Using graph theory to analyze gene network coherence

Francisco Gómez-Vela, Norberto Díaz-Díaz, José A Lagares, José A Sánchez, Jesús S Aguilar-Ruiz

School of Engineering, Pablo de Olavide University, Sevill

Motivation and Objectives

Gene networks (GNs) have become one of the most important approaches for modelling gene-gene relationships in Bioinformatics (Hecker et al, 2009). These networks allow us to carry out studies of different biological processes in a visual way.

Many GN inference algorithms have been developed as techniques for extracting biological knowledge (Ponzoni et al, 2007; Gallo et al, 2011). Once the network has been generated, it is very important to assure network reliability in order to illustrate the quality of the generated model. The quality of a GN can be measured by a direct comparison between the obtained GN and prior biological knowledge (Wei and Li, 2007; Zhou and Wong, 2011). However, these both approaches are not entirely accurate as they only take direct gene–gene interactions into account for the validation task, leaving aside the weak (indirect) relationships (Poyatos, 2011).

In this work the authors present a new methodology to assess the biological coherence of a GN. This coherence is obtained according to different biological gene-gene relationships sources. Our proposal is able to perform a complete functional analysis of the input GN. With this aim, graph theory is used to consider not only direct relationships but indirect ones as well.

Methods

The aim of our proposal is to evaluate the functional coherence of an input GN. The coherence is calculated according to current gene-gene interaction knowledge which is stored in public biological databases (DB). Thus, graph theory is applied with the aim of considering all gene-gene relationships (i.e. direct and indirect relationships) presented in the Input Network (IN).

Our approach works in various steps. First,the IN and the DB are converted into distance matrices (DM) using Floyd-Warshall algorithm (Asghar et al, 2012). This approach is a graph analysis method that solves the shortest path problem. This algorithm uses an adjacency matrix to compute the minimum path for every pair of genes. In this sense, the shortest path between two vertices is computed by incrementally improving an estimate on the shortest path between those vertices, until the estimate is optimal. Hence, the minimum distance of all gene pair combinations are computed and stored in DMin and DMdb, respectively. Furthermore, a distance threshold (δ) is used to exclude relationships that lack biological meaning. This threshold denotes the maximum distance to be considered as relevant in the DM generation process. Thus, if the minimum distance between two genes is greater than δ, then no path between the genes will be assumed.

Once the distance matrices have been obtained, they are combined to generate a new one. The new matrix, hereafter called Coherence Matrix (CM), contains the existing gap between the common genes in either the DMin and the DMdb.

$$CM= |DM_{IN} - DM_{DB}|$$

Where CM(i,j)= |DMin(i,j) – DMdb(i,j)| denotes the coherence of relationship between gene gi and gene gj with regard to the information stored in DB. Note that, relationships between genes within IN and DB will be only considered to generate CM. It is not possible to establish the quality of the rest of the relationships. DB contains no information to ascertain whether the relationships are biologically relevant or not.

According to the coherence values stored in CM and to an accuracy coherence level (θ), the differences and similarities between the GN and DB could be obtained. The differences are classified as false positives and false negatives, while the similarities as true positives and true negatives. Therefore, if CM(i,j) is greater than θ it will be considered as a false positive, while if it is less than or equal to θ, it will be computed as true positive. In case there is no path between gi and gj in the IN, neither in DB (IN(i,j)=DB(i,j)=infinite), it will be considered as a true negative. Nevertheless, if there is no path in IN but there is in DB, it will computed as a false negative.

Table 1: F-Measure and Accuracy values obtained by different input GN according to prior biological knowledge and chronologically sorted. The best results in each dataset are emphasized.

	Soinov		Bulashevska		Ponzoni (GRNCOP)		Gallo (GRNCOP2)	
	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy	F-Measure	Accuracy
BioGrid	0,42	0,27	0,79	0,65	0,9	0,82	0,86	0,75
KEGG	0,48	0,58	0,5	0,34	0,43	0,28	0,61	0,47
SGD	0,47	0,31	0,69	0,53	1	1	0,73	0,58
YeastNet	0,45	0,29	0,66	0,5	1	1	0,77	0,62

Results and Discussion

In order to assess the robustness of our proposal, we present a set of analysis of different yeast cell cycle networks using four prior biological knowledge data sets.

Input networks were produced applying four inference network techniques (Soinov et al, 2003; Bulashevska and Eils 2005; Ponzoni et al, 2007; Gallo et al, 2011) on the well-known yeast cell cycle expression data set (Spellman et al, 1998). Finally, the functional coherence of GNs generated is measured using our proposal according to the gene-gene interaction knowledge stored in BioGRID (Stark et al, 2010), KEGG (Kanehisa et al, 2012), SGD (Cherry et al, 2012) and YeastNet (Lee et al, 2007).

Multiple studies were carried out using different threshold value combinations. δ and θ have been modified from one to five, generating 25 diffe-rent combinations. The results show that the higher δ values, the greater is the noise introduced. Coherence level threshold (θ) shows similar behavior; the lower θ, the smaller is the noise. The most representative result, summarized in Table 1, was obtained for $\delta=4$ and $\theta=1$. This combination has a biological meaning. For each gene, only the interactions in a radius of four should be considered as relevant. Moreover, they ought to have a difference no greater than 1 to be considered as valid.

Table 1 shows that inference method proposed by Gallo (GRNCOP2) generates the most reliable result, although Ponzoni technique (GRNCOP) provides the best result in three of the four data sets. Soinov approach obtains the worst values.

These results are consistent with the experiment carried out in (Ponzoni et al, 2007) and (Gallo et al, 2011). GRNCOP was successfully compared with Soinov and Bulashevska approaches, while Gallo et al presented a detailed analysis of the performance of GRNCOP and GRNCOP2, where the last one shows the most stable result. These behaviors are also found in the obtained results. GRNCOP presents better coherence values than Soinov and Bulashevska in BioGrid, SGD and YeastNet. Similarly, GRNCOP2 obtains more stable values than GRNCOP, especially for F-measure.

References

1. Asghar A, et al (2012) Speeding up the Floyd–Warshall algorithm for the cycled shortest path problem. AppliedMathematics Letters 25(1): 1
2. Bulashevska S and Eils R (2005) Inferring genetic regulatory logic from expression data. Bioinformatics 21(11):2706.
3. Cherry JM, et al (2012) Saccharomyces Genome Database: the genomics resource of budding yeast.Nucleic Acids Research 40: D700-705. doi:10.1093/nar/gkr1029
4. Gallo C, et al (2011) Discovering time-lagged rules frommicroarray data using gene profile classifiers. BMCBioinformatics 12:123.
5. Hecker M, et al (2009) Gene regulatory network inference:Data integration in dynamic models – a review. Biosystems 96:86.
6. Kanehisa M, et al (2012) KEGG for integration and interpretationof large-scale molecular datasets. Nucleic AcidsResearch 40:D109-D114
7. Lee I, et al (2007) An improved, bias-reduced probabilistic-functional gene network of baker's yeast,Saccharomyces cerevisiae. PLoS ONE 2(10):e988.
8. Ponzoni I, et al (2007) Inferring adaptive regulationthresholds and association rules from gene expressiondata through combinatorial optimization learning.IEEE/ACM Transaction on Computation Biology andBioinformatics 4(4):624.
9. Poyatos JF (2011). The balance of weak and stronginteractions in genetic networks. PloS One 6(2):e14598.
10. Soinov L, et al (2003) Toward reconstruction of genenetworks from expression data by supervised learning. Genome Biology 4:R6.

11. Spellman PT, et al (1998). Comprehensive identificationof cell cycle-regulated genes of the yeastSaccharomyces cerevisiae by microarray hybridization.Molecular Biology of the Cell 9(12):3273.
12. Stark C, et al (2010) The BioGRID Interaction Database:2011 update. Nucleic Acids Research 39 (Database is sue):D698
13. Wei Z and Li H (2007). A Markov random field model fornetwork-based analysis of genomic data. Bioinformatics23(12):153 Zhou H and Wong L (2011). Comparativeanalysis and assessment of M.tuberculosis H37Rv protein-proteininteraction datasets. BMC genomics, 12 (Suppl 3):S20)
14. Zhou H and Wong L (2011). Comparative analysis and assessment of M. tuberculosis H37Rv protein-protein interaction datasets. *BMC genomics*, **12** (Suppl 3):S20.