Bayesian Network (BN) of Targeting Interactions in Chromatin: Reanalyzed

Ricky Lim

FNWI, Address Science Park 904,1098 XH Amsterdam

ABSTRACT

Motivation: Reconstruct the inferred BN of targeting interactions using the dataset from the previously published paper **?** in order to assess and identify its reproducibility.

Results: The previous inferred BN of targeting interactions was successfully reconstructed in which 50 out of 52 previously reported targeting interactions were recovered. The missing interactions were between BRM-Hist3.3 and D1-dm. Furthermore, the reconstructed BN80 also shares the similar network topology with the average of out-degree of 2.5. These results suggest a high level of the reproducibility of the previously published BN80 **?**.

Availability: A package containing codes and files for this study was available as a supplemented zip file, BNI_Reanalyzed.zip

Contact: Ricky.Lim@student.uva.nl

1 INTRODUCTION

Chromatin has been fascinating biologists for decades as it has been thought to provide an ideal platform for packaging of huge eukaryotic genome into a very limited nuclear space. This striking phenomenon suggests several rounds of compaction in order to fit the genome inside the eukaryotic nucleus. The initial round of compaction involves the wrapping of the genome around the core histone proteins forming a structure known as nucleosome. In addition to the core histones, other non-histone proteins, such as remodelers, packagers, histone modifiers, linkers, and many other different types of proteins, have been shown to associate together with nucleosomes. This association pulls nucleosomes closer together into a condensed complex structure known as chromatin, allowing the genome to be packed inside the nucleus, although the details of this process remains largely unknown (reviewed in **?**). Here, these non-histone proteins are referred to as chromatin components.

Besides the reported associations between these chromatin components with the genome, previous mapping studies have shown that they are not spread randomly along the genome but rather occupy discrete locations (reviewed in **?**). These observations raise the question of how they are being targeted specifically to these discrete locations along the genome. According to **?**, this specific targeting is dependent on the interactions with other chromatin components. This motivated their study to explore a network of interactions among chromatin components that may modulate their specific genome-wide binding patterns **?**.

The collection of binding maps of 43 broadly selected chromatin components from previous mapping studies in *Drosophila* offered the possibility to infer a model representing the network of targeting interactions in chromatin**?**. One of a variety of network inference algorithms that has been proposed for this purpose is Bayesian Network Inference (BNI) algorithm. BNI has two essential properties for this task. First is owing to the robustness of BNI to the existence of unobserved variables in a system with the assumption that most variables are (conditionally) independent of each other **?**. Here variables and system refer to chromatin components, and chromatin respectively. This is particularly important as the collected binding maps consist of only 43 chromatin proteins, although there are more chromatin proteins involved in targeting these proteins together in chromatin. With this assumption, BNI can handle explicitly the unobserved (unmeasured) chromatin proteins, and provide only the representation of a group of observed chromatin components that most probably influence (i.e., conditionally dependent) each other. Second is the ability to provide an indication of the flow of causality in the system,i.e, the most probable direction of targeting interactions with the assumption of no hidden common cause **?**. However, the causal patterns learned by BNI must be interpreted with caution, not as evidence but rather as the indication. For more details on this matter, the reader is referred to **??**.

Applying BNI algorithm, they inferred BN of targeting interactions among 43 chromatin components which provides predictions of interesting targeting interactions in chromatin coupled with experimental validation**?**. This network highlighted novel probable interactions, i.e, HP1 and HP1c each target HP3 competitively to different locations along the genome. Furthermore, on the basis of network topology analysis, they demonstrated a chromatin component that is mostly connected, i.e, Brahma. This component is interpreted as a key player in the targeting other chromatin components. These results reflected the usefulness of the inferred network to predict the non-canonocial targeting interactions and guide the experimental targets that provides biological insight.

However, this inferred network suffers from limitations of its construction, imposed mainly by noise arising from the heuristic characteristic of BNI algorithm. It is therefore critical to evaluate and identify the reproducibility of this inferred BN of targeting interactions in chromatin. In this study, attempts are made to reconstruct the BN of targeting interactions in chromatin using the same dataset and BNI algorithm in order to assess its reproducibility.

2 METHODS

Bayesian Network (BN)
A bayesian network (BN) is a probabilistic model of dependencies for a set of variables in order to understand a particular system**?**. In this study, the

BN is used as a model of the targeting interactions among all 43 chromatin components in chromatin. The core assumption of the BN is based on the notion of conditional independence.

In the network, the nodes correspond to the observed variables (in this study, i.e.,chromatin components) and directed edges denote dependencies between the variables. If there is an edge from node X to a node Y (X → Y), X is a parent of Y and the edge represents the probable targeting interactions between these two chromatin components, i.e., X and Y. This interaction is interpreted as follows, as previously described in **?** : The presence of X at its specific locations along the genome (X's loci) promotes the association of Y to these X's loci. In addition to that, this denotes a functional interaction rather than biochemical interaction, implying that the interaction is not necessarily a direct contact but it may be mediated by other components.

The formal statement of this BN network of targeting interactions is as follows: The BN for a set of variables $\chi = \{X_1, ..., X_{43}\}$ is a pair of $\langle G, \Theta \rangle$ where 1 to 43 denotes 43 chromatin components and G represents a directed acyclic graph (DAG) on χ and Θ is a joint probability distribution on χ. such that the notion of conditional independence (Markov assumption) is satisfied: Each variable is independent of its non-descendants, given its parents in G. The joint distribution Θ is defined as a product of conditional probabilities of each variable (θ), expressed as follows:

$$P(X_1, ..., X_{43}) = \prod_{i=1}^{43} P(X_i | \text{Parents}(X_i))$$

Parents(X_i)) represents a set of parents for each variable. For more details on model semantics, the reader is referred to **?**.

Learning BN of Targeting Interactions from Scratch
In this study, BN is learned from using the observational data, i.e, the dataset of the collection of 43 binding maps from scratch. Learning from scratch means without the prior biological knowledge about the absence or presence of edges(dependencies) between chromatin components. The dataset (D) is a matrix of dimension 4380(rows) × 43(columns). Each Row represents a genomic location(i.e, gene) and each column corresponds to a chromatin component. The problem of learning BN is stated as follows (adapted from **?**):

Problem Statement:
Given a dataset $D = \{\mathbf{x^1}, ...\mathbf{x^{4380}}\}$ of independent instances of χ assuming that each genomic location (i.e, gene) as an independent sample. Find a network B =$\langle G, \theta \rangle$ that best matches D.

The common approach for this problem is by introducing a scoring function to evaluate each possible inferred network with respect to the dataset, followed by searching for the optimal network according to the score **?, ?**.

Scoring
In this study, the bayesian-based scoring function, i.e., Bayesian Dirichlet equivalent (BDe) is applied. This function evaluates the posterior probability of a graph (network) given the dataset**?**, expressed generally as:

$$\text{Score}(G : D) = \log P(D|G) + \log P(G) + C$$

$P(D|G)$ is the probability of the data given a graph as the marginal likelihood, $P(G)$ is the prior probability of the network structure, and C is the normalizing factor (is not required to compare the score between networks). The marginal likelihood is computed by averaging the probability of the data over all possible parameter (Θ) assignments to G, as expressed below.

$$P(D|G) = \int P(D|G, \Theta).P(\Theta|G).\,\mathrm{d}\Theta$$

The BDe scoring function requires the variables to be discrete, $P(\Theta|G)$ to follow Dirichlet prior distribution and $P(G)$ to have uniform distribution in order to make that this function asymptotically consistent that guarantee score equivalence. For more technical details and explanations, the interested reader is referred to **??**. In this study, the continuous binding data for each chromatin component were discretized based on 95% percentile threshold in which the top 5% of the strongly bound genes are assigned to 1 (target) and the rest of the lower quantile to 0 (non-target) as described in **?**.

Searching
Finding the optimal network given the dataset is classified as NP-complete owing to the excessively large number of possible networks that can be constructed and evaluated given the dataset**?, ?, ?**. Hence obtaining the exact solution is not possible in a reasonable time at this current moment. For this reason, the heuristic approach (does not guarantee for the optimal solution, but it may give an acceptable solution in a fast manner for some instances) is proposed, such as simulated annealing and greedy-hill climbing. The principles of these heuristic methods are based on the successive random mutations (add or remove or reverse edges) to the initial network state and the iterative evaluation of the resulting mutated networks. Searching is repeated several times as specified by the user, each time with different initial networks and stopped with some criteria, e.g., after a certain number of networks is being evaluated **?**. However these heuristic methods suffers from the curse of noise in providing the solution. This is due to the random starts and mutations during the search. Nevertheless, these methods in some cases can still give close estimates(reviewed in **?**).

Simulated annealing is preferred over greedy-hill climbing as the former method accommodates an escape from local maxima which is the common problem of the latter method. The escape from local maxima is achieved by permitting some changes that may reduce the network's score at the initial phase of searching with a probability specified by a parameter T. As the search progresses, i.e., repeated several times, the T's value is lowered gradually. The main idea of simulated annealing is to allow the search algorithm exploring the search space at the start of the process so that the sensitivity of the final solution to the initial network state is reduced, thereby lowering the chances of being trapped at local maxima**?**.

van Steensel, et al., (2010) learned the BN of targeting interactions using bootstrapping approach. The principle of the bootstrapping approach is based on the generation of slightly perturbed dataset of the original dataset for a number of times and learn the BN from them. This results in the collection of many fairly reasonable model with respect to the data and these models are used to estimate the confidence of the features, i.e, are X and Y close neighbors(being a parent/child), therefore having the targeting interaction in all the networks?

Learning BN using Banjo
The learning procedure is run using the software application Banjo 2.0 (developed by A.J. Hartemink), in which the concepts of learning BN as described previously are implemented. For this learning of BN using Banjo search, BDe scoring function (default evaluator Choice) to evaluate BN, and simulated annealing with RandomLocalMove for network searching and other parameter settings were specified as described in supplementary data4.txt **?**. The bootstrapping procedure (adapted from **?**) is formally stated as follows:

- For $i = 1, ..., 1000$,

- Construct a slightly perturbed input data, D_i, generating 1000 input data, each input data matrix has a dimension of 4380 rows each selected by random sampling with replacement from D (original dataset) and 43 columns (chromatin components)

- Apply the learning procedure using Banjo search on D_i, generating $(G_1, ..., G_{1000})$in total of 1000 networks.

- For each feature of interest compute:

$$\text{Conf}(f) = \frac{1}{m} \sum_{i=1}^{m} f(G_i)$$

$$f(G) = \left\{ \begin{array}{ll} 1 & \text{if } f \text{ is a feature in G} \\ 0 & \text{if } f \text{ is absent} \end{array} \right\}$$

This gives the confidence values from bootstrapping, i.e, Bootstrap Confidence Score (BCS) for each possible pair of chromatin components (feature).

The bootstrapping procedure in this study was run overnight on 1.6 GHz Intel Core i5, processed using R software and stored as a text file.

Learning causality from the inferred BN
Initially, BN was constructed as a structured directed graph (DAG) to represent the dependence relationships between variables. As BN encodes the directions of the relations, this has motivated researchers to exploit further the possibility to infer the causal dependence from BN. Friedman, et al., 2000 proposed that the inferred BN may provide the causal dependencies between variables, although at best as indication but not evidence. Note that BN is not a causal network, although they both share similar mathematical network representation.

In the causal network, X → Y *(left)* ≠ Y → X *(right)*. This is interpreted as X causes Y *(left)*, in a way that manipulating the value of X (by mutations) affects the value of Y, but not vice versa and Y causes X *(right)* means that manipulating Y affects X. Whereas in BN, **X → Y** *(left)* ≡ **Y → X** *(right)* since BN only models the dependency between variables. They both (*left and right*) are equivalent and the directions of edges can't be determined, therefore they are represented in the inferred causal structure from BN as **X − Y**. They encode the same dependency statement, i.e, **X** and **Y** are dependent, and therefore does not imply the causal pattern. However, van Steensel and colleagues (2010) proposed that the flow of causality can still be predicted from the bootstrapping approach during learning BN under the assumption that no hidden variables for the common causes. The intuition is as follows: A certain direction of an edge may hold a small amount of information in 100 samples and this might be a spurious causality. But if a certain direction remains over a certain confidence threshold, in this case, over 800 out of 1000 samples, then they start to believe the direction of the dependency. Although the causal structure is still unknown but at best it can give the indication that may guide the design for experimental targets. Applying this approach, the directions of edges are predicted by counting the edge with each direction linking this pair of components (feature), i.e, the bootstrap score for each direction in the 1000 Banjo Networks.

BN80 reconstruction
For the reconstruction of BN80, 80% cutoff of confidence score (BCS) was selected as previously described in van Steensel, 2010. After applying 80% cutoff on the bootstrap.freq.parent2child or bootstrap.freq.child2parent, or bootstrap.freq.combined, the predicted targeting interactions of components were filtered and visualized using Cytoscape2.8.3 **?**, applying the layout algorithm, Edge-Weighted Force-Directed, in which the weighted edge maps the BCS. Nodes are organized as previously inferred BN. Note that the final BN80 of targeting interactions constructed by van steensel and colleagues (2010) is an average network of 1000 reasonable BN network (bootstrapping approach) with 80% cutoff on the bootstrap confidence score (BCS).

Network topology analysis
In this study, the degree (connectivity), i.e, the out-degree and out-degree distribution were analyzed as measures to permit the comparison of network topology with the previous published BN80 network. The out-degree and out-degree distribution are obtained by counting how many edges directed out from a node and counting the number from each node with a certain number of edges, respectively. The analysis was run in R using *igraph* package.

Mining previously known interactions from literature
The predicted targeting interactions among 43 components from the reconstructed BN80 was validated by mining the predicted pairs of chromatin components (*as input*) from literature using NCBI ESearch. This process is run using NATbox toolbox in R**?**. The tool is available in the supplemented folder (ProjectsBNI_chromatinRanalysisanalysis). This toolbox searches for the co-occurrences predicted pairwise interactions in the abstracts deposited in PUBMED database (version June,16 2012). The input search term follows as described by van Steensel, et al., 2010. The *output* of this validation is in the HTML format containing the retrieved pairs of components with hyperlinks to the PUBMED abstracts. The number of the retrieved HTML was counted by a python script, i.e, parseHTM.py (stored in the supplementary folder).

R Info
In this study, the analysis was run in R

- R version 2.15.0 (2012-03-30), `x86_64-apple-darwin9.8.0`
- Locale: `C/en_US.UTF-8/C/C/C/C`
- Base packages: base, datasets, grDevices, graphics, grid, methods, stats, utils
- Other packages: ggplot2 0.9.1, igraph 0.5.5-4, plyr 1.7.1, reshape 0.8.4
- Loaded via a namespace (and not attached): MASS 7.3-18, RColorBrewer 1.0-5, colorspace 1.1-1, dichromat 1.2-4, digest 0.5.2, labeling 0.1, memoise 0.1, munsell 0.3, proto 0.3-9.2, reshape2 1.2.1, scales 0.2.1, stringr 0.6, tools 2.15.0
- Environment run in Rstudio 0.95.265

3   RESULTS AND DISCUSSION

Reconstructed BN80
The attempts are made to reconstruct BN of targeting interactions among 43 chromatin components in chromatin that was previously published **?**, in order to identify and evaluate its reproducibility. The reconstruction was conducted using similar dataset and BNI algorithm as described in their methods. Note: Results presented here are meant to be comparative with the previously published BN80 **?**.

The BN was reconstructed using bootstrapping approach to compute the confidence scores for each predicted interaction as described in the method section above: *Learning BN using Banjo*. The result is presented as a supplementary dataset BootstraapScoreRL.txt PATH: BNI_chromatin/Ranalysis/analysis. After applying the cutoff of confidence score of 80%, BN80 network was reconstructed. 50 predicted targeting interactions of chromatin components were recovered in this reconstructed BN80. In the previous BN80, 52 predicted interactions were reported. This result indicates that the reconstructed BN80 has a high coverage (50 out of 52 predicted pairs), i.e, **96.15%** relative to the previous BN80.

Because the method to visualize the previous BN80 was not mentioned in their publication, the reported 52 predicted targeting interactions was visualized using similar method as described in the method section above *BN80 reconstruction* for comparison with the reconstructed BN80. Figure 1 shows the previous(A) and reconstructed BN80s (B) sharing similar targeting interactions with the exception of brm-His 3.3 and D1- dm. The missing chromatin component in the reconstructed BN80 was marked with red node color,i.e., His 3.3 in the previous BN80. This displays

the missing pair, i.e, the interaction between His 3.3 and Brm in the reconstructed BN80. The interaction between D1- dm was also absent in the reconstructed BN80 as the edge connecting them did not exist.

One of the possible reasons for the absence of these interactions in the reconstructed BN80 is due to the arbitrary cut- off of 80% of the bootstrapping score. The bootstrapping scores for these interactions, i.e, brm-His 3.3. and D1-dm are 79.9% and 77.8%, respectively. Therefore, applying this cut-off policy of 80% , i.e, ($> 80\%$) would filter out these interactions, although, the scores of these interactions are close to the cut-off. This indicates that the arbitrary characteristic of the cut-off policy emerges as the weakness of their BNI algorithm ? for its reproducibility.

Besides these two missing targeting interactions, figure 1 shows that 50 of the recovered predicted targeting interactions also share similar predicted causal patterns with the previous BN80. These results altogether suggest that the previous BN80 can be reconstructed with a relatively high coverage (96.15%) and similar predicted causality directions, thus indicating a high level of reproducibility of the previously published BN80.

Network analysis
Furthermore, the degree centrality as a common measure to inspect the network topology was also analyzed in order to investigate whether the reconstructed BN80 also shares similar network topology with the previous BN80.

In Figure 2, the distribution of the degree, i.e, the out-degree(edges directed out from a component) is shown, indicating that the most of the components in the reconstructed and previous BN80 have roughly similar number of edges, close to the average of 2.5. Biologically, this can be interpreted as most of the chromatin components involves in targeting 2 or 3 other components.

Whereas, chromatin components that have more or less edges, that far away from the average, are very rare. As shown in Figure 2, only one chromatin component that has either 0 or 6/5 out-degrees. Su(var)3-7 and brm are components with 0 and 6/5 degrees respectively (see figure 1). Note that in the reconstructed BN80 the highest out-degree is 5 while in the previous BN80 is 6. This is due to the missing pairwise interactions between brm-Hist3.3, as identified above. In the reconstructed BN80s, brm is also identified as the component with the highest degree serving as hub in the network as previously reported. The hub in the network is biologically interpreted as a central player in orchestrating the targeting interactions of many chromatin components.

The overall result for network analysis indicates that the reconstructed and previous BN80 also share similar network topology with respect to the degree-centrality with similar out-degree average of 2.5.

Retrieving predicted interactions from PubMed
BN80 provides predictions for the targeting interactions in chromatin. These predicted interactions include the non-canonical and also the known ones. In order to estimate the predictive performance of BN80, the known predicted interactions were queried in the PubMed database for the publications that mention these interactions as pair co-occurances in their abstracts. The results of this query were the lists of pairwise interactions of chromatin components in html format that include the hyperlinks
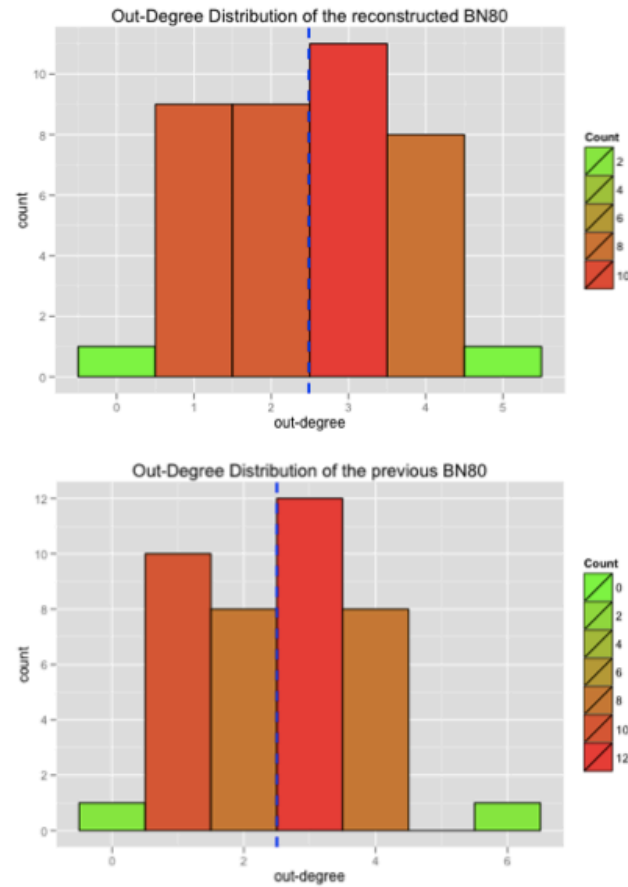

Fig. 2. Out-Degree Distribution. The dotted blue line displays the average out-degree

to the respective abstracts in PubMed (*pairs_paper* and *pairs_RL* for previous and reconstructed BN80s, respectively).

The PubMed query for the reconstructed BN80 shows that 42 predicted pairs of interactions were retrieved in PubMed database out of 50 predictions, hence, the PubMed recovery rate is 84%. The previous BN80 was reported to have 48% PubMed recovery. However this PubMed search was done using the old database, i.e, version October 28 2008. For this reason, the PubMed query was reanalyzed using the more recent version of database(2012) for comparison. The resulting query shows that 42 pairs were overlapping with the PubMed's abstracts out of 52 predicted interactions, thereby the recovery rate of the previous BN80 is 81%. These results, altogether, demonstrate an increase in the estimation of the predictive performance of the previous BN80 from 48% in 2008 to 81% in 2012, as more reported targeting interactions were deposited in PubMed and importantly, the predictive performance of the reconstructed BN80 (84%) is comparable with the previous BN80 (81%).

Furthermore, among all possible combinations of pairwise interactions of these 43 components, i.e, 903 interactions were also queried in PubMed in order to contrast with the predicted interactions derived by BN80s. This search's retrieved 552 out of 903 interactions, i.e, the recovery rate of 61.1%. Note that

A

B
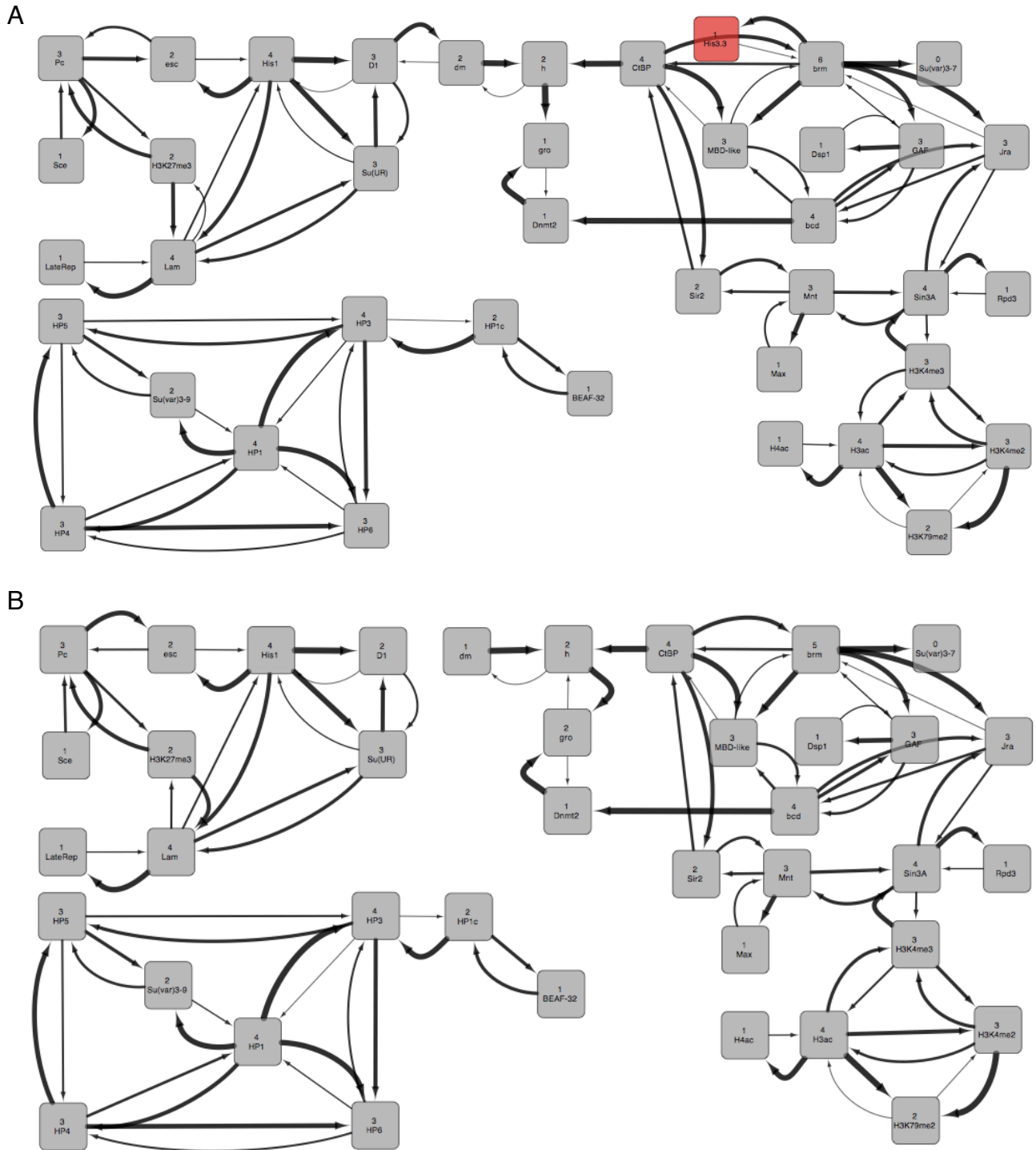
Fig. 1. Inferred BN80. A) from Paper **?** B) the reconstructed network. The number in each node(chromatin component) denotes the out-degree(number of edges directed out from the node). The thickness of the edges corresponds to the weights of the bootstrapping score

Fig. 3. PubMed Validation

predictions acquired by all possible combinations give rise to the increase of the number of predicted interactions, however, with a cost of a large proportion of these interactions were not retrieved from the PubMed database. Therefore, the predictive accuracy becomes lower bringing about the lower overall recovery rate. These results above indicates an higher predictive performance with respect to the accuracy of BN80s (81%-84%) in contrast when the predictions obtained by all combinations (61.1%). Although the margin (20% difference between BN80's and all combinations' predictions) is less striking than in the previous search using the old PubMed Database(2008), in which the recovery rates of previous BN80 and all pairs combinations are 48% and 13%, respectively, i.e., a margin of 35%.

However, the PubMed validation only provides a rough estimation, because the retrieved pair co-occurrences do not necessarily include targeting interactions between them. Therefore the found interactions in PubMed are still required to be manually curated by experts.

4 CONCLUSION

With the availability of the binding maps of chromatin components, and BNI algorithm as previously published, effort has been made to reconstruct the previously published BN80 of targeting interactions in chromatin. The results in this study indicate that the reconstructed BN80 has high coverage of previously reported targeting interactions, in which 50 out of 52 pairwise interactions were recovered. Furthermore, this reconstruction of BN80 also yielded a comparable network topology with the average out-degree of 2.5. These results, altogether, demonstrate the high quality of reproducibility of the previously published BN80.

Despite its high reproducibility, caution must be taken particularly for the application of this BN80 to infer causal directions of targeting interactions. This is due to the inherent nature of bayesian network that models only the dependencies of the variables not the flow of causality. Although the authors**?** suggest that using bootstrapping approach the most probable directions could also be modeled, however this is under a strict assumption that no hidden common cause between chromatin components. As BN80 was constructed using only 43 components while in reality more than 43 components are expected to orchestrate the targeting interactions in chromatin, therefore the model of causality direction would be more likely false due to hidden components that may mediate the causality directions.

Finally, one of the most exciting longer term possibility of this research line is to incorporate the temporal information, such as during different developmental stages into the network. This may transform the static network into dynamic and further equipped the network with possible feedbacks that may model the targeting interactions in chromatin closer to the real process underlying the targeting interactions in chromatin.

SOFTWARE ACKNOWLEDGEMENT

a. Banjo (2012). Bayesian Network Inference with Java Objects, developed by A.J.Hartemink.
Available from http://www.cs.duke.edu/ amink/software/banjo/
b. Csardi G, Nepusz T (2006). The igraph software package for complex network research, InterJournal, Complex Systems 1695. http://igraph.sf.net
c. Heckmann, M. (2011). OpenRepGrid - An R package for the analysis of repertory grids. University of Bremen, Germany, Diploma thesis.
d. G. van Rossum and F.L. Drake(eds) (2001). Python Reference Manual, PythonLabs, Virginia, USA. Available at http://www.python.org
e. H. Wickham (2007). Reshaping data with the reshape package. Journal of Statistical Software, 21(12)
f. H. Wickham (2009). ggplot2: elegant graphics for data analysis. Springer New York
g. R Development Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing,Vienna, Austria. ISBN 3-900051-07-0,
URL http://www.R-project.org.
h. RStudio (2012). RStudio: Integrated development environment for R (Version 0.96.122) [Computer software]. Boston, MA. Retrieved May 20, 2012. Available from http://www.rstudio.org/