Physica A 458 (2016) 313–328

Contents lists available at ScienceDirect

Physica A

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

Gene-based and semantic structure of the Gene Ontology as a complex network

Claudia Coronnello [a,b], Michele Tumminello [c], Salvatore Micciché [d,*]

[a] Fondazione Ri.MED, Via Bandiera 11, 90133, Palermo, Italy

[b] IBIM-CNR, Via Ugo la Malfa 153, 90146, Italy

[c] Università degli Studi di Palermo, Dipartimento di Scienze Economiche, Aziendali e Statistiche, Viale delle Scienze, Ed. 13, 90128, Palermo, Italy

[d] Università degli Studi di Palermo, Dipartimento di Fisica e Chimica, Viale delle Scienze, Ed. 18, 90128, Palermo, Italy

HIGHLIGHTS

- We study the projected network of terms starting from a bipartite terms/genes network.
- GO terms distinct from a semantic point of view might be linked in the above network.
- Such GO terms are in the same community when considering their gene content.
- This is important from a biomedical point of view, as it reveals relations amongst biological functions.

ARTICLE INFO

Article history:
Received 19 November 2015
Received in revised form 30 January 2016
Available online 13 April 2016

Keywords:
Complex systems
Networks
Bipartite system
Community detection
Ontology
Genes

ABSTRACT

The last decade has seen the advent and consolidation of ontology based tools for the identification and biological interpretation of classes of genes, such as the Gene Ontology. The Gene Ontology (GO) is constantly evolving over time. The information accumulated time-by-time and included in the GO is encoded in the definition of terms and in the setting up of semantic relations amongst terms. Here we investigate the Gene Ontology from a complex network perspective. We consider the semantic network of terms naturally associated with the semantic relationships provided by the Gene Ontology consortium. Moreover, the GO is a natural example of bipartite network of terms and genes. Here we are interested in studying the properties of the projected network of terms, i.e. a gene-based weighted network of GO terms, in which a link between any two terms is set if at least one gene is annotated in both terms. One aim of the present paper is to compare the structural properties of the semantic and the gene-based network. The relative importance of terms is very similar in the two networks, but the community structure changes. We show that in some cases GO terms that appear to be distinct from a semantic point of view are instead connected, and appear in the same community when considering their gene content. The identification of such gene-based communities of terms might therefore be the basis of a simple protocol aiming at improving the semantic structure of GO. Information about terms that share large gene content might also be important from a biomedical point of view, as it might reveal how genes over-expressed in a certain term also affect other biological processes, molecular functions and cellular components not directly linked according to GO semantics.

* Corresponding author.
E-mail address: salvatore.micciche@unipa.it (S. Micciché).

http://dx.doi.org/10.1016/j.physa.2016.03.062
0378-4371/

1. Introduction

The last decade has seen the advent and consolidation of ontology based tools for the identification and biological interpretation of classes of genes, such as the Gene Ontology (GO) [1]. GO allows for associating a gene to its biological functions and it also provides the information about the other genes which cooperate in performing such functions. As such, GO is a useful tool for exploiting the existence of sets of genes involved in a certain pathology. The GO is constantly evolving [2–4] over time. The information accumulated time-by-time and included in the GO is encoded in the definition of terms and in the setting up of semantic relations amongst terms. The semantic GO structure is mainly based on the knowledge of existing relations amongst biological functions, based on the available literature.

The GO is a natural example of bipartite complex system of terms and genes. One can therefore investigate the properties of the associated bipartite network, as well as the properties of its projected networks. Here we will be interested on the projected network of terms, i.e. a gene-based weighted network in which the nodes are the terms and a link between any two terms is set up whenever a gene is annotated in both terms [5]. Recently a methodology has been proposed that identifies preferential links in the projected network [6], i.e. links whose presence in the projected network cannot be explained in terms of a random co-occurrence of neighbors in the bipartite system. The resulting network is called statistically validated network (SVN). One aim of the present paper is to understand whether the semantic and the gene-based term networks share the same structural properties or not, even at the level of statistically validated networks. In fact, such approach might be the basis of protocols that are able to capture the relations amongst genes so that they can be profitably transferred at the level of terms.

Another way to compare the information encoded in the semantic GO structure with the one associated to the genes annotated in the terms is to investigate the network communities. Communities within the gene-based network are gathering GO terms that share a similar profile in terms of their annotated genes. Communities in the semantic network put together GO terms that share a similar profile in terms of their semantic relationships. Indeed, the idea of searching for communities of GO terms is not new. However, one usually looks for communities of the semantic GO network only [7–14]. Our approach is different. In fact, we use the information on the gene content of any GO term in order to create a statistically validated network of GO terms and then we partition it by using any standard community detection algorithm. The statistical characterization of these communities is then performed by using the information relative to the communities in the semantic network. As we will show, it turns out that in some cases GO terms present in the same community of the gene-based SVN are not joined by any semantic link. This shows that terms that appear to be distinct from a semantic point of view are instead connected when considering their gene content. The identification of communities in the gene-based SVN can therefore be the basis of a simple protocol able to fully exploit the possible relationships amongst terms, thus improving the knowledge of the semantic structure of GO.

The above results indicate that the gene-based SVN has a modular structure organized around the three main GO branches (BP, MF, CC). Such results refer to a small portion of the entire gene-based network, since the SVN only accounts for 4% of the whole set of GO terms. In order to verify if the SVN and the whole gene-based network share similar properties, we used an approach different from the community detection analysis, which is not feasible to investigate the behavior of the complete gene-based network. We then studied the spectral properties of the correlation matrix associated to the gene-based network. The analysis of the spectral properties of the whole correlation matrix confirms that the semantic distinction of the three branches (BP, CC and MF) is also a fingerprint for the gene-based network. Furthermore, the community structure of the gene-based SVN is confirmed by investigating the hierarchical structure of its terms. In fact, we find that the communities of the gene-based validated network are compatible with the clusters obtained by applying the average linkage clustering technique to the correlation matrix associated to its terms.

One final investigation regards the role of the gene annotations in the GO terms in the level of modularity of the gene-based network. Specifically, we investigated whether or not each single genes is preferentially annotated in one of the three branches. Our investigation shows that a crucial role is played by the semantic relations amongst terms. In fact, when we consider all the annotations, as inherited from the semantic GO links, we find 40% of genes preferentially annotated in CC. On the contrary, when we consider genes as annotated only in the most specific terms, genes annotated in different branches are compatible with the random null hypothesis of genes annotated uniformly among the GO branches.

As a by-product, we present a simple methodology that allows to have a first glance insight about the biological meaning of groups of GO terms. We have put on a statistical basis what any researcher first does when he obtains a list of GO terms that are somehow relevant in the analysis he is performing. The first thing to do is to read the definitions of the GO terms trying to figure out a possible "story" for the reason why the obtained terms are connected together. We have devised a procedure that helps in this direction by providing the most relevant "words" of the "story".

The paper is organized as follows: in Section 2.1 we illustrate the data considered in our investigation, in Section 2.2 we will briefly review the methodologies that allows the generation of statistically validated networks while the community characterization methodology is illustrated in Section 2.3. In Section 3 we will study the semantic and gene-based network and show GO terms that belong to the same gene-based network community and are not joined by any semantic link. Our conclusions are drawn in Section 4.