Reconstructing Gene Regulation Network based on Conditional Mutual Information

Bei Yang [a], Yaohui Xu [b]

School of Information Engineering, Zhengzhou University, Zhengzhou 450000, China

[a]iebyang@zzu.edu.cn, [b]ityaohuixu@163.com

Keywords: Mutual information, Conditional mutual information, Gene expression data, Conditions Relative Average Entropy.

Abstract. The purpose of gene regulatory network construction is to deduce the potential relationship between genes in gene expression data. In this paper, the algorithm CMICRAT is used to construct the gene-directed undirected graphs based on the mutual information and Conditional mutual information by the gene expression data, and then use the CRAE to determine the direction of the undirected graph. The experimental results show that the proposed algorithm can improve the accuracy of constructing the gene regulation network by experimentally validating the gene expression data of DREAM4 gene expression data in international biological competition.

1. Introduction

The construction of gene regulatory network, also known as reverse engineering [1] of gene regulatory network, aims to infer the potential relationship between genes in gene expression data. Gene regulatory networks can help people understanding the mechanism of gene regulation in the cell and revealing the mystery of life. The understanding of the relationship between gene regulations is initially based on biological experiments [2] to verify the interaction between genes, but this approach is costly, slow progress in research, becoming a bottleneck restricting the development of biological systems. With the development of biotechnology, such as gene chip microarray, the high-through put data provide us with an opportunity to understand the potential regulatory relationships of biological genes. The application of machine learning and data mining to the construction of gene regulatory networks has become one of the hotspots in bioinformatics.

2. Related Work

Bayesian network model [3] and the mutual information association model [4] are commonly used to construct the gene regulation network. The static Bayesian network model is introducing the joint probability distribution to construct the directed acyclic graph. Margolin [5] Using MI (Mutual Information) to calculate the mutual information between the two genes to measure the degree of association between genes in order to build a gene regulatory network. Sales [6] using mutual information to construct a gene regulatory network. But not suitable for measuring a gene at the same time by a number of gene regulation of the situation. Conditional Mutual Information (CMI) method was applied to Bayesian network construction and the non-linear dependence of genes was deduced in this paper [7].

In the process of gene regulation network construction, not only need to find a control relationship with the gene pairs, but also need to determine the direction of regulation between genes. This paper presents the algorithm CMICRAT to construct a gene regulatory network for gene expression data. Firstly, the non-directional network is constructed based on the mutual information and the conditional mutual information [7], and then the relative entropy of the condition is used to determine the direction of the edge of the undirected network.

3. Background

3.1 Conditional mutual information

Definition 1. For two random variables X and Y, the mutual information I (X, Y) was defined as:

$$\mathrm{I}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{X},\mathrm{Y}) \quad (1)$$

Where H (X) and H (Y) represent the entropy of the random variables X and Y respectively; H (X, Y) represents the joint entropy of the variable X and the variable Y.

$$\mathrm{I}(\mathrm{X},\mathrm{Y}) = \mathrm{H}(\mathrm{X}) + \mathrm{H}(\mathrm{Y}) - \mathrm{H}(\mathrm{X},\mathrm{Y}) \quad (2)$$

$$\mathrm{H}(\mathrm{X},\mathrm{Y}) = -\sum_{\mathrm{x}\in\mathrm{X},\mathrm{y}\in\mathrm{Y}} \mathrm{p}(\mathrm{x},\mathrm{y})\log\frac{\mathrm{p}(\mathrm{x},\mathrm{y})}{\mathrm{p}(\mathrm{x})\mathrm{p}(\mathrm{y})} \quad (3)$$

Where: p (x) represents the probability that the variable X takes x; $\mathrm{p}(\mathrm{x},\mathrm{y})$ represents the joint distribution probability of the variable X and variable Y.

The conditional mutual information [5] of the variable X and the variable Y under the variable Z condition $\mathrm{I}(\mathrm{X},\mathrm{Y}|\mathrm{Z})$ is recorded as

$$\mathrm{I}(\mathrm{X},\mathrm{Y}|\mathrm{Z}) = \sum_{\mathrm{x}\in\mathrm{X},\mathrm{y}\in\mathrm{Y},\mathrm{z}\in\mathrm{Z}} \mathrm{p}(\mathrm{x},\mathrm{y},\mathrm{z})\log\frac{\mathrm{p}(\mathrm{x},\mathrm{y}|\mathrm{z})}{\mathrm{p}(\mathrm{x}|\mathrm{z})\mathrm{p}(\mathrm{y}|\mathrm{z})} \quad (4)$$

Where $\mathrm{p}(\mathrm{x},\mathrm{y},\mathrm{z})$ represents the joint probability distribution of variables X, Y and Z; $\mathrm{p}(\mathrm{x}|\mathrm{z})$ represents the conditional probability that the variable x holds under the condition that the variable z is established; $\mathrm{p}(\mathrm{x},\mathrm{y}|\mathrm{z})$ represents the conditional probability that the variable x and the variable y are established under the condition that the variable z is satisfied.

3.2 Conditions Relative Average Entropy

Definition 2. For the random variable X and Y, its CRAE [8] is defined as:

$$\mathrm{CRAE}(\mathrm{X} \rightarrow \mathrm{Y}) = \frac{\mathrm{H}(\mathrm{Y}|\mathrm{X})}{\mathrm{H}(\mathrm{Y})\cdot|\mathrm{Y}|} \quad (5)$$

Where H(Y | X) represents the entropy of the variable Y under the condition of the given variable X. H (Y) represents the entropy of the variable Y; H (Y | X) represents the entropy of the variable Y under the condition of the given variable X

4. CMICRAT Algorithm

The algorithm CMICRAT mainly consists of two processes: first, the conditional mutual information is used to generate the unstructured control network, and then the direction of regulation of the gene pair is determined according to CRAE.

4.1 Generating an undirected gene regulatory network

Building an inbound control network consists of the following three steps [7]:

Input the gene expression data A, set the parameter θ to determine the independence, and generate the complete graph G according to the number of genes, k = -1.

Let k = k + 1: For a nonzero $\mathrm{G}(\mathrm{i},\mathrm{j}) \neq 0$, the genes associated with genes i and j are selected from the network. Calculate the number of these genes, recorded as m.

If m <k, the algorithm ends. If m ≥ k, L genes are selected from the m genes as the conditional genes, and they are $\mathrm{K} = [k_1, \cdots, k_l]$. Where K has a total of C_m^k species. Then we calculate the C_m^k L-order condition mutual information $\mathrm{I}(\mathrm{i},\mathrm{j}|\mathrm{K})$ and take their maximum value as $\mathrm{I}_{max}(i,j|K)$. $\mathrm{I}_{max}(i,j|K) < \theta$, $\mathrm{G}(\mathrm{i},\mathrm{j}) = 0$. Return (2)

4.2 Edge-oriented

Data discretization

Definition 3. The standard fraction $\sigma_{x,j}$ of the gene X at the time point t_j is defined as [9]:

$$\sigma_{x,j} = \frac{|x_j - \mu|}{\sigma} \quad (6)$$

Where μ represents the mean value of the gene expression for the total time of the gene X, σ is the standard deviation, and $x_j$represents the expression value of the gene X at the time t_j. Given the threshold k, if $\sigma_{x,j} \geq k$, then the gene X is expressed at time t_j, denoted as 1; otherwise the gene is not expressed,

Edge-oriented

The undirected network is determined by the CRAE.

For the two nodes X and Y in any edge of the network, if the $\mathrm{CRAE}(X \rightarrow Y) > \mathrm{CRAE}(Y \rightarrow X)$, then the existence of the node X and the node Y between the existence of $X \rightarrow Y$; if $\mathrm{CRAE}(X \rightarrow Y) \geq$ $\leq \mathrm{CRAE}(Y \rightarrow X)$, then the node X and the node Y are considered to be $Y \rightarrow X$.

5. Experimental Results

Data were generated using gene generator software GeneNetWeaver [10], which was used by the international competition for DREAM4, to generate gene data. In order to evaluate the performance of the algorithm CMICRAT, 10 genes of simulated gene data and standard network structure were generated.

The generated gene regulatory network is evaluated by the precision measure to measure the performance of the algorithm [11].

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP+FP}} \qquad (6)$$

Where TP represents the number of correct edges, the FP represents the number of false edges and FN represents the number of edges that exist in the real network but are not generated by the algorithm.

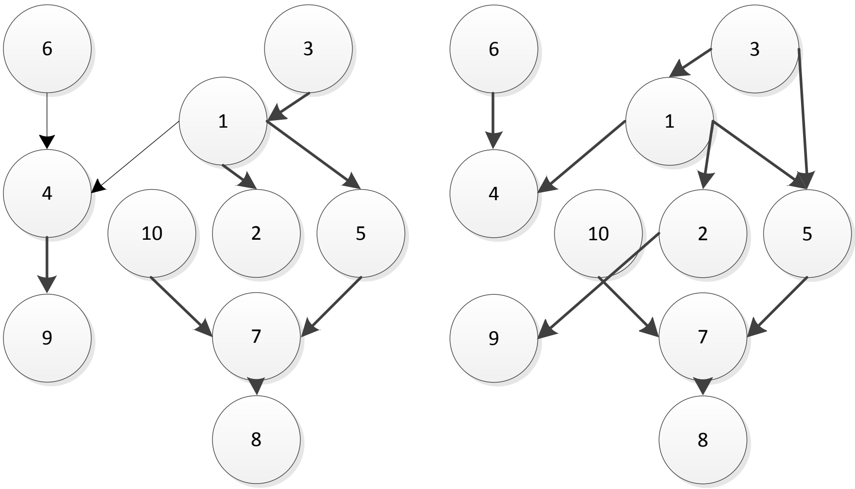

Fig. 1 Standard networks and generated networks

The left graph shows the standard structure of the network, which has 10 edges of 10 genes; the right graph shows the network structure generated by the CMICRAT algorithm.

Compared with the standard network of the nodes, the number of correct edges is 9, the number of redundant edges is 1, and the unbiased edge is 1, and the relative entropy is 10 orientations. Indicating that only one side of the wrong direction.

6. Conclusion

In this paper, we use the mutual information to eliminate the redundant edges and construct the undirected graphs of gene control networks. Compared with the use of mutual information as a measure, the accuracy rate is higher. The conditional relative mean entropy is used to determine the direction of gene regulation in the network and construct the final gene regulation network. Experiments show that the CMICRAT algorithm is used to construct the topology of the gene regulation network, which has high accuracy.

References

[1]. Vijesh N, Chakrabarti S K, Sreekumar J. Modeling of gene regulatory networks: A review [J]. Journal of Biomedical Science & Engineering, 2013, 06(1):223-231.

[2]. emmens K, Dhollander T, Bie T D, et al. Inferring transcriptional modules from ChIP-chip, motif and microarray data[J]. Genome Biology, 2006, 7(5):: R37.

[3]. Friedman N, Linial M, Nachman I, et al. Using Bayesian Network to Analyze Expression Data.[J]. Journal of Computational Biology, 2000, 7(3-4):601-620.

[4]. Wang J,Chen B, Wang Y, et al. Reconstructing regulatory networks from the dynamic plasticity of gene expression by mutual information.[J]. Nucleic Acids Research, 2013, 41(8):395-408.

[5]. Margolin A A, Nemenman I, Basso K, et al. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context[J]. BMC Bioinformatics, 2006, 7(1):: S7.

[6]. Sales G, Romualdi C. parmigene--a parallel R package for mutual information estimation and gene network reconstruction.[J]. Bioinformatics, 2011, 27(13):1876-1877(2).

[7]. Zhang X, Zhao X M, He K, et al. Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information.[J]. Bioinformatics (Oxford, England), 2012, 28(1):98-104.

[8]. Jiang J, Wang J, Yu H, et al. Poison Identification Based on Bayesian Network: A Novel Improvement on K2 Algorithm via Markov Blanket[M]// Advances in Swarm Intelligence. Springer Berlin Heidelberg, 2013:173-182.

[9]. Altman E I. Predicting financial distress of companies: revisiting the Z-Score and ZETAÂ® models[J]. Handbook of Research Methods & Applications in Empirical Finance, 2000.

[10]. http://dreamchallenges.org/project/dream4-in-silico-network-challenge/

[11]. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves[C]// International Conference on Machine Learning. DBLP, 2006:233-240.