Computational Biology and Chemistry 56 (2015) 142–151

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/compbiolchem

Research Article

Gene network coherence based on prior knowledge using direct and indirect relationships

Francisco Gómez-Vela*, José Antonio Lagares, Norberto Díaz-Díaz

School of Engineering, Pablo de Olavide University, Seville, Spain

ARTICLE INFO

Article history:
Received 22 July 2014
Received in revised form 6 March 2015
Accepted 20 March 2015
Available online 27 March 2015

Keywords:
Gene association networks
Biological knowledge
Gene network assessment
Biological validation
Heuristic algorithm

ABSTRACT

Gene networks (GNs) have become one of the most important approaches for modeling biological processes. They are very useful to understand the different complex biological processes that may occur in living organisms. Currently, one of the biggest challenge in any study related with GN is to assure the quality of these GNs. In this sense, recent works use artificial data sets or a direct comparison with prior biological knowledge. However, these approaches are not entirely accurate as they only take into account direct gene–gene interactions for validation, leaving aside the weak (indirect) relationships.

We propose a new measure, named gene network coherence (GNC), to rate the coherence of an input network according to different biological databases. In this sense, the measure considers not only the direct gene–gene relationships but also the indirect ones to perform a complete and fairer evaluation of the input network. Hence, our approach is able to use the whole information stored in the networks. A GNC JAVA-based implementation is available at: http://fgomezvela.github.io/GNC/

The results achieved in this work show that GNC outperforms the classical approaches for assessing GNs by means of three different experiments using different biological databases and input networks. According to the results, we can conclude that the proposed measure, which considers the inherent information stored in the direct and indirect gene–gene relationships, offers a new robust solution to the problem of GNs biological validation.

1. Background

In the last few years a huge amount of biological information has been analyzed by researchers in order to obtain useful knowledge. Currently, the analysis of this information and its representation are challenges that are faced by modeling techniques. Gene networks (GNs) are one of the most accepted tools for the representation of the genetic models in current bioinformatics studies, since they are able to show easily and visually the gene regulatory processes. GNs present the biological information as a graph, where genes are represented as nodes of the graph, and their relationships are presented as edges (Pavlopoulos et al., 2011). There are many works which use the GNs as a method to represent the regulatory gene processes. For example, the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000) is one of the most widely used repositories for analyzing relationships between genes. The metabolic pathways stored in KEGG contain knowledge about different biological processes in GNs structures.

The GN generation process is a crucial step in any study related to reconstruction of gene regulatory processes. Depending on the chosen method to generate the network, the model obtained may widely change. Hecker et al. (2009) mainly categorized the GNs into four approaches based on the algorithm used to obtain the models: Information theory-based networks, which usually use a correlation algorithm to establish gene–gene interactions depending on similar expression patterns among the genes (Butte and Kohane, 2000); logical networks (Boolean networks) (Bornholdt, 2008) involving only two states for the network genes: expressed and unexpressed; differential equation-based networks that describe gene expression changes as a function of the expression of other genes and environmental factors (Climescu-Haulica and Quirk, 2007); and finally, Bayesian networks (Needham et al., 2007), which reflect the stochastic nature of gene regulation and make use of the Bayes-rule.

Further, GNs can be represented in different levels of abstraction. These levels range from the detailed gene regulatory processes (like the metabolic pathways stored in KEGG where genes, proteins, compounds and metabolites interact on chemical reactions) to the

* Corresponding author. Tel.: +34 954967868; fax: +34 954348377.
E-mail addresses: fgomez@upo.es (F. Gómez-Vela), jalagrod@alu.upo.es (J.A. Lagares), ndiaz@upo.es (N. Díaz-Díaz).

http://dx.doi.org/10.1016/j.compbiolchem.2015.03.002
1476-9271/

gene association networks, where the relationships represent some kind of influence between the genes. In this sense, it is possible to transform a complex network into an association network rising its level of abstraction (Hecker et al., 2009). For example, in the work of Sales et al. (2012) a method to obtain gene association networks from the metabolic pathways of KEGG is presented.

Once the network is obtained, the next step is to ensure the reliability of the inferred model. In order to evaluate the GNs quality, synthetic data (Van den Bulcke et al., 2006) and/or prior biological knowledge (Li and Li, 2008) are used to assess the quality of the relationships presented in the network.

On the one hand, synthetic or artificial data represent the simulation of a real biological data set. This method is based on producing an artificial data set according to a previously known network. The simulated gene expression values are stored in a data set in order to be used as input for the GN inference method. Finally, the performance of the algorithm is tested by a comparison of the two networks, the real previous network and the generated one. Different tools, presented in the literature, may be used for producing these artificial data sets (Schaffter et al., 2011). For example, Gill et al. (2010) used synthetic data sets generated by the Syntren tool (Van den Bulcke et al., 2006) in order to provide a recipe for conducting a differential analysis of networks constructed from microarray data and proposed formal statistical test for gene networks.

Despite the fact that synthetic data are useful to rate the performance of algorithms, they cannot reproduce completely the complex internal features of real biological processes since they are only based on simulations of real processes.

On the other hand, the use of real biological knowledge provides great reliability in the correctness of the models validated, since this approach allows a direct comparison of them with highly accepted data. This validation, widely used in literatures (Aguilar-Ruiz et al., 2011; Kamburov et al., 2012; Nakamura et al., 2012), is usually carried out by using biological databases such as Gene Ontology (Ashburner et al., 2000), SGD (Cherry et al., 2012) or YeastNet Lee et al. (2007) as gold standard.

In spite of the fact that the methods presented above are useful for the validation process, they present a fundamental lack. These methods do not use all the information stored in a GN, as they only consider the strong relationships leaving aside the indirect ones (Wei and Li, 2007). The relationships stored in a GN can be classified into two groups (Poyatos, 2011): direct (or strong) which connect one pair of genes directly; indirect (or weak) where two genes are connected by a path with multiple edges. Due to this, a direct comparison is not completely accurate because it ignores weak relationships.

Furthermore, the use of the indirect relationships is useful to mitigate another important problem in the validation of GNs. Inference algorithms are not always able to generate a network that reproduce the internal features of biological processes to identify the relationships between genes (Muddana et al., 2006). This issue may occur due to problems in the input data set, and are oblivious to the inference algorithm (Zeisel et al., 2010; Draghici et al., 2006).

In this paper, the authors present a new method to assess the biological coherence of gene association networks by using the knowledge stored in biological databases. The method obtains a measure, named GNC (gene network coherence), which is able to consider not only the direct relationships, but also the indirect ones to perform a complete and fairer evaluation of the input network.

To show the performance of GNC, we used some representatives networks from the *Saccharomyces cerevisiae* and the *Homo sapiens* in three different experiments. The results achieved show that GNC outperforms the classical approaches for assessing GNs and offers a new robust solution to the problem of GNs biological validation.

2. GNC approach

Our approach is based on the use of direct and indirect relationships presented in a GN for the evaluation of its biological quality. GNC rates the biological coherence of an input network regarding a specific biological database (DB). Depending on the type of the information to be validated, i.e. protein–protein interaction (PPI), co-expression relationships, gene regulatory relationships or even all of them, a different database should be selected.

The methodology entails three consecutive steps (see Fig. 1), which are described below.

2.1. Step 1: obtaining adjacency matrices

In the first step, the information of the gene–gene relationships presented in the input network, and the information of the relationships from the selected biological database, are represented as adjacency matrices (*AM*). The adjacency matrix of a GN of N genes is the $N \times N$ upper-diagonal matrix where the entry a_{ij} is 1 if a direct edge exists from gene g_i to gene g_j, and 0 if there is no direct edge between them. An example is depicted in Fig. 1, where two *AM*s are obtained from the input network and the biological database, respectively.

After the *AM*s have been obtained, the Floyd–Warshall algorithm (Asghar Ainia, 2012) is used to calculate the minimum path for every pair of genes. Hence, the minimum distance of all gene pair combinations is computed and stored into two distance matrices; one for the input network (DM_{IN}) and another for the database (DM_{DB}). Note that indirect relationships are considered to generate both distance matrices. This process is depicted in the step 1 of Fig. 1.

2.2. Step 2: obtaining the coherence matrix

Once the distance matrices have been obtained, they are combined to generate a new one that stores the coherence of the gene–gene relationships from the input network regarding the biological database. The new matrix, hereafter called Coherence Matrix (*CM*), is a $V \times V$ upper-diagonal matrix where V is the number of genes considered. This matrix stores the existing gap among the common genes in both matrices (DM_{IN} and DM_{DB}). The method uses only the common genes between the input network and the database because there is no information to rate the quality of the interactions from genes in *IN* that are not present in *DB*. With this pruning, the problems associated with the different networks size are overcome. Each value of the coherence matrix is calculated by a distance function (1) that is described at following:

Definition 1. Biological coherence of relationship between $gene_i$ and $gene_j$: Given the relationship distances between $gene_i$ and $gene_j$ in the input network and the database, respectively ($DM_{IN}(i, j)$ and $DM_{DB}(i, j)$), each entry of the *CM* ($CM(i, j)$) is calculated as follows:

$$CM(i,j) = \frac{1}{(|\alpha - \beta| \min\{\alpha, \beta\} + 1)} \quad (1)$$

where α and β denote entries from the input network and database *DM*s, respectively. That is, $\alpha = DM_{IN}(i, j)$, $\beta = DM_{DB}(i, j)$.

Eq. (1) presents a distance function to obtain the coherence of the relationships from the input network. This equation uses the same concept presented in the topology-based distance function presented by Dougherty (2007), where the function is showed as the summation of the difference between the distances of the relationships (see Section 5 for more details). In contrast to this, our approach is able to score not only the existing gap between the networks, but also weight this gap based on its relevance (distance