STRUCTURING A GENE NETWORK USING A MULTIRESOLUTION INDEPENDENCE TEST

Takayuki Yamamoto, Tetsuya Takiguchi and Yasuo Ariki

Department of Computer and System Engineering, Kobe University, Japan
krush-groove@me.cs.scitec.kobe-u.ac.jp, {takigu, ariki}@kobe-u.ac.jp

ABSTRACT

In order to structure a gene network, a score-based approach is often used. A score-based approach, however, is problematic because by assuming a probability distribution, one is prevented from finding other dependent relationships with other genes. In this research, we structured a gene network from observed gene expression data using a multiresolution independence test and a conditional independence test, which is the non-parametric method proposed by Margaritis for learning the structure of Bayesian networks without making any probability distribution assumptions. The experimental results achieved an improvement in sensitivity of 0.05, and an improvement in specificity of 0.01.

Index Terms— gene network, Bayesian network, conditional independence test, non-parametric

1. INTRODUCTION

In recent years, one of the most important research problems in bioinformatics involves discovering out the mechanisms that form the basis of gene networks. A number of different frameworks for gene network modeling, Bayesian networks [1], Boolean networks [2] and differential equations [3], and so on, have been proposed. The gene expression data obtained from a DNA microarray is used to find the structure of gene networks, but this data generally contains a lot of noise and outliers, and the number of samples is small. In this research, we chose Bayesian networks, which enabled us to structure networks with only a few samples.

In order to structure Bayesian networks, the score-based approach [4] is often used. The score-based approach must assume a probability distribution, thus preventing one from finding other dependent relationships with other genes. The independence-based approach [5] is another method for structuring Bayesian networks without the assumption of a probability distribution, but this method is problematic because it is sensitive to noise and outliers. To solve this problem, we incorporate a multiresolution independence test [6], which enables us to structure gene networks from unreliable samples, without any effect from noise or outliers, into the method for structuring Bayesian networks.

This paper is organized as follows. In section 2 we introduce a method to structure Bayesian networks; in section 3, the multiresolution independence test and the conditional independence test using its test [7] are explained; in section 4, experiments we performed are explained; and in section 5, we summarize our research.

2. NETWORK STRUCTURING APPROACH

There are two general classes of algorithms used to structure Bayesian networks. The first is called the score-based approach [4], and the second is called the independence-based approach [5].

2.1. Score-based approach

The score-based approach employs a search in the space of all possible legal structures guided by a heuristic function. The search procedure maximizes the score, usually by hill-climbing. Other search techniques, such as a genetic algorithm, have also been used. This algorithm is problematic in that a probability distribution must be assumed and the structure tends toward the local optimum.

2.2. Independence-based approach

The independence-based approach uses the fact that the structure of a Bayesian network implies a set of conditional independence. This property is exploited by conducting a number of statistical conditional independence tests on the data and using the results to make inferences about the structure. A set of possible structures that satisfy the conditional independencies found in the data is constrained, and it is inferred that the structure is the only possible one. This algorithm is problematic because it is sensitive to noise and outliers.

3. INDEPENDENCE TEST

The independence-based approach enables us to structure Bayesian networks without the assumption of a probability

distribution, but this method is sensitive to noise and outliers. To solve this problem, we use a multiresolution independence test and a conditional independence test that makes us of the multiresolution independence test.

3.1. Multiresolution independence test

First, we describe the case for testing independence at a single, fixed resolution. We denote the resolution as $R \equiv I \times J$, and divide the scatter plot of variables X and Y into $I \times J$ domains. Also, we denote the counts of each domain as $c_1, \cdots, c_K$, $K \equiv IJ$, the sample size of data set as N, the probability of each cell as $p_1, \cdots, p_K$, and the set of grid boundaries along the axes as $\mathbf{B}_R$. The probability of the data set $\mathbf{D}$ is the likelihood of the cell counts, namely,

$$Pr(\mathbf{D}|p_{1\cdots K}, \mathbf{B}_R, R) = N! \prod_{k=1}^{K} \frac{p_k^{c_k}}{c_k!} \tag{1}$$

Since the parameter p_k is unknown, we use a prior distribution $Pr(p_{1\cdots K})$ to cover for parameter p_k. We choose the Dirichlet distribution, which is conjugated prior to the multinomial, as a prior distribution.

$$Pr(p_{1\cdots K}) = \Gamma(\gamma) \prod_{k=1}^{K} \frac{p_k^{\gamma_k - 1}}{\Gamma(\gamma_k)}. \tag{2}$$

where $\gamma = \Sigma_{k=1}^{K} \gamma_k$ and $\Gamma(x)$ is the gamma function. The positive numbers $\gamma_{1\cdots K}$ of this distribution are its hyperparameters. Given Eq. (1), (2), we get

$$\begin{aligned} Pr(\mathbf{D}) &= \int Pr(\mathbf{D}|p_{1\cdots K}) Pr(p_{1\cdots K}) dp_{1\cdots K} \\ &= \frac{\Gamma(\gamma)}{\Gamma(\gamma + N)} \prod_{k=1}^{K} \frac{\Gamma(\gamma_k + c_k)}{\Gamma(\gamma_k)} \end{aligned} \tag{3}$$

When assuming that our data have been produced by one of two classes of models, one representing independence (M_I) and one not ($M_{\neg I}$), we get

$$\begin{aligned} Pr(\mathbf{D}) &= Pr(\mathbf{D}|M_I) Pr(M_I) \\ &+ Pr(\mathbf{D}|M_{\neg I}) Pr(M_{\neg I}). \end{aligned} \tag{4}$$

We denote model M_I's prior probability as $Pr(M_I) \equiv \rho$, and $M_{\neg I}$'s prior probability as $Pr(M_{\neg I}) = 1 - \rho$. Eq. (4) is transformed by Bayes' theorem to get

$$Pr(M_I \mid \mathbf{D}) = 1 / \left[1 + \frac{1 - \rho}{\rho} \frac{Pr(\mathbf{D}|M_{\neg I})}{Pr(\mathbf{D}|M_I)} \right]. \tag{5}$$

The $Pr(\mathbf{D} \mid M_{\neg I})$ of the dependent model that contains IJ parameters is given by Eq. (6),

$$\begin{aligned} Pr(\mathbf{D}|M_{\neg I}) &= \frac{\Gamma(\gamma)}{\Gamma(\gamma + N)} \prod_{k=1}^{K} \frac{\Gamma(\gamma_k + c_k)}{\Gamma(\gamma_k)} \\ &\equiv \boldsymbol{\Upsilon}(\mathbf{C}_K, \gamma_K). \end{aligned} \tag{6}$$

For the independent model, we assume two multinomial distributions, one each along the X and Y axes, that contain J and I parameters, respectively. The data likelihood is given by Eq. (7).

$$\begin{aligned} Pr(\mathbf{D}|M_I) &= \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{i=1}^{I} \frac{\Gamma(\alpha_i + c_i)}{\Gamma(\alpha_i)} \\ &\times \frac{\Gamma(\beta)}{\Gamma(\beta + N)} \prod_{j=1}^{J} \frac{\Gamma(\beta_j + c_j)}{\Gamma(\beta_j)} \\ &\equiv \boldsymbol{\Upsilon}(\mathbf{C}_I, \alpha_I) \boldsymbol{\Upsilon}(\mathbf{C}_J, \beta_J) \end{aligned} \tag{7}$$

In Eq. (6) and (7), $\alpha = \Sigma_{i=1}^{I} \alpha_i$, $\beta = \Sigma_{j=1}^{J} \beta_j$, and $\gamma = \Sigma_{k=1}^{K} \gamma_k$. Assuming the Dirichlet distribution is uniform, we choose $\alpha_i = \beta_j = \gamma_k = 1$ for all i, j, k. Given Eq. (5), (6) and (7), we get the formula for the posterior probability of independence at resolution R.

$$Pr(M_I|\mathbf{D}) = 1 / \left[1 + \frac{(1 - \rho) \boldsymbol{\Upsilon}(\mathbf{C}_K, \gamma_K)}{\rho \boldsymbol{\Upsilon}(\mathbf{C}_I, \alpha_I) \boldsymbol{\Upsilon}(\mathbf{C}_J, \beta_J)} \right] \tag{8}$$

Then, we employ a Bayesian approach and average over the possible choices, weighted by their posterior.

$$\begin{aligned} Pr(M_I|R_{max}, \mathbf{D}) = \int & Pr(M_I|\mathbf{B}_R, R_{max}, \mathbf{D}) \\ & Pr(\mathbf{B}_{R_{max}}|R_{max}, \mathbf{D}) d\mathbf{B}_{R_{max}} \end{aligned} \tag{9}$$

To compute the inner integral, we should ideally average over all possible histogram boundary placements along the X and Y axes. We assume a uniform prior distribution $Pr(\mathbf{B}_R \mid R)$ over the grid boundary placement.

3.2. Conditional independence test

Our procedure for testing for the conditional independence of X and Y given Z can be summarized in the following three steps:

1. Subdivide the Z axis into m bins resulting in a partition of the data set $\mathbf{D}$ of size N into $\mathbf{D}_1, \mathbf{D}_2, \cdots \mathbf{D}_m$.

2. Measure the conditional independence in each bin by performing an independence test for X and Y ($Pr(M_I \mid \mathbf{D})$), using the multiresolution independence test.

3. Combine the conditional independence from each bin into a single number.

In these steps, the basis of testing for conditional independence in each bin and the algorithm of testing for conditional independence, called the recursive-median algorithm, are described below.

3.2.1. Testing for conditional independence in each bin

Theorem 1 $If(\{X,Y\} \perp Z),\ then\ (X \perp Y \mid Z)\ if\ and\ only\ if (X \perp Y)$.

We can use the above theorem as follows. If $Pr(\{X,Y\} \perp Z \mid \mathbf{D}_i) = 1$, the conditional independence of X and Y given Z is the same as $Pr(X \perp Y \mid \mathbf{D}_i)$, according to the theorem. If $Pr(\{X,Y\} \perp Z \mid \mathbf{D}_i) = 0$, nothing can be said about the conditional independence of X and Y given Z without actually conducting a conditional test. This is because the distribution of $\{X,Y\}$ is certain to change with Z within the bin, making the theorem inapplicable. Therefore, in this case, the posterior is taken equal to the prior probability $Pr(X \perp Y \mid Z, \mathbf{D}_i) = \rho = 0.5$. From the above fact, denoting $(\{X,Y\} \perp Z) \equiv U$, $(X \perp Y) \equiv I$, $(X \perp Y|Z) \equiv CI$, using the theorem of total probability, $Pr(CI \mid \mathbf{D}_i)$ is:

$$\begin{aligned} Pr(CI|\mathbf{D}_i) &= Pr(CI|U,\mathbf{D}_i)Pr(U|\mathbf{D}_i) \\ &+ Pr(CI|\neg U,\mathbf{D}_i)Pr(\neg U||\mathbf{D}_i) \\ &= Pr(I|\mathbf{D}_i)Pr(U|\mathbf{D}_i) \\ &+ \rho(1 - Pr(U|\mathbf{D}_i)) \end{aligned} \quad (10)$$

This test can be used for both $Pr(U \mid \mathbf{D}_i)$ and $Pr(I \mid \mathbf{D}_i)$ since they are not conditional. Therefore, we now have a way of estimating the posterior probability of conditional independence from the results of two unconditional independence tests, in each bin of a given discretization of the Z-axis.

3.2.2. Recursive-median algorithm

The recursive-median algorithm is shown in Fig. 1. The names of variables X, Y and Z and a data set $\mathbf{D}$ are input into the algorithm. It starts by calculating the measure of posterior probability of independence I and U using a single interval along the Z-axis that contains the entire data set. It then splits the data set along the Z-axis at the median, producing two non-overlapping intervals containing the same number of points and recursively calculates the above process for each of the two subsets. When only one point remains, the recursion reaches its base case. In this case, 0.5 is returned both for I and U, since both the independent and dependent model are supported by the single data point equally well. The recursive calculation results, I_1, I_2 and U_1, U_2 are then combined into I' and U'. At the end of the run on the entire data set, the returned value of U can be discarded or used as a measure of confidence in the main result, if desired. We conclude conditional independence if and only if $I/\rho \geq 1$.

(I, U) = Recursive-Median $(X, Y, Z, \mathbf{D})$

if $\mid \mathbf{D} \mid \leq 1$
  return$(0.5, 0.5)$
U = Pr$(\{X,Y\} \perp Z \mid \mathbf{D})$
I = Pr$(X \perp Y \mid \mathbf{D}) \times U + \rho \times$ (1 - U)
$z^* = median\ (\mathbf{D}, Z)$
$\mathbf{D}_1$ = { points j of $\mathbf{D}$ such that $z_j \leq z^*$ }
$\mathbf{D}_2$ = { points j of $\mathbf{D}$ such that $z_j > z^*$ }
(I_1, U_1) = Recursive-Median $(X, Y, Z, \mathbf{D}_1)$
(I_2, U_2) = Recursive-Median $(X, Y, Z, \mathbf{D}_2)$
$f_1 = (z^* - z_{min})/(z_{max} - z_{min})$
$f_2 = (z_{max} - z^*)/(z_{max} - z_{min})$
$I' = exp(f_1 \ln I_1 + f_2 \ln I_2)$
$U' = exp(f_1 \ln U_1 + f_2 \ln U_2)$
if $U > U'$
  return$(0.5, 0.5)$
else
  return$(0.5, 0.5)$

Fig. 1. The recursive-median algorithm

4. EXPERIMENTS

We structured a gene network from the observed gene expression data of a yeast cell cycle obtained from GEO (Gene Expression Omnibus). We use only genes on the cell cycle pathway (Fig. 2) in KEGG (Kyoto Encyclopedia of Genes and Genomes) Database. The score-based approach, the independence-based approach, and the independence-based approach using the multiresolution independence test (the network structured by this method is shown in Fig. 3.) were used for the method to structure a gene network and these methods were compared (Fig. 4) based on sensitivity (the number of correctly achieved edges divided by the number of target edges) and specificity (the number of correctly achieved non-edges divided by the number of target non-edges). We used hill-climbing for the score-based approach's search method, a Bayesian information criterion (BIC) for the score-based approach's score, and the correlation coefficient for the independence-based approach's independence. In Fig. 4, the blue bar denotes the results of the score-based approach, the red bar denotes the results of the independence-based approach, and the green bar denotes the results of the independence-based approach using the multiresolution independence test.

Based on the results, the sensitivity and specificity of the independence-based approach using the multiresolution independence test achieved the highest value. These results indicate that the multiresolution independence test enables us to structure gene networks from unreliable samples without effect of noise or outliers.

5. SUMMARY

We incorporated a multiresolution independence test into the method used to structure a Bayesian network, structured a

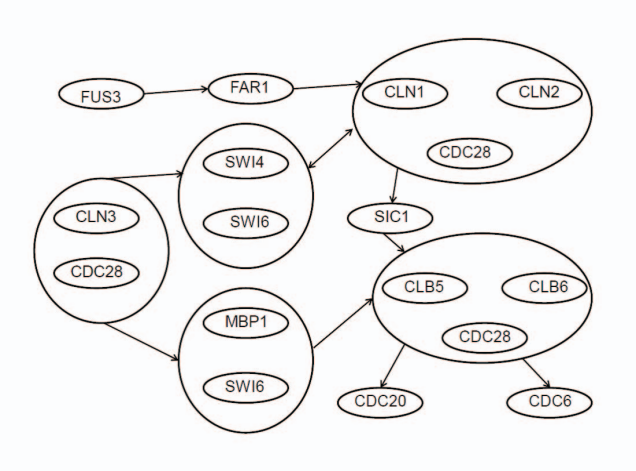

Fig. 2. Target network

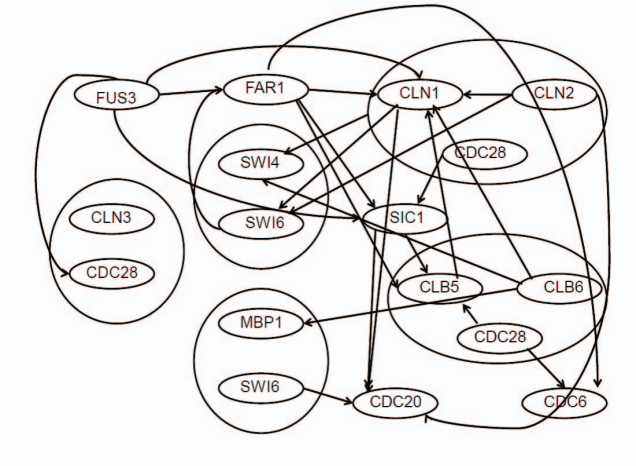

Fig. 3. Network constructed by independence-based approach and multiresolution independence test

gene network from the observed gene expression data, and compared this approach with existing methods. The results show an improvement in the network structured by the proposed method, proving the new method's potential as a gene network structuring method.

6. REFERENCES

[1] Friedman, N., Linial, M., Nachman, I. and Pe'er, D.: " Using Bayesian Networks to Analyze Expression Data", *Journal of Computational Biology*, Vol. 7, pp. 601-620 (2000).

[2] Akutsu, T., Miyano, S. and Kuhara, S.: " Algorithm for identifying Boolean networks and related biological networks based on matrix multiplication and fingerprint function ", *Journal of Computational Biology*, Vol. 7, pp. 331- 343 (2000).

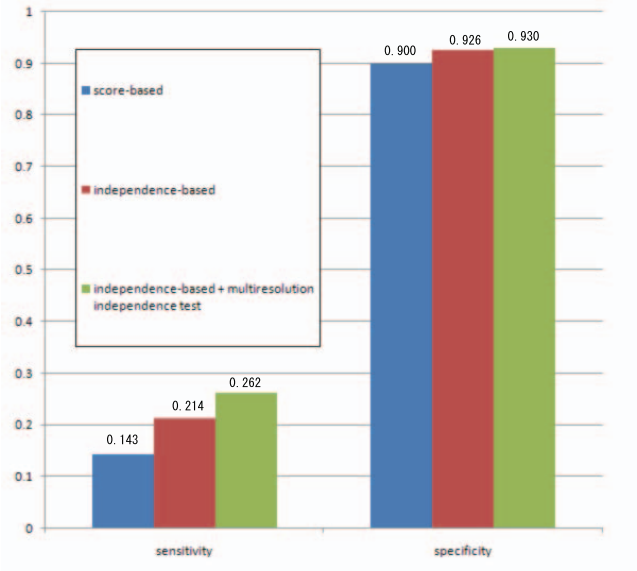

Fig. 4. Comparison of the three methods

[3] Chen, T., He, H. L. and Church, G. M.: "Modeling gene expression with differential equations ", *Proc. Pacific Symposium on Biocomputing*, pp. 29-40 (1999).

[4] Lam, W. and Bacchus, F.: " Learning Bayesian belief networks: an approach based on the MDL principle ", *Computational Intelligence*, Vol. 10, pp. 269-293 (1994).

[5] Spirtes, P., Glymour, C., and Scheines, R.: " Causation, Prediction ", and Search. Adaptive Computation and Machine Learning Series, *MIT Press*, 2nd edition (2000).

[6] Margaritis, D. and Thrun, S.: " A Bayesian Multiresolution Independence Test for Continuous Variables ", *Proc. Uncertainty in Artificial Intelligence* (*UAI*), pp. 346-353 (2001).

[7] Margaritis, D.: "Distribution-Free Learning of Bayesian Network Structure in Continuous Domains ", *Proc. the Twentieth National Conference on Artificial Intelligence* (*AAAI*), pp. 825-830 (2005).