COMPARATIVE STUDY OF GRN'S INFERENCE METHODS BASED ON FEATURE SELECTION BY MUTUAL INFORMATION

Fabrício M. Lopes[1,2]*, David C. Martins-Jr*[1] *and Roberto M. Cesar-Jr*[1]

[1]Institute of Mathematics and Statistics, University of São Paulo, Brazil
{fabriciolopes,davidjr,cesar}@vision.ime.usp.br
[2]Federal University of Technology - Paraná, Brazil
fabricio@utfpr.edu.br

ABSTRACT

Feature selection is a crucial topic in pattern recognition applications, especially in the genetic regulatory networks (GRNs) inference problem which usually involves data with a large number of variables and small number of observations. In this context, the application of dimensionality reduction approaches such as those based on feature selection becomes a mandatory step in order to select the most important predictor genes that can explain some phenomena associated with the target genes. Given its importance in GRN inference, many feature selection methods (algorithms and criterion functions) have been proposed. However, it is decisive to validate such results in order to better understand its significance. The present work proposes a comparative study of feature selection techniques involving information theory concepts, applied to the estimation of GRNs from simulated temporal expression data generated by an artificial gene network (AGN) model. Four GRN inference methods are compared in terms of global network measures. Some interesting conclusions can be drawn from the experimental results.

1. INTRODUCTION

Feature selection is one of the main approaches for dimensionality reduction that selects a small subset of the original features to represent the observed patterns. In genomic expression signals (mRNA concentrations), there are two main goals in performing feature selection [1]. One is to eliminate irrelevant genes from the classification (or prediction) in order to enhance its performance. The other is to discover the structure of the genetic networks or the mechanisms responsible for some biological phenomenon of interest (e.g. progress or repression of a disease).

The cell control is a result of a multivariate activity of genes. Thus, for disease treatment design and drugs creation purposes, there is a strong motivation for multivariate interaction modeling. Genetic regulatory networks (GRN) control a cell in the genomic level, determining the transcription of some genes to mRNA with a variable intensity (expression), which works as a model for protein synthesis [2]. Genes and proteins act together, composing complex regulatory networks. GRNs describe these regulatory processes and the molecular reaction of a cell to several stimuli. High-throughput techniques for measurement of RNA concentrations and proteins allow new approaches for the study of such networks. The analysis of these data sets requires sophisticated methods.

The genomic expressions can be derived from time series and time independent (steady state) data. This work concentrates on the analysis of feature selection techniques for inference of relationships among genes based on time series data. A myriad of feature selection algorithms and criterion functions have been proposed to infer such relationships [3–6]. However, how to validate the network identification results? One way to objectively assess such algorithms is to apply them to computational gene network models for which the mechanisms are completely known [7, 8].

In this context, we have developed an experimental comparison of feature selection methods based on information theory concepts for GRN inference purpose. The comparative study proposed here involves three methods from the package *minet* implemented in R/Bioconductor [9] and one method implemented in Java [6] which applies a classical wrapper feature selection algorithm (sequential floating forward selection - SFFS [10]) guided by penalized mean conditional entropy (MCE) [5].

The ground truth for the comparison of the four used methods aforementioned is defined by the application of the artificial genetic network (AGN) model [8], which uses theoretical models of complex networks [11–13] to define the AGN topology. The dynamics of the AGN is given by applying transition functions on an input signal. These functions are used to simulate temporal expression data. A similarity metric based on a confusion matrix [14] is taken into account to compare the inferred GRNs against the ground truth.

Next section presents a brief description of the feature selection methods used in the comparative analysis for GRNs inference. Section 3 discusses the AGN model to generate the ground truth and the simulated expression signals, while Section 4 describes the similarity measures adopted to compare the inferred and the expected networks. Some comparative results are presented and discussed in Section 5. Section 6 concludes this text, including possible directions of this work.

2. FEATURE SELECTION METHODS ADOPTED FOR COMPARATIVE ANALYSIS

We describe the assessed methods based on information theory to infer connectivity among elements in a GRN. The mutual information is a measure of shared information between two variables defined by means of the entropy (H):

$$I(X_i, X_j) = H(X_i) + H(X_j) - H(X_i, X_j) \quad (1)$$

where $H(X_i) = \sum_{x_i \in X_i} P(X_i) log P(X_i)$ and $H(X_i, X_j) = \sum_{x_i \in X_i, x_j \in X_j} P(X_i, X_j) log P(X_i, X_j)$.

We have chosen four methods which are freely available on the Internet. Three methods are available within the package *minet* under R/Bioconductor [9], while the fourth technique is available as part of a feature selection graphical environment implemented in Java [6].

2.1. CLR Algorithm

This technique extends the relevance networks approach which associates an edge between two genes X_i and X_j if the mutual information $I(X_i, X_j)$ is greater than a given threshold. In the CLR (Context Likelihood of Relatedness) algorithm [15, 16], the mutual information is computed for every pair of genes, deriving a score related to the empirical distribution of the mutual information values. This score is given by $z_{ij} = \sqrt{z_i^2 + z_j^2}$ where

$$z_i = \max\left(0, \frac{I(X_i, X_j) - \mu_i}{\sigma_i}\right), \quad (2)$$

μ_i is the mean and σ_i is the standard deviation of X_i.

2.2. ARACNE

ARACNE (Algorithm for the Reconstruction of Accurate Cellular Networks) [17] is based on the Data Processing Inequality [18], which states that, if X_i interacts with X_k through X_j, then

$$I(X_i, X_j) \leq \min(I(X_i, X_j), I(X_j, X_k)). \quad (3)$$

The method starts by applying the relevance networks approach (associating edges to pairs of genes with high mutual information [16]). After that, it analyzes each triple of genes in order to eliminate the edge with lowest mutual information in cases where the difference between the lowest and the second lowest mutual information is greater than a given threshold (indirect interaction removal).

2.3. MRNET

This method [9] employs the maximum relevance / minimum redundancy (MRMR) [19] feature selection for inference of GRNs. For every gene placed as a target (Y), it applies a sequential selection procedure in which, at each step, the partial solution set $\mathbf{Z}$ of features is updated by adding the feature X_i that maximizes the difference $u_i - r_i$, where u_i is the relevance term given by $I(X_i, Y)$ and r_i is the redundancy term given by

$$r_i = \frac{1}{|\mathbf{Z}|} \sum_{X_j \in \mathbf{Z}} I(X_i, X_j), \quad (4)$$

which analyzes the redundancy of X_i to each selected variable $X_j \in \mathbf{Z}$.

2.4. SFFS+MCE

The sequential floating forward selection [10] is a wrapper approach that selects or removes a feature according to some criterion function that evaluates subsets instead of just comparison of feature pairs. The criterion function adopted here is the mean conditional entropy by penalization of rarely observed instances [5, 6] given by

$$H(Y|\mathbf{Z}) = \frac{M - N}{s} H(U(0, |D| - 1)) + \sum_{\mathbf{z} \in \mathbf{Z}: P(\mathbf{z}) > \frac{1}{s}} P(\mathbf{z}) H(Y|\mathbf{z}), \quad (5)$$

where D is a discrete set of values assumed by the genes (e.g. 0,1 for binary case), $U(0, |D| - 1))$ is the uniform distribution function applied to the values of D, s is the number of data samples and N is the number of instances $\mathbf{z} \in \mathbf{Z}$ with $P(\mathbf{z}) > \frac{1}{s}$ (more than one observation). Because higher mean conditional entropies lead to lower mutual information, the SFFS is guided to minimize this criterion function. The SFFS+MCE is applied for every gene placed as target.

3. AGN MODEL

The AGN model [8] is composed of three main components: (1) topology, (2) network state and (3) transition functions. The topology of an AGN is defined by theoretical complex networks models [11–13]. We have adopted uniformly-random Erdös-Rényi (ER) and the scale-free Barabási-Albert (BA).

The AGN model is a complex network $G = (V, E)$ formed by a set $V = \{v_1, v_2, \ldots, v_N\}$ of nodes or "genes", connected by a set $E = \{e_1, e_2, \ldots, e_M\}$. It is important to keep the same average number of connections of nodes k for comparative analysis between ER and BA. In this way, to keep k fixed for the ER model, the probability p of connecting each pair of nodes is given by $p = \frac{k}{N-1}$. The BA topology follows a *linear preferential attachment* rule, i.e., the probability of the new node v_i to connect to an existing node v_j is proportional to the degree of v_j. Therefore, the nodes of ER networks have a random pattern of connections and BA networks have some nodes highly connected while others have few connections.

Each gene can assume a value from a discrete set D (in this work, $D = \{0, 1\}$, i.e., on/off) that represents its states. The network state s at time t is determined by $s_t = \{v_{1,t}, v_{2,t}, \ldots, v_{N,t}\}$. The transition functions F are defined by logic circuits, one for each gene of the network $v_{i,t+1} = F(u_{ki,t})$, in which $u_{ki} \in G$ represents the k genes (predictors) that have input edges to v_i (target). The transition functions are defined by considering a deterministic model [20], i.e, the networks remain fixed in the choice of k inputs nodes, chosen logic circuits and

chosen predictors, during all instants of time. The dynamics of an AGN is determined by applying the transition functions to an arbitrary initial state $s_0 = \{v_1 = 0, v_2 = 1, \ldots, v_N = 1\}$ during T time instants (signal size), i.e., the target state at time $t_{i+1}, i = 0, 1, \ldots, T-1$ is obtained by observing the predictor states at t_i and applying its respective logic circuit. As a result, we have the simulated temporal signals of length T, which are used for the network identification methods presented in Section 2.

4. VALIDATION METRIC

In order to quantify the similarity between A (AGN model networks) and B (inferred networks), we adopted the validation metric based on a confusion matrix [14] (Table 1).

Table 1. Confusion matrix. TP = true positive, FN = false negative, FP = false positive, TN = true negative.

Edge	Inferred	Not Inferred
Present	TP	FN
Absent	FP	TN

The networks are represented in terms of their respective adjacency matrices M, such that each edge from node i to node j implies $M(i,j) = 1$, with $M(i,j) = 0$ otherwise. The measures considered in this work are calculated as follows:

$$Similarity(A,B) = \sqrt{TPR \cdot TNR}$$
$$TPR = \frac{TP}{(TP+FN)}, \quad TNR = \frac{TN}{(TN+FP)}, \tag{6}$$

We consider the geometrical average $Similarity(A,B)$ between the ratios of correct ones (TPR) and correct zeros (TNR), observing the ground truth matrix A and the inferred matrix B. Observe that both coincidences and differences are taken into account by these indices, implying the maximum similarity to be obtained for indices values near 1.

5. EXPERIMENTAL RESULTS

In this section two distinct complex network models are confronted in order to analyze the importance of the topology for network inference methods. Random networks (ER) having uniform distribution on the node degree and scale-free networks (BA) having a power law distribution on the node degree have been tested. Another objetive is to investigate the impact of average degree variation in both models.

For all experiments, the two network models (BA and ER) with 100 nodes were used. The average degree k varied from 1 to 5 and the simulated temporal expression were generated using 10 randomly chosen initial states, each one with length 30. These expressions were concatenated into one single signal of size 300. The experimental results present the median obtained from 50 simulations for each network architecture and k value, using the default parameters of the methods [6, 9].

Figures 1 and 2 show the similarity (described in Section 4) between the inferred networks and AGN-based networks in terms of the average node degrees by considering, respectively, ER and BA architectures. Clearly we can observe that only SFFS+MCE method has an important decrease of the similarity rate by increasing the average degree. This behavior was expected due to the fact that the generated network has more connections for larger k values, i.e., the signal of the target gene is determined by the composition of Boolean functions from more predictors, generating sophisticated boolean functions, which are more difficult to identify. Interestingly, the MRNET method presents a slight improvement of results from $k = 1$ to $k = 4$ considering ER topology. Considering BA topology, the MRNET performance behavior is inverted, i.e., it presents a slight decreasing of the similarity rate from $k = 1$ to $k = 4$.

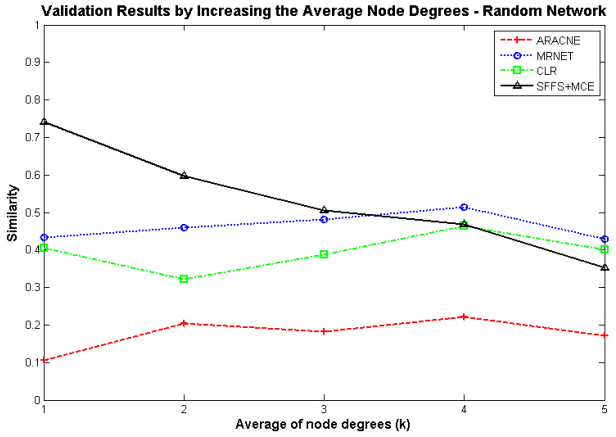

Figure 1. Network identification rate considering the increasing average node degrees, using the uniformly-random Erdös-Rényi network architecture (ER).

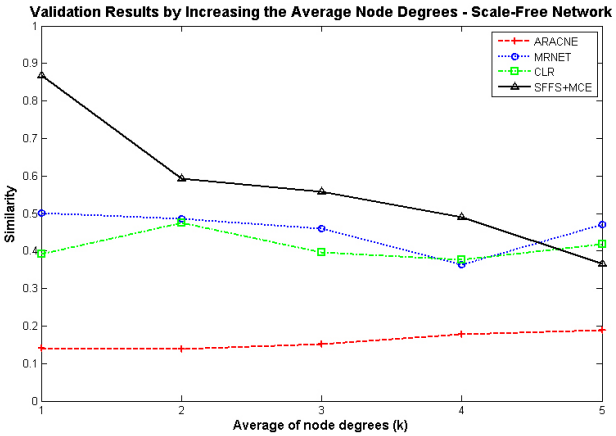

Figure 2. Network identification rate considering the increasing average node degrees, using the scale-free Barabási-Albert network architecture (BA).

For both topologies, SFFS+MCE presents best results for networks with small average input degree ($k \leq 3$ for ER and $k \leq 4$ for BA), achieving 86% of similarity for BA and 74% for ER, when $k = 1$. The MRNET performs best for large k ($k \geq 4$) for ER and $k = 5$ for BA. The CLR method presents a behavior closely related to MRNET, but presenting slightly lower similarity rates. The ARACNE method presented the lowest results in all experiments.

An important fact to take into account by analyzing these results is that dynamical systems like biological systems are in the frontier between non-chaotic and chaotic behavior. The nodes of such systems present a degree of prediction between 2 and 3 in average [21].

6. CONCLUSION

This work proposed a comparative analysis in order to evaluate four GRN inference methods based on feature selection by mutual information. We highlighted the importance of the network topology, the augmentation of average degree of nodes (complexity of the network) and the measurement of the similarity rate of network inference by these methods.

The results were obtained by applying the inference methods to the estimation of GRNs from simulated temporal expression data generated by an artificial gene network (AGN) model. The results indicate that the network topology was important for the SFFS+MCE method in terms of similarity rate. The importance of topology was also observed in other methods, especially in MRNET that presents completely opposite similarity rate behavior from one topology to the other.

Considering the average degree of nodes, SFFS+MCE presents best results for both topology networks with small average degree, while MRNET performs best for large average degree. The CLR method presents a behavior closely related to MRNET, but presenting slightly lower similarity rates. The ARACNE method presented the lowest similarity rates in all experiments. The results indicate that SFFS+MCE is more appropriate for analysis of biological systems than the other three methods compared.

This work initiates the inference analysis by using large number of time observations, which is desirable. In a further work, small number of observations will be analyzed in comparison to these experimental results. The next stage of this work is to consider complex networks measurements [13] (local and global) in order to refine the inference network analysis. Other relevant improvement is to include some uncertainty in the transition functions, i.e., to use stochastic transition functions and to measure its effect on network inference methods. Other methods also could be included in the comparative results.

7. ACKNOWLEDGMENTS

This work was supported by FAPESP, CNPq and CAPES.

8. REFERENCES

[1] I. Shmulevich and E. R. Dougherty, *Genomic Signal Processing*, Princeton Univ. Press, New Jersey, 2007.

[2] A. Kelemen, A. Abraham, and Y. Chen, *Computational Intelligence in Bioinformatics*, Springer, 2008.

[3] S. Liang, S. Fuhrman, and R. Somogyi, "Reveal, a general reverse engineering algorithm for inferenceof genetic network architectures," in *Pacific Simposium on Biocomputing*, 1998, vol. 3, pp. 18–29.

[4] S. Keles, M. van-der Laan, and M. B. Eisen, "Identification of regulatory elements using a feature selection method," *Bioinformatics*, vol. 18, no. 9, pp. 1167–1175, September 2002.

[5] J. Barrera, R. M. Cesar-Jr, D. C. Martins-Jr, R. Z. N. Vencio, E. F. Merino, M. M. Yamamoto, F. G. Leonardi, C. A. B. Pereira, and H. A. Portillo, *Methods of Microarray Data Analysis V*, chapter Constructing probabilistic genetic networks of *Plasmodium falciparum*, from dynamical expression signals of the intraerythrocytic development cycle, pp. 11–26, Springer-Verlag, 2007.

[6] F. M. Lopes, D. C. Martins-Jr, and R. M. Cesar-Jr, "Feature selection environment for genomic applications," *BMC Bioinformatics*, vol. 9, no. 451, 2008.

[7] P. Mendes, W. Sha, and K. Ye, "Artificial gene networks for objective comparison of analysis algorithms," *Bioinformatics*, vol. 19, no. Suppl 2, pp. 122ii–129, 2003.

[8] F. M. Lopes, R. M. Cesar-Jr, and L. F. Costa, "AGN simulation and validation model," in *Advances in Bioinformatics and Computational Biology*. August 2008, vol. 5167 of *Lecture Notes in Computer Science*, pp. 169–173, Springer-Verlag.

[9] P. E. Meyer, F. Lafitte, and G. Bontempi, "minet: A R/Bioconductor package for inferring large transcriptional networks using mutual information," *BMC Bioinformatics*, vol. 9, no. 461, 2008.

[10] P. Pudil, J. Novovičová, and J. Kittler, "Floating search methods in feature selection," *Pattern Recogn. Lett.*, vol. 15, no. 11, pp. 1119–1125, 1994.

[11] R. Albert and A. L. Barabási, "Statistical mechanics of complex networks," *Rev. Mod. Phys.*, vol. 74, no. 1, pp. 47–97, 2002.

[12] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[13] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, "Characterization of complex networks: a survey of measurements," *Advances in Physics*, vol. 56, no. 1, pp. 167–242, 2007.

[14] L. F. Costa, M. Kaiser, and C. C. Hilgetag, "Predicting the connectivity of primate cortical networks from topological and spatial node properties," *BMC Systems Biology*, vol. 1, no. 16, 2007.

[15] J. Faith, B. Hayete, J. Thaden, I. Mogno, J. Wierzbowski, G. Cottarel, S. Kasif, J. Collins, and T. Gardner, "Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS Biology*, vol. 5, no. 1, pp. 259–265, 2007.

[16] A. Butte and I. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," in *Pacific Symposium on Biocomputing*, 2000, pp. 418–429.

[17] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla-Favera, and A. Calfano, "Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, no. Suppl 1, 2006.

[18] T. M. Cover and J. A. Thomas, "Elements of information theory," in *Wiley Series in Telecommunications*. John Wiley & Sons, New York, NY, USA, 1991.

[19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. on PAMI*, vol. 27, no. 8, pp. 1226–1238, 2005.

[20] S. A. Kauffman, "Metabolic stability and epigenesis in randomly constructed nets," *Journal of Theoretical Biology*, vol. 22, no. 3, pp. 437–467, March 1969.

[21] S. A. Kauffman, *The Origins of Order*, Oxford University Press, 1993.