Analysis of the GRNs Inference by Using Tsallis Entropy and a Feature Selection Approach

Fabrício M. Lopes[1,2], Evaldo A. de Oliveira[1], and Roberto M. Cesar-Jr[1]

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

Abstract. An important problem in the bioinformatics field is to understand how genes are regulated and interact through gene networks. This knowledge can be helpful for many applications, such as disease treatment design and drugs creation purposes. For this reason, it is very important to uncover the functional relationship among genes and then to construct the gene regulatory network (GRN) from temporal expression data. However, this task usually involves data with a large number of variables and small number of observations. In this way, there is a strong motivation to use pattern recognition and dimensionality reduction approaches. In particular, feature selection is specially important in order to select the most important predictor genes that can explain some phenomena associated with the target genes. This work presents a first study about the sensibility of entropy methods regarding the entropy functional form, applied to the problem of topology recovery of GRNs. The generalized entropy proposed by Tsallis is used to study this sensibility. The inference process is based on a feature selection approach, which is applied to simulated temporal expression data generated by an artificial gene network (AGN) model. The inferred GRNs are validated in terms of global network measures. Some interesting conclusions can be drawn from the experimental results, as reported for the first time in the present paper.

Keywords: Tsallis entropy, feature selection, inference, validation, gene regulatory networks, bioinformatics.

1 Introduction

In general, living organisms can be viewed as networks of molecules connected by chemical reactions. More specifically, the cell control involves the activity of several related genes, in which the relationship among them is known or not. Gene regulatory networks (GRNs) are used to indicate the interrelation among genes in the genomic level [1]. Such information is very important for disease treatment design, drugs creation purposes and to understand the activity of living organisms in the molecular level. In this way, there is a strong motivation for GRNs inference.

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 473–480, 2009.

High-throughput techniques for measurement of mRNA concentrations allow the simultaneous verification of cell components activity (state) in multiple instants of time. Some methods were proposed for modeling and identification of GRNs from expression data sets [2,3,4,5,6,7,8,9,10,11]. This work focuses attention in a particular method proposed by Barrera *et al* [7] in order to evaluate the application of Tsallis entropy [12] in the GRNs inference problem by using an artificial gene network (AGN) model [13]. This technique is based on a feature selection algorithm that minimizes entropy measures as the criterion function.

The Tsallis entropy has been stood out in the last years as a generalization of the Shannon entropy [14]. This is not only due to its applications [15], but also due to its theoretical foundation [16]. Its use becomes important on systems with long-range interactions (which causes long-range correlations), a particular feature of nonextensive systems. But, are the gene regulatory networks nonextensive? How to interpret them in this context? In order to investigate these questions, the present work addresses the problem about the inference and extensivity of GRNs by applying information theory. We also analyze the quality and limitations of the adopted method [7] to infer network topologies.

Next section presents a brief description of the AGN model to generate the ground-truth and the simulated expression signals. Section 3 presents the network inference method, while Section 4 describes the similarity measures adopted to compare the inferred and the ground-truth networks. Experimental results are presented and discussed in Section 5. Section 6 concludes this text with possible future directions of this work.

2 AGN Model

The AGN model [13] is composed of three main components: (1) topology, (2) network state and (3) transition functions. The topology of an AGN may be defined by theoretical complex networks models [17,18,19]. We have adopted the uniformly-random Erdös-Rényi (ER) and the scale-free Barabási-Albert (BA) models.

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, in order 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 while BA networks have some nodes highly connected and others with 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 $s_t = \{v_{1,t}, v_{2,t}, \ldots, v_{N,t}\}$, called the state vector.

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 number of predictors and its influence (measured by edges) on target genes are defined by considering a deterministic model [20], *i.e.*, the networks remain fixed in the choice of k input 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 $s_1 = \{v_1 = 1, v_2 = 0, \ldots, v_N = 1\}$ during T time instants (*i.e.*, the signal size). The target state at time $t_{i+1}, i = 2, 3, \ldots, T$ is hence obtained by observing the predictor states at t_i and by applying its respective logic circuit. As a result, we have the simulated temporal signals of length T, which are used for the network identification method presented in Section 3.

3 Network Inference

The use of entropy functions to infer gene interaction network topology from time series has been showed a promising tool [7, 21]. Of course, the precision of the inference depends on the information available and the suitability of its use.

The inference method used in this work is described in [7], in which the entropy [14] of the temporal gene expression was employed as a criterion function in a feature selection [22] approach to identify the network topology. The main idea is to select the predictors subset that minimizes the entropy of each target gene from its expressions profiles.

In this context, network inference is modeled as a series of feature selection problems, one for each gene. The inference method starts by fixing the target gene Y. The time series determined by gene expressions are used to build a table of conditional probabilities of the classes Y given the patterns $\mathbf{X}$ that minimizes a criterion function. The classes are defined by the target values at time $t + 1$, while the patterns are defined by the predictors values at time t.

The search space is normally very large, so that an exhaustive search cannot be performed. In our approach, the *Sequential Forward Floating Search* (SFFS) [23] algorithm was applied for each target gene in order to select the subset of genes $\mathbf{X}$ that minimizes the criterion function given by Equation (1). As defined in [24], the penalized mean conditional entropy of Y given all the possible instances $\mathbf{x} \in \mathbf{X}$ is given by:

$$H(Y \mid \mathbf{X}) = \frac{\alpha(M - N)\ H(Y)}{\alpha M + s} + \frac{\sum_{i=1}^{N}(f_i + \alpha)\ H(Y \mid \mathbf{X} = \mathbf{x_i})}{\alpha M + s}, \qquad (1)$$

where M is the number of possible instances of the feature vector $\mathbf{X}$, N is the number of observed instances (the number of non-observed instances is given by $M - N$), f_i is the absolute frequency (number of observations) of $\mathbf{x_i}$ and s is the number of samples. The α constant is the penalty weight for non-observed instances ($\alpha = 1$ in the present work).

Once we are interested to better understand the method, mainly about its performance given a data set, we focus on the entropy function and use the generalized entropy proposed by Tsallis [12,25]. The Tsallis entropy has been studied and applied by many researchers in different approaches (information theory [16],

thermodynamics [26]) and systems. Its suitability has been proved [27], mainly where the Shannon is not recommended, *i.e.*, for long-range interactions between the nodes. As defined in [24], by the inclusion of such generalization in Equation (1), the conditional entropy $H(Y \mid \mathbf{X} = \mathbf{x_i})$ becomes:

$$H(Y \mid \mathbf{X} = \mathbf{x_i}) = \frac{1}{q-1}(1 - \sum_{y \in Y}(P(y \mid \mathbf{x_i}))^q), \tag{2}$$

where $P(y \mid \mathbf{x_i})$ is the conditional probability of y given the observation of an instance $\mathbf{x_i} \in \mathbf{X}$.

The Tsallis entropy generalizes the Shannon entropy and can be used as a functional set by the parameter q which is defined as an entropic parameter that characterizes the degree of nonextensivity. For $q < 1$ the entropies are superextensives and for $q > 1$ the entropies are subextensives. Furthermore, when $q = 1$ the entropies are extensive and the Shannon form is completely recovered[1].

Lower values of H yield better feature subspaces (the lower H, the larger is the information gained about Y by observing $\mathbf{X}$). In this way, the SFFS is guided to minimize the criterion function in Equation (1). The selected features are taken as predictor genes for each target gene, which are used to link the genes and thus to recover the network topology.

The next section describes the similarity measures adopted to compare the inferred and the AGN-based networks.

4 Validation

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 [28] (Table 1).

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

Edge	Inferred in B	Not Inferred in B
Present in A	TP	FN
Absent in A	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:

$$\begin{gathered} Similarity(A,B) = \sqrt{TPR \cdot TNR}\,, \\ TPR = \frac{TP}{(TP+FN)}, \quad TNR = \frac{TN}{(TN+FP)}. \end{gathered} \tag{3}$$

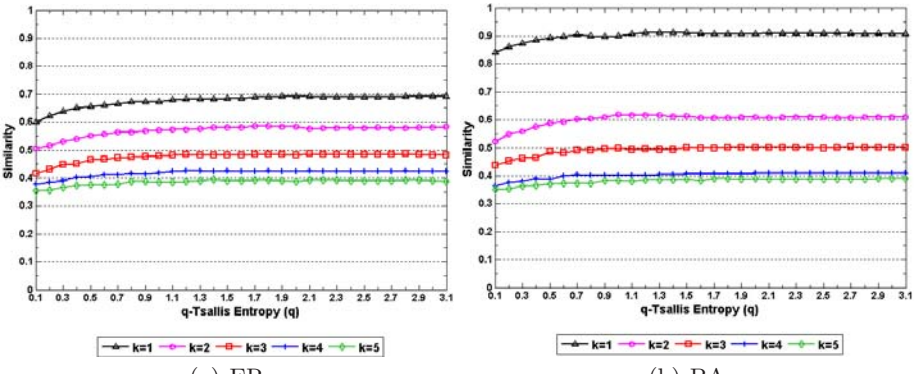
[1] This can be easily obtained by taking the limit $q \to 1$ in the Equation (2).

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 values near 1.

5 Experimental Results

This section presents the experimental results by applying Tsallis entropy [25] for GRNs inference, as presented in Section 3, by considering three main aspects: (1) Variation of parameter q of Tsallis entropy; (2) two different complex network topologies (ER) and (BA); (3) complexity of networks in terms of average node degree (k).

For all experiments, the two network models (BA and ER) with 100 nodes were used. The q parameter of Tsallis entropy varied from 0.1 to 3.1 in steps of 0.1, and the average node degree k varied from 1 to 5 in steps of 1. The simulated temporal expression was generated using 10 randomly chosen initial states, each one with length 30. These expressions were concatenated into a single signal of size 300. The experimental results were obtained from 50 simulations for each network topology and k value, using the default parameters of the method [24].

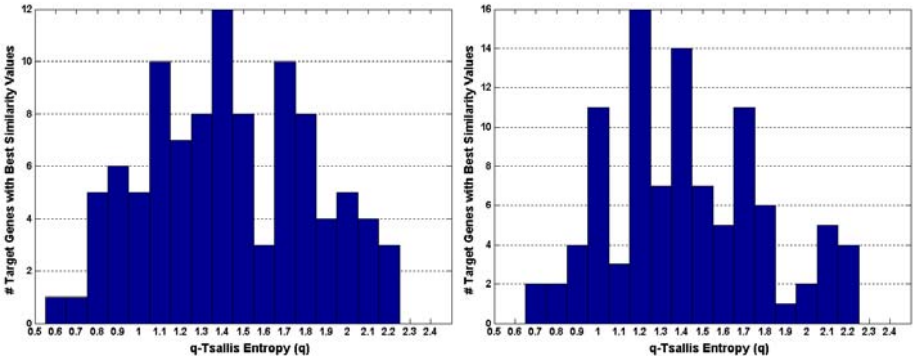


(a) ER (b) BA

Fig. 1. Network identification rate by increasing q of the Tsallis entropy, using: (a) uniformly-random Erdös-Rényi (ER) and (b) scale-free Barabási-Albert (BA)

Figures 1 (a) and (b) show the median values of similarity (described in Section 4) between the inferred networks and AGN-based networks in terms of q of the Tsallis entropy and the average node degrees (k). These figures present a soft increase of similarity rate by increasing the q for all average degrees k. This result suggests a dependence of the method accuracy on the parameter q.

In order to better investigate this behavior, Figures 2 (a) and (b) present the histograms of the frequency of target genes with best similarity rate found for each q value, by considering respectively, ER and BA network topologies.



(a) ER (b) BA

Fig. 2. Histogram of the frequency of target genes with best similarity value per q, over all average node degree k., using: (a) uniformly-random Erdös-Rényi (ER) and (b) scale-free Barabási-Albert (BA). All average node degree k was considered.

Table 2. The best results found for $q = 1$ and for all q values (global) by considering: (a) uniformly-random Erdös-Rényi network topology (ER) and (b) scale-free Barabási-Albert network topology (BA)

(a) ER

q	k	**TP**	**FP**	**TN**	**FN**
1	1	175	195	9630	0
global		179	121	9700	0
1	2	224	137	9639	0
global		228	62	9710	0
1	3	231	136	9633	0
global		241	71	9688	0
1	4	208	184	9608	0
global		218	106	9676	0
1	5	205	206	9578	11
global		210	139	9643	8

(b) BA

q	k	**TP**	**FP**	**TN**	**FN**
1	1	114	257	9629	0
global		114	193	9693	0
1	2	206	102	9692	0
global		206	27	9767	0
1	3	283	77	9636	4
global		290	20	9689	1
1	4	250	130	9508	112
global		289	78	9548	85
1	5	255	135	9423	187
global		283	105	9451	161

Figure 2 (a) presents higher frequency when $q = 1.4$, and the distribution is concentrated between $q = 0.6$ and $q = 2.2$. On the other hand, Figure 2 (b) presents higher frequency when $q = 1.2$, and the distribution is concentrated between $q = 0.7$ and $q = 2.2$. These results reinforce the fact that the variation of q is important for the method accuracy, *i.e.*, the nonextensivity of the networks.

In order to estimate the improvement of the accuracy by varying q, Tables 2 (a) and (b) present the comparisons of commonly used Shannon entropy $q = 1$ and the best global results obtained by the variation of q. In general, it is possible to notice that global results exhibit an improvement of accuracy *w.r.t* $q = 1$ for all average node degrees (k) in both network topologies (ER and BA). In particular, the number of false positives (FP) presents higher improvement of accuracy, achieving 55% of reduction in false positives for ER when $k = 2$ and 74% for BA when $2 \leqslant k \leqslant 3$.

6 Conclusion

This work described a comparative analysis in order to evaluate the application of Tsallis entropy in a GRN inference method based on a feature selection approach by considering three main aspects: (1) variation of the parameter q (degree of nonextensivity) of Tsallis entropy; (2) two different complex network topologies; (3) complexity of networks in terms of average node degree (k).

The results indicated a valuable improvement of the accuracy of the GRNs inference by using the Tsallis entropy. This improvement was observed in both kinds of networks (ER and BA) and also for different degrees of complexities k (average gene degree). We have found the best similarity values on the range $0.6 \leqslant q \leqslant 2.2$, where the degree of nonextensivity q around 1.4 performs better results. In fact, the results have shown that tested networks tend to be a little subextensive ($q > 1$).

These results can be seen as a first stage to better understand the inference of network topologies by information theory approaches, *i.e.*, by using entropy criteria. The main point is the possibility of nonextensivity of the networks and, therefore, the entropy related methods dependence on that.

The next stage of this work is to consider complex networks models and measurements [19] (local and global) and more precise similarity measures between two networks [29] in order to refine the analysis of the inference method and the nonextensivity of the networks. 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 method. Other methods that apply information theory for GRNs inference could be included in the comparative results and analysis of nonextensivity of the networks.

Acknowledgments

This work was supported by FAPESP, CNPq and CAPES.

References

1. Hovatta, I., et al.: DNA microarray data analysis, 2nd edn. CSC, Scientific Computing Ltd. (2005)
2. Liang, S., Fuhrman, S., Somogyi, R.: Reveal: a general reverse engineering algorithm for inference of genetic network architectures. In: PSB, pp. 18–29 (1998)
3. Weaver, D.C., et al.: Modeling regulatory networks with weight matrices. In: PSB, pp. 112–123 (1999)
4. Butte, A., Kohane, I.: Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: PSB, pp. 418–429 (2000)
5. Keles, S., van-der Laan, M., Eisen, M.B.: Identification of regulatory elements using a feature selection method. Bioinformatics 18(9), 1167–1175 (2002)
6. Shmulevich, I., et al.: Probabilistic boolean networks: a rule-based uncertainty model for gene regulatory networks. Bioinformatics 18(2), 261–274 (2002)
7. Barrera, J., et al.: 2. In: Constructing probabilistic genetic networks of Plasmodium falciparum from dynamical expression signals of the intraerythrocytic development cycle, pp. 11–26. Springer, Heidelberg (2007)

8. Margolin, A.A., et al.: Aracne: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC Bioinformatics 7(suppl. 1) (2006)
9. Faith, J., et al.: Large-scale mapping and validation of escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS Biology 5(1), 259–265 (2007)
10. Meyer, P.E., Kontos, K., Lafitte, F., Bontempi, G.: Information-theoretic inference of large transcriptional regulatory networks. EURASIP Journal on Bioinformatics and Systems Biology, 1–9 (2007)
11. Zhao, W., Serpedin, E., Dougherty, E.R.: Inferring connectivity of genetic regulatory networks using information-theoretic criteria. IEEE/ACM Trans. on Comput. Biology and Bioinformatics 5(2), 262–274 (2008)
12. Tsallis, C.: Possible generalization of Boltzmann-Gibbs statistics. Journal of Statistical Physics 52(1), 479–487 (1988)
13. Lopes, F.M., Cesar-Jr, R.M., Costa, L.F.: AGN simulation and validation model. In: Bazzan, A.L.C., Craven, M., Martins, N.F. (eds.) BSB 2008. LNCS (LNBI), vol. 5167, pp. 169–173. Springer, Heidelberg (2008)
14. Shannon, C.E.: A mathematical theory of communication. Bell System Technical Journal 27, 379–423, 623–656 (1948)
15. Issue, S.: Nonextensive statistical mechanics: new trends, new perspectives. Europhysics News 36(6), 185–231 (2005)
16. Abe, S.: Tsallis entropy: how unique? Cont. Mech. Therm. 16(3), 237–244 (2004)
17. Albert, R., Barabási, A.L.: Statistical mechanics of complex networks. Rev. Mod. Phys. 74(1), 47–97 (2002)
18. Newman, M.E.J.: The structure and function of complex networks. SIAM Review 45(2), 167–256 (2003)
19. Costa, L.F., Rodrigues, F.A., Travieso, G., Boas, P.R.V.: Characterization of complex networks: a survey of measurements. Adv. in Phys. 56(1), 167–242 (2007)
20. Kauffman, S.A.: Metabolic stability and epigenesis in randomly constructed nets. Journal of Theoretical Biology 22(3), 437–467 (1969)
21. Lopes, F.M., Martins-Jr, D.C., Cesar-Jr, R.M.: Comparative study of GRN's inference methods based on feature selection by mutual information. In: GENSIPS, pp. 1–4. IEEE Computer Society, Los Alamitos (2009)
22. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification, 2nd edn. Wiley, Chichester (2000)
23. Pudil, P., Novovičová, J., Kittler, J.: Floating search methods in feature selection. Pattern Recogn. Lett. 15(11), 1119–1125 (1994)
24. Lopes, F.M., Martins-Jr, D.C., Cesar-Jr, R.M.: Feature selection environment for genomic applications. BMC Bioinformatics 9(451), 1–21 (2008)
25. Tsallis, C.: Nonextensive Statistical Mechanics and its Applications. Lecture Notes in Physics. Springer, Heidelberg (2001)
26. Velazquez, L., Guzmán, F.: Remarks about the Tsallis formalism. Phys. Rev. E 65(4), 046134.1–046134.5 (2002)
27. Tsallis, C.: Nonadditive entropy: the concept and its use. The European Physical Journal A 40(3), 257–266 (2009)
28. Costa, L.F., et al.: Predicting the connectivity of primate cortical networks from topological and spatial node properties. BMC Systems Biology 1, 1–16 (2007)
29. Dougherty, E.R.: Validation of inference procedures for gene regulatory networks. Current Genomics 8(6), 351–359 (2007)