. UCSC Known Genes were extracted and then mapped to Affymetrix U133P2 probe IDs using the tables provided in the UCSC Genome Browser [23]. Genes without corresponding U133P2 IDs were removed. If multiple genes map to the same U133P2 ID, only one was retained. We also removed genes from chromosomal regions marked with "random" or genes from haplotype regions.

B. Interactive information of PTMs

In order to derive interactive information of histone modifications, we firstly created tag profiles in promoter regions ($TSS \pm 1kbp$) for each modification and nucleosome positioning. At first, each region was divided into non-overlapping 200bp bins. Then, sequence tag locations were shifted +65bp for hits on the positive strand and -65bp for hits on the negative strand. The bin for each tag was determined by the middle of the tag. Each tag profile can then be represented as a 10-dimensional vector $TagProf = (bin_1, \ldots, bin_{10})$, where bin_i $(i = 1, \ldots, 10)$ is the logarithm of the number of tags belonging to the bin i^{th}. If a bin has no tag, it was assigned the value of 0. Interactive information of each modification at a promoter region was measured as $InterInfo = Correlation(ModTagProf, NucTagProf)$, where $ModTagProf$ and $NucTagProf$ are the tag profiles of the modification and nucleosome positioning at that region, respectively. In our work, we employed non-parametric Spearman's rank correlation since it has no prior assumption on data distribution.

C. Bayesian Networks

1) Definition: A Bayesian network for a set of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ is a probabilistic model consisting of two components [36], [37]:

- A network structure S, which is a directed acyclic graph, representing conditional (in)dependence relationships among variables in **X**
- A set P of local probability distributions associated with each variable.

Markov and Faithfulness conditions guarantee that these two components, (S, P), encode a joint probability distribution on **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. The number of parents of each variable is usually small so a BN provides a compact and convenient way to represent a joint probability distribution. In our work, we learnt a BN on discrete variables, which means that local probability distributions P can be represented by Conditional Probability Tables (CPTs). Such a table specifies the probability a variable takes a certain value given the values of its parents.

2) Learning BN structure: As mentioned above, a BN contains two components. Consequently, there are two steps in learning a BN model from data: parameter learning, which learns the local probability distributions P, and structure learning, which learns the structure S. Because the main target of our work is to uncover the (in)dependence relationships among the PTMs so we are interested in the latter learning problem.

We employed the score-based search method to learn a BN structure representing (in)dependence relationships among PTMs. 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 (a graph without any edges), 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 [34], a greedy hill-climbing search was used as our search strategy. To escape from local maximum, a simulated annealing approach was used.

To score a candidate network, we used a Bayesian scoring metric, which was originated from [35], and further developed by [36] as BDe (Bayesian metric with Dirichlet prior and equivalence) 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 meanings but correspond to prior knowledge for the parameters. When no prior knowledge is available, they can be estimated as $N_{ijk}^{'} = N/(r_i qi$ 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.

D. Bootstrapping and Selection of The Cut-off Threshold

As the search-and-score method may output a different network on each run (in our work, we only selected one graph with the highest score as the output), we employed the bootstrapping method, proposed by Friedman *et al.*[14], to estimate the confidence level for each edge in the resulting network. Given a dataset D of N instances, we created a new dataset $D^{'}$ by resampling from D with replacement N times. Then a BN was inferred on $D^{'}$. These two steps of resampling and inferring a BN were repeated m times, generating m different BNs. The confidence level of each edge was estimated as the proportion of networks containing that edge. A threshold, named τ, was chosen to decide whether an edge was included in the resulting network or not.

Because there is no positive training data about the relationships among histone modifications, we derived the following method to select a reasonable value for τ. At first, we randomly split the input dataset D into two equal parts, D_1 and D_2, T times. At each time, two bootstrapped BNs corresponding D_1 and D_2, namely $partialBN_1$ and $partialBN_2$, were learned as above-described. Then, we defined a measure, named acc_i, as:

$$acc_i = \frac{\sharp(partialBN_1 \cap partialBN_2)}{\sharp(partialBN_i)}, i = \{1, 2\}$$

where the denominator is the number of edges of $partialBN_i$ and the numerator is the number of edges that appear in both $partialBN_1$ and $partialBN_2$.
The selection criteria was then chosen as:

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

where $Var(acc_i)$ is the variance of acc_i after T times of data splitting and network learning steps above. It is easy to see that acc_i ($i = \{1, 2\}$), and therefore $Stability$ are the functions of τ. We then chose τ that gives $Stability$ the smallest value as the cut-off threshold.

III. Results

A. Interactive Information of Histone Modifications Reflect Their Regulatory Effects

Interactive information of different histone modifications corresponding to the gene set was calculated as described in *Section 2.2.* After this step, we received a table containing 10572 rows and 41 columns, where each row corresponds to one gene, each column corresponds to one of 41 features (38 histone modifications and 3 proteins), and each element corresponds to the interactive information of the feature at the promoter region of the gene. Histograms of interactive information (Figure 1) showed that, modifications associated with transcription activation (such as H3K4Me1/3) and elongation (H3K36Me3), play more important role in shaping nucleosome organization at promoters than repressive (H3K9Me3, H4K20Me3) and less-well-understood PTMs (H3K79Me1/2). This is consistent with the hypothesis that, in order to activate the genes nucleosomes at promoters, considered as the barrier to transcriptional machinery, should be destabilized or even evicted from original locations and this can be accomplished with the help of multitude of modifications on nucleosomes. Notably, the histogram of H2A.Z showed that it does not have much effect on nucleosome organization despite being a well-known euchromatin mark. This is consistent with what was reported in [24].

B. Inference of Histone Modification Network

- *Data Preprocessing.* The contingency table described in previous section was used as the input for the network inference algorithm. Using interactive information provides a natural way for data scaling because all values will be in the range of $[-1, 1]$. Our network model only accepts discrete variables so the next step is to discretize data into discrete values. In our work, each feature was discretized into 3 categories using interval discretization scheme. We chose the region of $\pm 1kbp$ around the TSS as well as the 3-category

Table I
CRITERIA FOR THRESHOLD SELECTION

τ	*Stability*
0.5	0.0135
0.55	0.0101
0.6	**0.0077**
0.65	0.0163
0.7	0.0197
0.75	0.0466
0.8	0.0245
0.85	0.049
0.9	0.0313

discretization scheme in our analysis because it has been shown elsewhere [16] that these choices could give a reasonable result on the data set.

- *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 *Section 2.3*. Empirical running showed that, with more than $1,500,000$ search iterations the network score was not significantly improved, so each search was limited to this number of iterations.
- *Threshold Derivation.* The procedure described in *Section 2.4* was employed to derive a reasonable threshold for confidence level related with each inferred relation (an edge of the resulting BN in this case). We randomly split input data into two equal datasets 10 times ($T = 10$), resulting in 20 datasets. For each dataset, we ran bootstrap procedure on it 100 times ($N = 100$) and derived a corresponding consensus BN. Each edge in the consensus BN has a related confidence score, measured by the number of times it appears in 100 bootstrapped BNs. Threshold τ was chosen in the range of $[0.5; 0.9]$ with the step of 0.05. Table I shows the values of the selection criterion $Stability$ with corresponding values of τ. The value of τ that gives $Stability$ the smallest value ($\tau = 0.6$, $Stability = 0.0077$) was chosen as the cut-off threshold. Finally, all input data was used to infer the structure of histone modification network. Bootstrap procedure was run 1000 times to identify confidence scores for each edge of the resulting network. After setting the threshold $\tau = 0.6$, we received a BN containing 50 edges (Figure 2), representing functional relationships among various histone modifications.

C. Network Analysis for Discovering Crosstalks among Histone Modifications

Although the modification state of each gene is likely to appear differently in particular cell type or condition, the dependencies found by our network model might reflect functional relationships among various histone modifications, which in general could be the same in different cell types or under different conditions.

The resulting network contained three root nodes (node without incoming edge), two are active modifications (H3K18Ac, H3K27Ac), and one is elongation-related mark (H3K79Me1). Among them, H3K18Ac and H3K27Ac are likely to play a central role, having out degrees (number of edges pointing away from the node) of 4 and 8 (highest allover the network), correspondingly. It included several chains of active modifications, such as $H3K4Ac \rightarrow H4K91Ac \rightarrow H4K16Ac$, $H2BK120Ac \rightarrow H2BK20Ac \rightarrow H4K5Ac \rightarrow H4K8Ac \rightarrow H4K12Ac$ downstream of H3K18Ac, and $H3K4Me2 \rightarrow H3K9Me1 \rightarrow H3K27Me1$, $H3K9Ac \rightarrow H3K4Me3 \rightarrow PolII$ downstream of H3K27Ac. A recent work has also reported about the essential role of H3K18/K27Ac in ligand-induced PolII recruitment on, and activation of, nuclear receptor target genes [31]. Five modifications, H3K79Me1/2/3, H2BK5Me1, H3R2Me1, which were reported having quite similar diffuse profiles [13], were found closely associated in our model by two chains, $H3K79Me1 \rightarrow H3K79Me2 \rightarrow H3K79Me3$ and $H3K79Me1 \rightarrow H4K20Me1 \rightarrow H2BK5Me1 \rightarrow H3R2Me1$, suggesting that our method could identify the relationships of not only concentrated modifications but also dispersed epigenomic domains. The chain of H3K79Me1, a less-well-understood modification, and H3K79Me2, an elongation mark, and H3K79Me3, a repressive mark in human, suggests a directional equilibrium among these modifications.

Another important node, even not present at the root level, is histone variant H2A.Z, which also had out degree of 4. Our model inferred that H2A.Z is synergistically influenced by H3K18/27Ac, and itself influences H3K4Me3. It is known that H2A.Z is an important component of euchromatin, whose function is to antagonize the repressive chromatin state. How it is deposited to specific sites and whether this process happens randomly or not, however, remain elusive. Raisner *et al.* [29] have shown that preventing acetylation by mutating specific lysine residues of histone H3 and H4 in yeast would cause defect in H2A.Z enrichment at several loci. Consistent with this, our model suggested that acetylation of H3K18/27 may play a critical role in regulating the deposition of H2A.Z onto chromatin. Moreover, as previously reported [25], cells lacking H2A.Z also had a reduced H3K4Me3 level at many promoters, in other words, H3K4Me3 depended on H2A.Z for its enrichment at promoters. Our network model confirmed this dependency by the link $H2A.Z \rightarrow H3K4Me3$, showing its advantage over previous models, which suggested this link in reverse direction, i.e. $H3K4Me3 \rightarrow H2A.Z$ [16], [19].

From inferred relationships, we found that H3K4Me3 may be influenced not only by H2A.Z but also by H3K27Ac, either directly or indirectly (through H3K9Ac). Tie *et al.* [26] have reported about the correlation between H3K27Ac and H3K4Me3 by observing their similar profiles at most of investigated sites. Their observation was supported by the fact that, Trithorax protein (TRX) and histone acetyltransferase CBP together acetylates H3K27, and TRX itself is a histone methyltransferase that specifically trimethylates histone H3 on lysine 4 (H3K4Me3) [27]. Another work has also shown that the abundance and "degree" of H3K4 methylation were dependent on histone H3 acetylation [28]. These evidences provide support for the links between H3K27Ac and H3K4Me3.

The resulting model also agreed with previous ones in some confirmed relationships, such as $H3K4Me3 \rightarrow PolII$ and $H3K79Me2 \rightarrow H4K20Me1$ [16], [19]. The former was supported by the fact that PolII binding is affected by the protein trxG, which catalyzes H3K4Me3 [32], [33]. The latter was supported by a research on mouse embryonic stem (ES) cell, which showed that deficiency of Dot1L, a H3K79 methyltransferase, would cause reduced levels of H3K9Me and H4K20Me [30].

In addition to the relationships that were already confirmed in literature, many new, unconfirmed relationships among histone modifications were suggested by our model, such as H4K20Me1 and H2BK5Me1, H3K4Me2 and H2A.Z, H3K4Me1 and H2BK5Me1, PolII and CTCF, which were also reported in previous works [16], [19]. Taken together, it showed some overlap with, but was not identical to, currently existing models. Moreover, it correctly recovered the relationships representing important crosstalks among various histone modifications and other chromatin binding factors. This result confirms the validity and efficiency of our network model.

IV. Conclusion

Chromatin is a highly compact structure for organizing genomic material inside the cell nucleus. However, it is not a passive entity but plays important roles in many DNA-mediated processes. The histone proteins of this structure are subject to numerous chemical modifications, which may act independently or in concert to contribute to regulatory functionality of chromatin. Elucidation of functional relationships among these modifications will, therefore, significantly improve our understanding of critical cellular processes, such as transcription or pathogenesis.

We have proposed a novel Bayesian Network-based computational approach to reconstructing histone modification network. Our network model was built on interactive information of histone modifications, which was measured by the correlation between histone modification and nucleosome positioning. We also derived an unbiased method for inferring the structure of static Bayesian networks. When applied on human CD4+ T cell Chip-Seq data, our method outperformed previous ones in recovering confirmed relationships. It also suggested many new ones, which could help deepen our understanding of regulatory mechanism by PTMs.

Acknowledgment

The first author is supported by Japanese Government Scholarship (Monbukagakusho) to study in Japan.

References

[1] Luger, K. *et al.*, (1997) Crystal structure of the nucleosome core particle at 2.8 A resolution, *Nature*, **389**, 251-60.

[2] Kouzarides, T., (2007) Chromatin modifications and their function, *Cell*, **128**(4), 693-705.

[3] Fraga, M.F. *et al.*, (2005) Loss of acetylation at Lys16 and trimethylation at Lys20 of histone H4 is a common hallmark of human cancer, *Nature Genet.*, **37**, 391-400.

[4] Mikkelsen, T.S. *et al.*, (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells, *Cell*, **448**(7151), 553-60.

[5] Bernstein, B.E. *et al.*, (2007) The mammalian epigenome, *Cell*, **128**(4), 669-81.

[6] Hawkins, R.D. *et al.*, (2010) Distinct epigenomic landscapes of pluripotent and lineage-committed human cells, *Cell*, **6**(5), 479-91.

[7] Strahl, B.D., and Allis C.D., (2000) The language of covalent histone modifications, *Nature*, **403**(6765), 41-5.

[8] Wang, Z. *et al.*, (2008) Combinatorial patterns of histone acetylations and methylations in the human genome, *Nat Genet.*, **40**(7), 897-903.

[9] Barski, A. *et al.*, (2007) High-resolution profiling of histone methylations in the human genome, *Cell*, **129**(4), 823-37.

[10] Hon, G. *et al.*, (2008) ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome, *PLoS Comput Biol.*, **4**(10), e1000201.

[11] Ucar, D. *et al.*, (2011) Combinatorial chromatin modification patterns in the human genome revealed by subspace clustering, *Nucleic Acids Res.*, **4**(10), e1000201.

[12] Zang, C. *et al.*, (2009) A clustering approach for identification of enriched domains from histone modification ChIP-Seq data, *Bioinformatics*, **25**(15), 1952-8.

[13] Song, Q. *et al.*, (2011) Identifying dispersed epigenomic domains from ChIP-Seq data, *Bioinformatics*, **25**(15), 1952-8.

[14] Friedman, N. *et al.*, (2000) Using Bayesian networks to analyze expression data, *J Comput Biol.*, **7**(3-4), 601-20.

[15] van Steensel, B. *et al.*, (2009) Bayesian network analysis of targeting interactions in chromatin, *Genome Res.*, **20**(2), 190-200.

[16] Yu, H. *et al.*, (2008) Inferring causal relationships among different histone modifications and gene expression, *Genome Res.*, **18**(8), 1314-1324.

[17] Chickering, D.M., (1995) A transformational characterization of equivalent Bayesian network structures, *Proceedings of 11th Conference on Uncertainty in Artificial Intelligence*,87-89, Morgan Kaufmann Publishers.

[18] Zhang, J., and Spirtes, P., (2005) A Transformational Characterization of Markov Equivalence between DAGs with Latent Variables, *Proceedings of 21th Conference on Uncertainty in Artificial Intelligence*,667-674, UAUI Press.

[19] Lv, J. *et al.*, (2010) Discovering Cooperative Relationships of Chromatin Modifications in Human T Cells Based on a Proposed Closeness Measure, *PLoS One*, **5**(12), e14219.

[20] Henikoff, S., (2008) Nucleosome destabilization in the epigenetic regulation of gene expression, *Nat Rev Genet.*, **9**(1), 15-26.

[21] Cosgrove, M.S. *et al.*, (2004) Regulated nucleosome mobility and the histone code, *Nat Struct Mol Biol.*, **11**(11), 1037-1043.

[22] Schones, D.E. *et al.*, (2008) Dynamic regulation of nucleosome positioning in the human genome, *Cell*, **132**(5), 887-98.

[23] Karolchik, D. *et al.*, (2004) The UCSC Table Browser data retrieval tool, *Nucleic Acids Res.*, **32**, D493-6.

[24] Hartley, P.D., and Madhani, H.D. (2009) Mechanisms that specify promoter nucleosome location and identity, *Cell*, **173**(3), 445-58.

[25] Meneghini, M.D. *et al.*, (2003) Conserved histone variant H2A.Z protects euchromatin from the ectopic spread of silent heterochromatin, *Cell*, **112**(5), 725-36.

[26] Tie, F. *et al.*, (2009) CBP-mediated acetylation of histone H3 lysine 27 antagonizes Drosophila Polycomb silencing, *Development*, **136**(18), 3131-41.

[27] Smith, S.T. *et al.*, (2004) Modulation of heat shock gene expression by the TAC1 chromatin-modifying complex, *Nat Cell Biol.*, **6**(2), 162-7.

[28] Nightingale, K.P. *et al.*, (2007) Cross-talk between histone modifications in response to histone deacetylase inhibitors: MLL4 links histone H3 acetylation and histone H3K4 methylation, *J Biol Chem.*, **282**(7), 4408-16.

[29] Raisner, R.M. *et al.*, (2005) Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin, *Cell*, **123**(2), 233-48.

[30] Jones, B. *et al.*, (2008) The histone H3K79 methyltransferase Dot1L is essential for mammalian development and heterochromatin structure, *PLoS Genet.*, **4**(9), e1000190.

[31] Jin, Q. *et al.*, (2011) Distinct roles of GCN5/PCAF-mediated H3K9ac and CBP/p300-mediated H3K18/27ac in nuclear receptor transactivation, *EMBO J.*, **30**(2), 249-62.

[32] Schuettengruber, B. *et al.*, (2007) Genome regulation by polycomb and trithorax proteins, *Cell*, **128**(4), 735-45.

[33] Schwartz, Y.B. and Pirrotta, V. (2007) Polycomb silencing mechanisms and the management of genomic programmes, *Nat Rev Genet.*, **8**(1), 9-22.

[34] Chickering, D.M. *et al.*, (1994) Learning Bayesian Networks is NP−hard, *Technical report*, Redmond, WA: Microsoft Research.

[35] Cooper, G. and Herskovits, E. (1992) A Bayesian method for the induction of probabilistic networks from data, *Machine Learning*, **9**, 309-47.

[36] Heckerman, D. *et al.*, (1995) Learning Bayesian networks: The combination of knowledge and statistical data, *Machine Learning*, **20**, 197-243.

[37] Jensen, F.V. and Nielsen, T.D. (2007) Bayesian Networks and Decision Graphs (second edition), Springer-Verlag, New York.

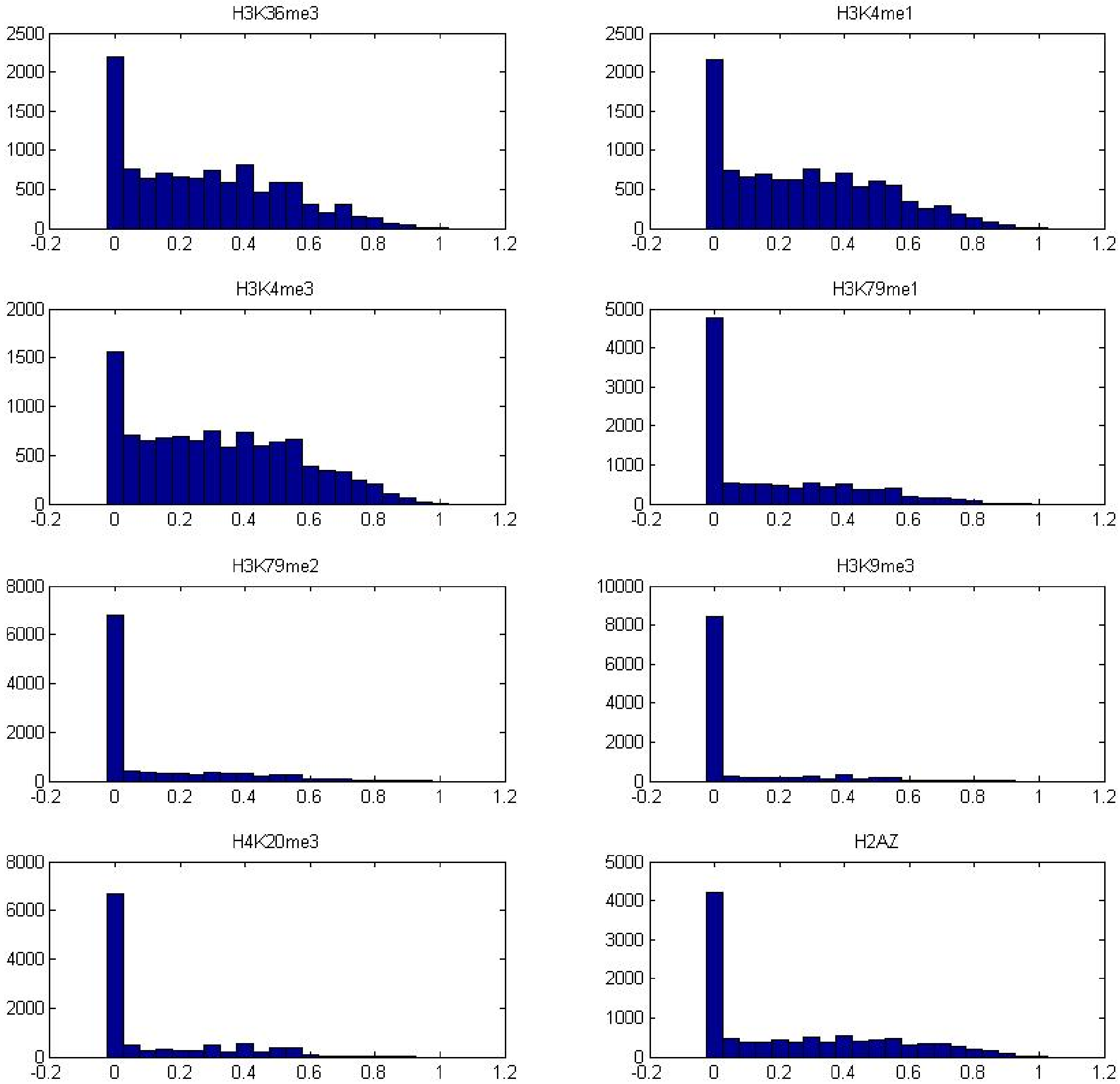

Figure 1. Histograms of interactive information of PTMs with distinct regulatory roles (transcription activation- and elongation-associated marks (H3K4Me1/3, H3K36Me3), repressive marks (H3K9Me3, H4K20Me3), and less-well-understood marks (H3K79Me1/2)) and histone variant H2A.Z. X-axis represents absolute interactive information value, Y-axis represents corresponding frequency

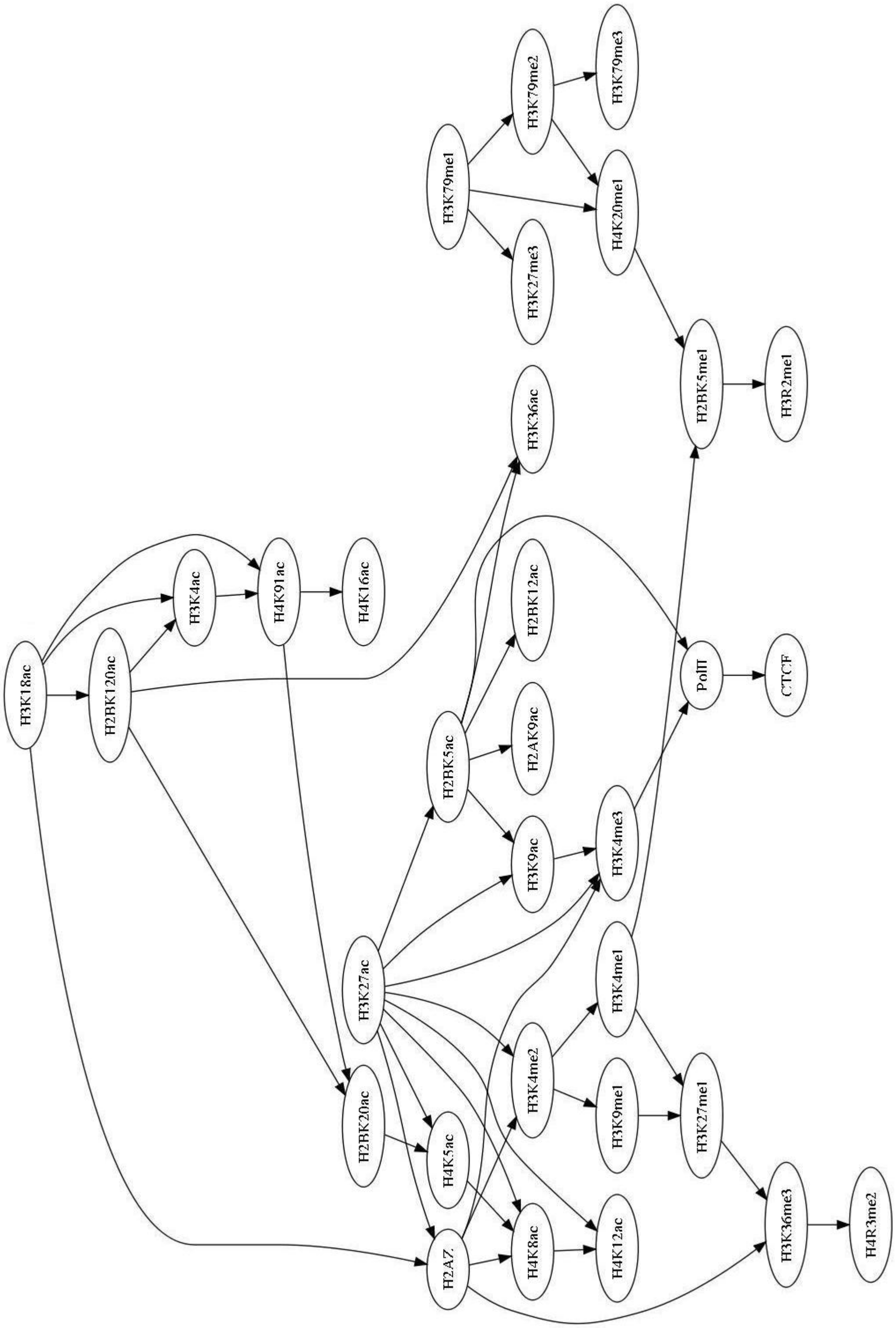

Figure 2. BN model of PTM interactions inferred from interactive information