Learning Bayesian Networks Structure Based Part Mutual Information for Reconstructing Gene Regulatory Networks

Qingfei Meng[1,2], Yuehui Chen[1,2](✉), Dong Wang[1,2](✉), and Qingfang Meng[1,2]

[1] School of Information Science and Engineering,
University of Jinan, Jinan 250022, China
{yhchen, ise_wangd}@ujn.edu.cn
[2] Shandong Provincial Key Laboratory of Network
Based Intelligent Computing, Jinan 250022, China

Abstract. As a kind of high-precision correlation measurement method, Part Mutual Information (PMI) was firstly introduced into Bayesian Networks (BNs) structure learning algorithm in the paper. Compared to the general search scoring algorithm which set the initial network as an empty network without edge, our training algorithm initialized the network structure as an undirected network. That meant that our initial network identified the genes related to each other. And then the following algorithm only needed to determine the direction of the edges in the network. In the paper, we quoted the classic K2 algorithm based on Bayesian Dirichlet Equivalence (BDE) scoring function to search the direction of the edges. To test the proposed method, We carried out our experiment on two networks: the simulated gene regulatory network and the SOS DNA Repair network of Ecoli bacterium. And via comparison of different methods for SOS DNA Repair network, our proposed method was proved to be effective.

Keywords: Gene regulatory networks · Bayesian Networks · Part Mutual Information · K2 algorithm · BDE scoring function

1 Introduction

The advances on high-throughput technology enabled us to obtain a great deal of gene expression data during the past ten years [1]. Meanwhile, a large number of models and methods emerged for accurately reconstructing gene regulatory networks. For example, the boolean network, the system of differential equations, artificial neural networks, bayesian network, and so on. Among these, the Bayesian Networks become a main model in the research of gene regulatory networks for it's advantages, including handling incomplete data sets, fully combing the prior knowledge of the domain, visual image, and so on.

Generally there were two kinds of algorithms for Bayesian network structure learning: Search scoring method and correlation analysis method. The search scoring method usually introduced a scoring function $S(G, D)$ and then used this function to

D.-S. Huang et al. (Eds.): ICIC 2017, Part II, LNCS 10362, pp. 647–654, 2017.
DOI: 10.1007/978-3-319-63312-1_57

evaluate each possible network structure for finding an optimal solution from all possible network structures. The correlation analysis method captured the dependency between nodes via independence test to study the network structure. In the study, we quoted the classic K2 algorithm [2] based on Bayesian Dirichlet Equivalence (BDE) scoring function to learn Bayesian Networks structure. The innovation of our method lied in introducing Part Mutual Information (PMI) to initialize the Bayesian Network structure.

The method based on the Mutual Information (MI) firstly obtained the mutual information matrix by calculating the mutual information between all possible gene pairs and then constructed the regulatory network through the mutual information matrix [3]. Although the method achieved good results, the limitation of this method was that it could not distinguish the direct and indirect association between genes with a high false positive. So Wang et al. [4] proposed a method based on conditional mutual information (CMI), which could distinguish the direct and indirect relationships between genes, and greatly reduced the false positive rate. However, the model based on conditional mutual information also had some limitations. For example Chen et al. [5] pointed out that there is a false negative problem based on conditional mutual information. When we measured the dependency between variables X and Y given variable Z, CMI could not correctly measure the direct association or dependency if X (or Y) was strongly associated with Z. In order to solve this problem, a method based on partial mutual information [5] was proposed, which solved the false positive problem of mutual information model and the false negative problem of conditional mutual information.

In the paper, we took advantage of partial mutual information to determine the relationship between variables for initialing Bayesian network structure. Then use K2 algorithm based on BDE scoring function to search the optimal network structure. Through two groups of experiments, our proposed method was illustrated to be effective.

2 Method

2.1 Part Mutual Information

Assuming that X and Y were two random variables, MI was defined on the basis of an extended Kullback–Leibler (KL) divergence D [6]:

$$MI(X;Y) = D(p(x,y)\|p(x)p(y)) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{1}$$

where $p(x,y)$ was the joint probability distribution of X and Y and $p(x)$ and $p(y)$ were the marginal distributions of X and Y, respectively. Clearly, MI in Eq. 1 was evaluated against the ‘mutual independence’ of X and Y, which was defined as

$$p(x)p(y) = p(x,y) \tag{2}$$

CMI for variables X and Y given Z could further detect nonlinearly direct association and was defined as:

$$\begin{aligned} CMI(X;Y|Z) &= D(p(x,y,z)\|p(x|z)p(y|z)p(z)) \\ &= \sum_{x,y,z} p(x,y,z)\log\frac{p(x,y|z)}{p(x|z)p(y|z)} \end{aligned} \tag{3}$$

So the conditional independence of X and Y given Z, which was defined as:

$$p(x|z)p(y|z) = p(x,y|z) \tag{4}$$

To analogy Eqs. (2), (4), partial independence of X and Y given Z was defined as:

$$p^*(x|z)p^*(y|z) = p(x,y|z) \tag{5}$$

where $p^*(x|z)$ and $p^*(y|z)$ were defined [7] as

$$p^*(x|z) = \sum_y p(x|z,y)p(y), p^*(y|z) = \sum_x p(y|z,x)p(x) \tag{6}$$

So

$$PMI(X;Y|Z) = D(p(x,y,z)\|p^*(x|z)p^*(y|z)p(z)) \tag{7}$$

Or as

$$PMI(X;Y|Z) = \sum_{x,y,z} p(x,y,z)\log\frac{p(x,y|z)}{p^*(x|z)p^*(y|z)} \tag{8}$$

2.2 Bayesian Networks

The Bayesian Networks were a directed acyclic graph $B = (G,\Theta)$ which shown the probabilistic dependency relationship between variables. $G = (V,E)$ was a directed acyclic graph and Θ was the collection of conditional probability table (CPT). $V = (V1,\cdots,Vn)$ was a set of nodes and each node represented a field random variable Xi.

$$E = \{<Xi,Xk> |Xi,Xk \in V, i \neq k\} \tag{9}$$

Where $<Xi,Xk>$ indicated the directed edge. $Xi \rightarrow Xk$ indicated the probabilistic dependency relationship between variables Xi and Xk. Xi was the father node of Xk. $Pa(Xk)$ indicated the parent node set of variable Xk.

For every variable Xk, it's value xk had the following parameter $\theta xk|pa(Xk) = P(xk|pa(Xk))$, which showed the probability of xk occurrence under the condition of $Pa(Xk)$. So the joint conditional probability distribution on a given set of variables of Bayesian Networks was as:

$$P(X1, X2, \cdots, Xn) = \prod_{i=1}^{n} P(Xi|Pa(Xi)) \tag{10}$$

2.3 K2 Algorithm

Cooper [2] put forward K2 algorithm based on Bayesian Dirichlet Equivalence (BDE) scoring function to learn Bayesian Networks structure. The K2 algorithm used the way of hill-climbing search to learn the network structure.

According to prior knowledge, We first obtained the initial network. Then, Use the search operator including reducing edges, increasing edges and reserving edges to modify the current network to get candidate network. According BDE scoring function, Select the fractional optimal network to replace the current network, and then continue the search until obtaining the best network.

The BDE scoring function an approximation of the marginal likelihood function under the condition of large sample. The BDE scoring function was defined as:

$$S(G, D) = LgP(D|G) + LgP(G) - \frac{d}{2}\log m \tag{11}$$

Where $P(D|G)$ was an edge distribution of dataset D and was the probability averaging of data from D. $P(G)$ was the prior probability of the network structure G. $\frac{d}{2}\log m$ was a penalty function for sample size m. Because of the addition of a penalty term, the BIC scoring function avoided overfitting.

2.4 Evaluation Index

In our study, five criterions (True Positive Rate (TPR), False Positive Rate (FPR), Positive Predictive (PPV), Accuracy (ACC) and F-score) were used to test the performance of the proposed method. Their definition was given as follows:

$$TPR = \frac{TP}{TP + FN} \tag{12}$$

$$FPR = \frac{FP}{FP + TN} \tag{13}$$

$$PPV = \frac{TP}{TP + FP} \tag{14}$$

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

$$F - score = \frac{2PPV * TPR}{PPV + TPR} \tag{16}$$

Where True Positive (TP) meant that edges in real networks were identified as edges in the model. False Positive (FP) meant that edges not in real networks were identified as edges in the model. True Negative (TN) meant that edges not in real networks were not identified as edges in the model. False Negative (FN) meant that edges in real networks were not identified as edges in the model.

3 Experimental Results and Analysis

3.1 Experiment 1: Simulated Gene Regulatory Network

Figure 1 showed a simulated gene regulatory network which was modeled by a S-system model [8]. The model consisted of n non-linear ordinary differential equations and the generic form of equation i is given as follows:

$$X_i^{'}(t) = \alpha_i \prod_{j=1}^{N} X_j^{g_{ij}}(t) - \beta_i \prod_{j=1}^{N} X_j^{h_{ij}}(t) + \varepsilon$$

Where X_i was a vector element of dependent variable, N was the number of variables, α_i and β_i was vector elements of non-negative rate constants, g_{ij} and h_{ij} were matrix elements of kinetic orders and random Gaussian noises (ε) were added to each equation independently.

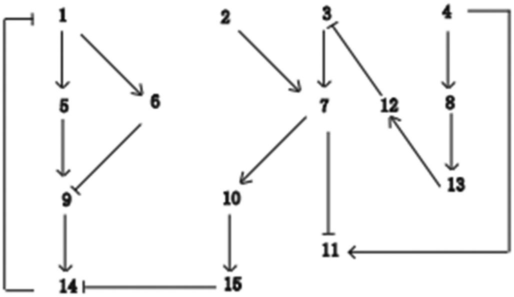

Fig. 1. A simulated gene regulatory network containing 15 genes

Table 1 listed the parameters of S-system containing 15 genes. The initial value is random and we respectively got the simulated data by 20 time points, .30 time points and 50 time points.

Table 1. Parameters of the S-system

α_i	1.0
β_i	1.0
$g_{i,j}$	$g_{1,14} = -0.1$, $g_{3,12} = -0.2$, $g_{5,1} = 1.0$, $g_{6,1} = 1.0$, $g_{7,2} = 0.5$, $g_{7,3} = 0.4$, $g_{8,4} = 0.2$, $g_{9,5} = 1.0$, $g_{9,6} = -0.1$, $g_{10,7} = 0.3$, $g_{11,4} = 0.4$, $g_{11,7} = -0.2$, $g_{12,13} = 0.5$, $g_{13,8} = 0.6$, $g_{14,9} = 1.0$, $g_{14,15} = -0.2$, $g_{15,10} = 0.2$, other $g_{i,j} = 0.0$
$h_{i,j}$	1.0 if $i = j$, 0.0 otherwise

From Table 2, we can see the TPR and FPR obtained by our proposed algorithm.

Table 2. Performances of the proposed model to the simulated data

Number of time points	TPR	FPR
20	0.63	0.57
30	0.77	0.41
50	1.0	0.12

3.2 Experiment 2: SOS DNA Repair Network of E.Coli Bacterium

The datasets from SOS DNA Repair network of E.coli bacterium [9] contained four experiments under various light intensities ($5Jm^{-2}$, $5Jm^{-2}$, $20Jm^{-2}$, $20Jm^{-2}$). Each experiment (http://www.weizmann.ac.il/mcb/UriAlon/ Papers/SOSData) consisted of 50 time points evenly spaced by 6 min and referred to eight genes: uvrD, lexA, umuD, recA, uvrA, uvrY, ruvA and polB. Figure 2 displayed the known real regulation among 8 genes.

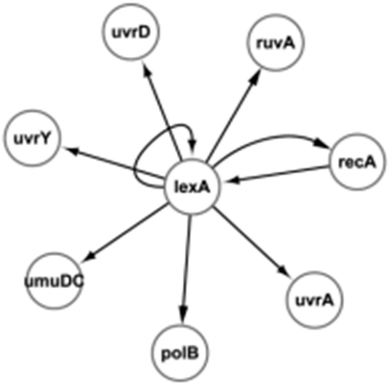

Fig. 2. The true SOS network with eight genes and 9 interactions

In the experiment, we firstly normalized the gene expression data for each gene to the interval [0, 1] using

$$x_i^{'}(t) = \frac{x_i(t) - \min_i}{\max_i - \min_i}$$

Where $x_i(t)$ is the actual measured gene expression level of gene i at time point t. min_i and $\max_i$ respectively were the minimum and maximum of gene i expression level. Table 3 showed the detail comparison between proposed method and other methods including S-system [10], ODEs [11], RNN [12] and DBN [13]. The results showed that our proposed method performed better than other methods.

Table 3. Comparison of different methods for SOS dataset

Item	S-system	ODEs	RNN	DBN	Proposed Method
TP	5	6	5	4	8
FP	10	15	2	5	7
TPR	0.5556	0.6667	0.55556	0.4444	0.8889
FPR	0.20833	0.3125	0.041667	0.10417	0.2692
PPV	0.3333	0.28571	0.71429	0.44444	0.5333
ACC	0.6667	0.57895	0.80702	0.75439	0.7714
F-score	0.41667	0.4	0.625	0.44444	0.6667

4 Conclusions

In the paper, the Part Mutual Information was introduced into the training algorithm of the Bayesian Networks structure, which meant that the initial network has identified the genes related to each other and the later searching process with K2 algorithm only needed to determine the direction of the edges in the Bayesian Networks. We carried out our experiments on two networks: the simulated gene regulatory network and the SOS DNA Repair network of Ecoli bacterium. And via comparison of different methods for SOS DNA Repair network, our proposed method was proved to be effective.

Acknowledgment. This work was supported by the National Key Research and Development Program of China (2016YFC106000), the National Natural Science Foundation of China (Grant No. 61671220, 61640218, 61201428), the Shandong Distinguished Middle-aged and Young Scientist Encourage and Reward Foundation, China (Grant No. ZR2016FB14), the Project of Shandong Province Higher Educational Science and Technology Program, China (Grant No. J16LN07), the Shandong Province Key Research and Development Program, China (Grant No. 2016GGX101022).

References

1. Cho, K.H., Choo, S.M., Jung, S.H., et al.: Reverse engineering of gene regulatory networks. IET Syst. Biol. **1**(3), 149–163 (2007)
2. Cooper, G.F., Herskovits, E.: A Bayesian method for the induction of probabilistic networks from data. Mach. Learn. **9**(4), 309–347 (1992)
3. Ji, Z., Wu, D., Zhao, W., et al.: Systemic modeling myeloma-osteoclast interactions under normoxic/hypoxic condition using a novel computational approach. Sci. Rep. **5** (2015)
4. Wang, B., Zhang, J., Chen, P., et al.: Prediction of peptide drift time in ion mobility mass spectrometry from sequence-based features. BMC Bioinform. **14**(8), S9 (2013)
5. Altay, G., Emmertstreib, F.: Revealing differences in gene network inference algorithms on the network level by ensemble methods. Bioinformatics **26**(14), 1738–1744 (2010)
6. Bao, W., Chen, Y., Wang, D.: Prediction of protein structure classes with flexible neural tree. Bio-Med. Mater. Eng. **24**(6), 3797–3806 (2014)
7. Zhao, J., Zhou, Y., Zhang, X., et al.: Part mutual information for quantifying direct associations in networks. Proc. Nat. Acad. Sci. **113**(18), 5130–5135 (2016)
8. Schreiber, T.: Measuring information transfer. Phys. Rev. Lett. **85**(2), 461–464 (2000)
9. Anzing, D., Balduzzi, D., Grosse-Wentrup, M., Schölkopf, B.: Quantifying causal influences. Ann. Stat. **41**(5), 2324–2358 (2013)
10. Kimura, S., Ide, K., Kashihara, A.: Inference of S-system models of genetic networks using a cooperative coevolutionary algorithm. Bioinformatics **21**, 1154–1163 (2005)
11. Ronen, M., Rosenberg, R., Shraiman, B.I., et al.: Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. Proc. Natl. Acad. Sci. U.S.A. **99**(16), 10555 (2002)
12. Noman, N., Iba, H.: Reverse engineering genetic networks using evolutionary computation. In: Genome Informatics International Conference on Genome Informatics, PubMed, pp. 205–214 (2005)
13. Ji, Z., Wang, B., Deng, S.P., et al.: Predicting dynamic deformation of retaining structure by LSSVR-based time series method. Neurocomputing **137**, 165–172 (2014)
14. Kimura, S., Sonoda, K., Yamane, S., et al.: Function approximation approach to the inference of reduced NGnet models of genetic networks. BMC Bioinform. **9**(1), 23 (2008)
15. Xu, R., Wunsch, D.C., Frank, R.L.: Inference of genetic regulatory networks with recurrent neural network models using particle swarm optimization. IEEE/ACM Trans. Comput. Biol. Bioinform. **4**(4), 681–692 (2007)
16. Perrin, B.E., Ralaivola, L., Mazurie, A., Bottani, S., Mallet, J., Buc, D.F.: Gene network inference using dynamic bayesian networks. Bioinformatics **19**(Suppl. 2), 138–148 (2003)