Gene expression complex networks: synthesis, identification and analysis

FABRÍCIO MARTINS LOPES[1,3], ROBERTO M. CESAR-JR.[1], LUCIANO DA F. COSTA[2]
[1]IME-USP, São Paulo, Brazil
[2]IFSC-USP, São Paulo, Brazil
[3]UTFPR, Paraná, Brazil

Abstract

An important question in computational biology is how genes are regulated and interact through gene networks. A method for the identification of gene networks from temporal data using pattern recognition techniques has been introduced in [1]. We have been developing algorithms to analyze the generated networks in terms of complex network concepts such as hubs and communities [2]. In order to validate the proposed algorithms, we have also developed a new approach to generate artificial gene networks (AGN). The AGN have two important components: the network topology and the predictors / target joint distribution. The present abstract reports the first results regarding the three following aspects for this framework: (1) AGN model generation; (2) gene network identification; (3) complex network analysis.

The AGN model generation starts by creating the topology of the complex networks. In this work we used scale-free networks as described in [2]. The generation of the AGN topology is carried out by first defining the AGN size followed by the edges generation process as in a scale-free network. Once the topology is generated, each gene (i.e. network vertex) is associated to input and output edges. The state of each gene is a function of its predictors, i.e. the genes associated to the input edges. Let X be a random vector representing the state of the predictors of a given gene. The state of such a gene, represented by a random variable Y, is defined as a function of the joint probability distribution $P(X,Y)$. Once the AGN topology is created, a distribution $P(X,Y)$ is randomly generated for each gene. The network is then simulated by random generation of initial states followed by the subsequent application of the created distributions for each gene. The time signals obtained for each gene for different initial states are obtained.

The next step is to submit the simulated data as input to the PGN identification method described in [1]. The PGN identification problem is modeled as a series of feature selection problems, one for each gene. The selected features are taken as predictor genes for each target gene. Hence, the selected predictors are used to link the genes and thus recover the network topology. In our approach, the Sequential Forward Search (SFS) algorithm has been adopted. The criterion function is the mean conditional entropy of $P(X,Y)$. The recovered network is then analyzed so that complex network measures can be extracted and statistically analyzed.

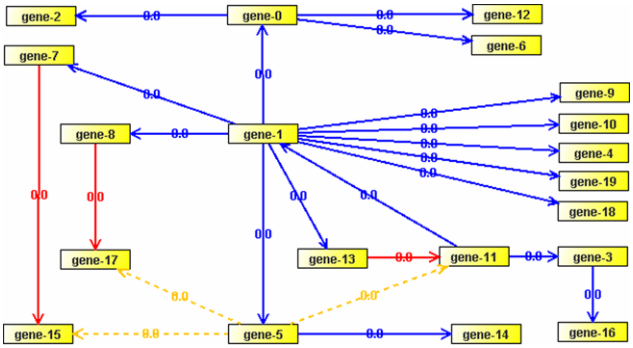

Figure 1 - An example of recovered network. The blue indicates predictors correctly found, the orange are false-positives and the red represents the predictors that did not found (negatives).

References

[1] Barrera, J.; Cesar-Jr, R. M.; Martins-Jr. D. C.; et al.; A new annotation tool for malaria based on inference of probabilistic genetic networks. *In Proceedings of the 5th International Conference for the Critical Assessment of Microarray Data Analysis (CAMDA '04)*, pp. 36–40, Durham, NC, USA, November 2004.

[2] Costa, L. da F.; Rodrigues, F. A.; Travieso, G.; Boas, P. R. V. Characterization of complex networks: a survey of measurements. *Advances in Physics*, v. 56, pp. 167-242, 2007.