Comparison of Gene Co-expression Networks and Bayesian Networks

Saurabh Nagrecha[1], Pawan J. Lingras[2], and Nitesh V. Chawla[1]

[1] Department of Computer Science and Engineering,
University of Notre Dame, Indiana 46556, USA
[2] Department of Mathematics and Computing Science, Saint Mary's University,
Halifax, Nova Scotia, Canada B3H 3C3

Abstract. Inferring genetic networks is of great importance in unlocking gene behaviour, which in turn provides solutions for drug testing, disease resistance, and many other applications. Dynamic network models provide room for handling noisy or missing prelearned data. This paper discusses how Dynamic Bayesian Networks compare against coexpression networks as discussed by Zhang and Horvath [1]. These shall be tested out on the genes of yeast *Saccharomyces cerevisiae*. A method is then proposed to get the best out of the strengths of both models, namely, the causality inference from Bayesian networks and the scoring method from a modified version of Zhang and Horvath's method.

1 Introduction

Biological processes, and by extension life, emerge from processes at the most basic level of the cellular structure- genes and proteins. A highly structured system of networks is responsible for information flow through the cell.

The central dogma of biology suggests mechanisms of information transfer in biological networks. This requires for us to consider genes, proteins, and their mutual interactions. DNA replication, transcription and translation are a few of these processes via which information is transferred. Gene coexpression analysis aims to provide increasingly reliable interaction models of biological systems. We restrict our model to that of a system of genes interacting with each other via expression. The nodes represent the individual genes, edges represent interactions within the system. These networks may be directed or undirected, cyclic or acyclic.

Gene expression studies usually start with microarray experiments where the expression levels of thousands of genes can simultaneously be measured. Microarray gene expression experiments are done with specimens of known heritage. These are exposed to a controlled environment with variables like nutrition, illumination, presence of various concentration of drugs. These experiments typically generate large matrices of gene expression levels. This data is usually quite noisy and may have missing values.

This data is then used to answer questions about regulatory mechanisms of gene expression. The authors demonstrate the performance of Bayesian Networks

A. Selamat et al. (Eds.): ACIIDS 2013, Part I, LNAI 7802, pp. 507–516, 2013.

as compared to coexpression networks, validated against curated gene interaction data.

2 Literature Review

A number of sophisticated methods which answer specific questions have been developed and proposed through the past two decades. Groundbreaking work by Spellman *et al.* in 1998 [2] on yeast genes using microarray hybridization techniques opened the field of systems biology and made it possible to perform scalable operations on genetic datasets. Applications ranging from the humble yeast to the Human Genome Project ultimately aim to create a "Rosetta Stone" to decipher the mystery that biological systems pose [3].

Approaches using Boolean Networks [4][5], and the next logical step Artificial Neural Networks have been proposed [6]. Methods using independent component analysis, and then self organized maps were used by Dragomir [7] were employed to solve the problem of class discovery.

The model used here builds on the approach by Murphy and Mian in [8]. Their method deals with Bayesian (belief) Networks as discussed in [9]. It unifies and generalizes models of boolean networks, Hidden Markov Models, and other widely accepted models. Boolean networks and Hidden Markov Models can be shown to be interconvertible with suitable assumptions of an intermediate state vector. Markov chains come associated with an inherent transition matrix (T) and if $T(i, j) = 0$, then this means that the system cannot make the transition from state i to state j. This kind of representation is unsuitable for sparse, discrete models- the kind we're considering here. So, we do not consider Boolean Networks or HMMs.

The use some or all of the aforementioned methods in yeast genes (those of the *Saccharomyces cerevisiae*) is of specific interest because it is fully sequenced, and widely researched. Bilu and Linial's [10] work proposes a hierarchical clustering through the metric "BLAST" which is a measure of similarity in genes. A functional prediction is then performed so as to validate the clustered genes.

Yeast genes are studied using Bayesian Networks by Friedman, *et al* in [11]. This Bayesian Network is put through a validation of known experimental results. The procedure is suited to cell cycle expressions and is thus of direct importance to our proposed method.

The system of coexpression networks inferred via a modification of the methods by Zhang and Horvath [1] for each timeframe is considered as an instance in a Markov Chain. This is then collapsed into a Bayesian Network (as justified above) using the networks discussed by Friedman *et al* in [11].

3 Study Data and Experimental Design

3.1 Study Data

As a consequence of the extensive nature of DNA microarray experiments, a "genomic" viewpoint on gene expression is provided. Data from microarray experiments on *Saccharomyces cerevisiae* by Spellman *et al.* is used here to demonstrate the methods proposed. This dataset contains 76 gene expression measurements of mRNA levels of 6177 *S. cerevisiae* ORFs. This data represents six time series under different cell cycle synchronization methods. According to Spellman *et al.* about 800 genes exist whose expression varies over different stages of the cell cycle. This data contains about 6% missing values which shall be dealt with slightly differently in the methods discussed below.

This data contains real values from the experiments. Usually, this is discretized for most purposes into 3 categories: *underexpressed* (-1) *baseline/normal* (0) and, *overexpressed* (1), depending on whether the gene is expressed lower than, similar to, or greater than the control, respectively. The thresholds for such discretization may be arrived at by setting the average from across the experimental data or from other independent experiments.

3.2 Coexpression Networks

Coexpression networks treat each gene as an individual node and connections between two such nodes depict the nature of interaction between the two genes. These interactions depend on the complexity of the model chosen. For instance, one could choose binary edges to denote presence (edge weight=1) or absence (edge weight=0) of interaction. Softer thresholding methods enable us to define weighted edges in the coexpression network. Adjacency functions which return such weights need to be defined accordingly. The parameters for these are sought using biologically motivated criteria, viz. the scale-free topology criterion[12,13].

Measures of Gene Similarity. Data is often taken in the form of raw expression levels where missing data usually results in loss of valuable information. A modified version of the data is used in this case instead. Exploiting the temporal nature of the time-series data, a noise eliminating curve-fit is implemented to take care of missing values and to smoothen out noisy kinks. This results ina relatively noiseless and more reliable correlation score. The similarity between each pair of genes is denoted by the measure s_{ij}. The absolute value of the Pearson correlation coefficient $s_{ij} = |cor(i, j)|$, or a shifted-and-scaled version $s_{ij} = \frac{1+cor(i,j)}{2}$ of it are often used here. The aim is to arrive at a similarity measure lying between 0 and 1. The similarity matrix thus arrived at, is denoted by $S = [s_{ij}]$

The Adjacency Function. To transform the similarity matrix into an adjacency matrix, the adjacency function is applied. The choice of the adjacency function decides whether we have soft (resulting in a weighted network) or hard thresholding (resulting in an unweighted network). The adjacency function is required to be a monotonically increasing function which maps the interval [0,1] into [0,1]. Hard thresholding for example works as below:

$$a_{ij} = signum(s_{ij}, \tau) \equiv \begin{cases} 1 \text{ if } s_{ij} \geq \tau \\ 0 \text{ if } s_{ij} < \tau \end{cases}$$

Soft thresholding is implemented so as to mitigate the loss of information incurred by hard thresholding. Two types of soft thresholding methods are often used: The sigmoid function

$$a_{ij} = sigmoid(s_{ij}, \alpha, \tau 0) \equiv \frac{1}{1+e^{-\alpha(s_{ij}-\tau 0)}}$$

and the power adjacency function

$$a_{ij} = power(s_{ij}, \beta) \equiv |s_{ij}|^{\beta}$$

Opinion on methods of estimating parameters for these functions varies widely. Methods suggesting the usage of p-value instead of the correlation coefficient in order to impose a hard threshold are commonly used. For soft thresholding methods, a detailed treatment using scale-free topology criteria is shown in [1]

Node Similarity/Dissimilarity. The coexpression analysis aims to identify tightly connected subsets of nodes. Out of many dissimilarity measures defined by authors, the toplogical overlap of two nodes [14] was shown to be useful in biological networks. For unweighted networks, the measure can be shown as below:

$$\omega_{ij} = \frac{l_{ij}+a_{ij}}{min\{k_i,k_j\}+1-a_{ij}}$$

where $l_{ij} = \sum a_{iu}a_{uj}$ and $k_i = \sum a_{iu}$. This may as well be extended to weighted networks. Here, in the case of $\omega_{ij} = 1$, the node with the lesser degree satisfies two conditions: 1) all of its neighbors are also neighbors of the other node and 2) it is connected to the other node. On the contrary, $\omega_{ij} = 0$, if i and j are unconnected and the two nodes do not share any neighbors. The topological overlap matrix is thus $\Omega = |\omega_{ij}|$ and is non-negative and symmetric. The dissimilarity measure is simply $d^{\omega}_{ij} = 1 - \omega_{ij}$. This matrix is the one which leads to distinctly clustered gene modules.

3.3 Bayesian Networks

Friedman *et al.*'s method treats the data with no prior assumptions of biological knowledge. It initially treats the measurements as independent samples from a distribution, ignoring the temporal aspect of the measurement. This is compensated by introducing an additional variable to denote the cell cycle phase. This variable is of key significance in all the networks learned and is forced to be a root in all the networks learned. This translates to the expression levels of the genes being dependent on the cell cycle phase.

Mathematical Formalism. A Bayesian Network is a representation of a joint probability distribution, comprising two components: the topological component G is a directed acyclic graph (DAG) whose vertices correspond to the random variables $X_1, ..., X_n$ and the second being Θ , the conditional distribution for each variable, given its parents in G. These components combined form a unique distribution in the space of the random variables $X_1, ..., X_n$. In association with the $Markov\ Assumption$, which states that each variable X_i is independent of its nondescendants, given its parents in G, the graph is a compact representation of the joint probability distribution, thus economizing on the number of parameters. The chain rule of probabilities and properties of conditional independencies help us decompose this into the $productform$ as below:

$$P(X_1, ..., X_n) = \prod P(X_i|\mathbf{Pa}^G(X_i))$$

where $\mathbf{Pa}^G(X_i)$ is the set of parents of X_i in G. The conditional distributions $P(X_i|\mathbf{Pa}^G(X_i))$ for each variable X_i are denoted by parameters specified in θ.

In specifying the conditional probability distributions, it is usual for one to represent the input random variables as continuous, discrete or mixed (in keeping with our initial mention of how the expression data is represented and subsequently interpreted). Continuous variables are usually represented using multivariate linear Gaussian distributions as $P(X|u_1, ..., u_k \sim N(a_0 + \sum a_i \cdot u_i, \sigma^2))$. Here the normally distributed random variable X's mean linearly depends on the values of its parents. If all the variables have similar Gaussian conditional distributions, then the joint distribution is a multivariate Gaussian. Discrete variables can be represented by multinomial distributions. This makes the free parameters exponential in the number of parents. Hybrid networks contain a mixture of continuous and discrete variables. These are of little relevance here and discussed in greater detail in Friedman *et al.*'s paper.

Learning Bayesian Networks. Learning a Bayesian Network can be stated as a problem as follows. Given a training set D=$\{X_1, ..., X_n\}$ of independent instances of $\mathcal{X}$, find a network $B = < G, \Theta >$ that best matches D. The problem is that of an optimization in the space of directed acyclic graphs in. The number of such graphs is superexponential in the number of variables involved.

An algorithm by Friedman, Nachman and Pe'er, called the *Sparse Candidate* algorithm is an efficient search procedure which focuses on certain relevant regions of the search space. We can identify a few key candidate parent node genes based on local statistics (like correlation) and then restrict the search to only those networks where these candidates are the parent nodes, thus slicing down the search space considerably and quickly. Possible over-restriction of the search space can be dodged by implementing an iterative search for the optimum set of initial candidates. The descendents are then found after ascertaining these candidates [8].

3.4 Experimental Design

The models arrived at using the above discussed methods are compared based on their performance with respect to biological data that has been experimentally validated. This provides a comparative study of how robust each method is, both individually, as well as when compared to each other. The performance may be based on the models' ability to identify important topological structures and causal patterns. Some of which are described below:

Dominant Genes. Genes directly involved in initiation and control of cellular cycles can be perceived as nodes of prime importance. Accuracy in predicting properties of such nodes is a cursory measure of robustness of the model.

Prominent Motifs. Extending the previous argument further topologically, it could be said that not just key nodes, but some motifs play an important role in cellular processes. The degree of isomorphism of the cell cycles (which would appear as subgraphs) identified by the methods with respect to the original motifs in the pre-validated data could be considered a more sophisticated extension of the above metric.

Markov Relations. Functional relations which make biological sense can be inferred using Markov chains as the criteria of identification. Functionally related genes which can be represented as Markov chains are an important feature. High confidence Markov relations have been known to concur with experimental validation. More interestingly, among high confidence Markov chains, one can often find conditional independence i.e. a group of highly correlated genes. [11]

Due to the extensive research performed on *Saccharomyces cerevisiae*, such data is readily available in the works of Spellman, etc. Thus the two methods are pitted head to head against each other. The validation is carried out for the most prominent genes in the organism and subsequent inferences are made.

4 Implementation

A sample data-set consisting of 12 key genes 77 states of observation (with no missing values) was used to test out most of the methods. This dataset required no cleanup due to the choice of the genes. The main network inference was carried out with the data of 6178 genes with 77 different observation states. This data contains numerous missing values.

4.1 Data Cleaning

In inferring a Bayesian Network out of such an incomplete database as the *S.cerevisiae* expression data, one is presented with a choice between ignoring the missing values and adapting dynamic models. A compromise between these two methods is chosen by using a modified dataset which now comprises of 58 observation instances instead of the original 77, the upper quartile of the missing values being omitted. This subset is further trimmed to provide a complete dataset without missing data for any of the constituent genes. This elminates the requirement to assume the presence of any further fictitious nodes and simplifies the complexity of the problem significantly.

Prevalidated interaction data is obtained from the “Saccharomyces Genome Database”. The data of relevance here is in the form of arcs constituting respective gene names. This contains a few repeated interactions, which are eliminated.

4.2 Bayesian Networks

The Bayesian networks can be formed using two approaches- score based structure learning algorithms like hill-climbing and Tabu search, and using constraint based structure learning algorithms. Constraint-based algorithms offer flexibility in setting thresholds for false positive, Type I statistical errors; but at the same time their execution time is greater than that of score based structure learning algorithms. Under similar values of α, the various constraint based algorithms seem to perform equally in terms of predicting arcs. A precision-recall analysis is done by observing the effect of varying α.

4.3 Coexpression Networks

The coexpression networks are formed using Pearson Correlation Coefficient or Mutual Information based scores using a dynamically derived β (not to be confused with the Type II statistical error). This method, unlike Bayesian Network inference, is robust to missing data (as long as the data stays within statistically significant margins). Here we adhere to the Pearson Correlation Coefficient based methods and implement hard thresholding using random resampling methods by bootstrapping the data. In doing so, we assume a Gaussian-like distribution of the expression values (which it does resemble closely). The final clusters themselves are of little significance for this particular analysis and we focus merely on the edges obtained.

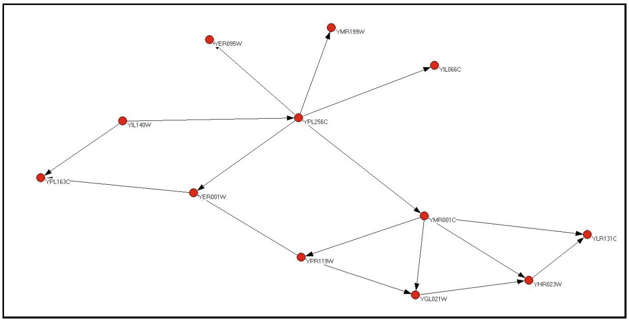

Fig. 1. The network formed out of the 12 genes considered. This Bayesian Network reflects the causal relationships exhibited in the gene cycles.

5 Inferences and Conclusions

5.1 Preliminary Inferences

Networks made using the toy dataset using coexpression analysis and Bayesian Network models concur in their predictions and with the validated dataset. This may be attributed to the small sample size of the subset chosen. The network exhibited is as per Fig. 1.

The final network consisting of 2361 genes and 7182 edges are displayed in the Pajek visualization as shown in Fig.2.

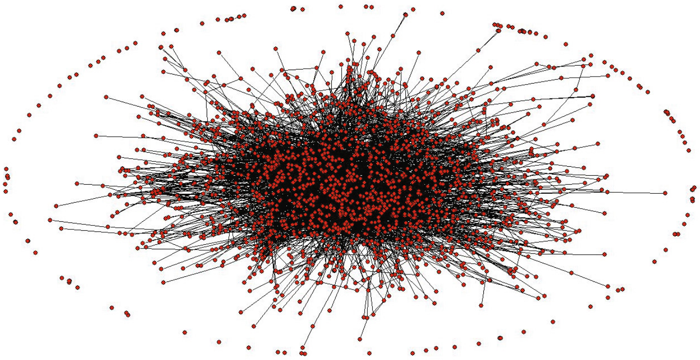

Fig. 2. Network comprising 2361 genes of *S.cerevisiae*. A Pajek visualization of the coexpression network.

Results of the Bayesian networks reveal the following key Markov relations in particular as per Table 2.

Table 1. Performance statistics for the different models. Note that the terms "precision" and "recall" are used in the context of the Bayesian Network being the relevant data and the coexpression network is the retrieved data.

Set A	*Set B*	*Precision*	*Recall*
Validated Data	Bayesian Network	0.95	0.67
Validated Data	Coexpression Network	0.62	0.53
Bayesian Network	Coexpression Network	0.83	0.77

5.2 Validation and Inferences

We perform a precision-recall analysis between the 3 set of arcs: the prevalidated data, the Bayesian Network, the coexpression network. (as shown in Table 1) The Bayesian Network is seen to be a closer estimater of gene interactions than the coexpression network due to superior precision and relatively higher recall. Relations that are prominently expressed in the data appear in all 3 models. The biological interpretations of the interactions are either spatial or functional in nature. Despite this, few of the high confidence functional interactions predicted may be considered false positives if arrived at using a Gaussian model, as it uses correlation values. This problem does not arise in the multinomial model, whose salient results are as per Table 2.

Table 2. List of Top Markov Relations, Multinomial Experiment

Confidence	*Gene 1*	*Gene 2*	*Notes*
1.0	YKL163W-PIR3	YKL164C-PIR1	Close locality on chromosome
0.985	PRY2	YKR012C	Close locality on chromosome
0.985	MCD1	MSH6	Both bind to DNA during mitosis
0.98	PHO11	PHO12	Both nearly identical acid phosphatases
0.975	HHT1	HTB1	Both are histones
0.97	HTB2	HTA1	Both are histones
0.94	YNL057W	YNL058C	Close locality on chromosome
0.94	YHR143W	CTS1	Homolog to EGT2 cell wall control, both involved in cytokinesis
0.92	YOR263C	YOR264W	Close locality on chromosome
0.91	YGR086	SIC1	Homolog to mammalian nuclear ran protein, both involved in nuclear function
0.9	FAR1	ASH1	Both part of a mating type switch, expression uncorrelated
0.89	CLN2	SVS1	Function of SVS1 unknown
0.88	YDR033W	NCE2	Homolog to transmembrame proteins suggest both involved in protein secretion
0.86	STE2	MFA2	A mating factor and receptor
0.85	HHF1	HHF2	Both are histones
0.85	MET10	ECM17	Both are sulte reductases
0.85	CDC9	RAD27	Both participate in Okazaki fragment processing

5.3 Conclusions

In this paper, a comparison has been made between Bayesian Network and coexpression networks on the basis of performance in predicting the structure of the expression network of the genome for baker's yeast. This is done without any prior knowledge of biology involved; in fact, biologically viable and plausible interactions stem out of the predicted models. A more throughly biologically supervised global topological treatment has been discarded in favor of learning the finer interaction structure. Evidently, Bayesian networks emerge as a more informative tool to determine the causal structure of such interactions.

References

1. Zhang, B., Horvath, S., et al.: A general framework for weighted gene co-expression network analysis. Statistical Applications in Genetics and Molecular Biology 4(1), 1128 (2005)
2. Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Molecular Biology of the Cell 9(12), 3273–3297 (1998)
3. Ideker, T., Galitski, T., Hood, L.: A new approach to decoding life: systems biology. Annual Review of Genomics and Human Genetics 2(1), 343–372 (2001)

4. Akutsu, T., Miyano, S., Kuhara, S., et al.: Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In: Pacific Symposium on Biocomputing, vol. 4, pp. 17–28. World Scientific, Maui (1999)
5. Hakamada, K., Hanai, T., Honda, H., Kobayashi, T.: A preprocessing method for inferring genetic interaction from gene expression data using boolean algorithm. Journal of Bioscience and Bioengineering 98(6), 457–463 (2004)
6. Weaver, D., Workman, C., Stormo, G., et al.: Modeling regulatory networks with weight matrices. In: Pacific Symposium on Biocomputing, vol. 4, pp. 112–123. World Scientific, Maui (1999)
7. Dragomir, A., Mavroudi, S., Bezerianos, A.: Som-based class discovery exploring the ica-reduced features of microarray expression profiles. Comparative and Functional Genomics 5(8), 596–616 (2005)
8. Murphy, K., Mian, S., et al.: Modelling gene expression data using dynamic bayesian networks. Technical report, Computer Science Division, University of California, Berkeley, CA (1999)
9. Pearl, J.: Bayesian networks (2011)
10. Bilu, Y., Linial, M.: The advantage of functional prediction based on clustering of yeast genes and its correlation with non-sequence based classifications. Journal of Computational Biology 9(2), 193–210 (2002)
11. Friedman, N., Linial, M., Nachman, I., Pe'er, D.: Using bayesian networks to analyze expression data. Journal of Computational Biology 7(3-4), 601–620 (2000)
12. Barabási, A., Oltvai, Z.: Network biology: understanding the cell's functional organization. Nature Reviews Genetics 5(2), 101–113 (2004)
13. Rzhetsky, A., Gomez, S.: Birth of scale-free molecular networks and the number of distinct dna and protein domains per genome. Bioinformatics 17(10), 988–996 (2001)
14. Ravasz, E., Somera, A., Mongru, D., Oltvai, Z., Barabási, A.: Hierarchical organization of modularity in metabolic networks. Science 297(5586), 1551–1555 (2002)