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 Miccichè [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.

© 2016 Elsevier B.V. All rights reserved.

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

http://dx.doi.org/10.1016/j.physa.2016.03.062
0378-4371/© 2016 Elsevier B.V. All rights reserved.

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.

2. Data and methodology

2.1. The Gene Ontology database

We consider only the human part of the Gene Ontology (GO). To this end we downloaded the *gene_association.goa_human.gz* file, release 1.224 with GOC validation date 20/02/2012. This is the file that accounts for the association between terms and genes. Together with that we have also downloaded the *gene_ontology_edit.obo* file, release 1.1.2667 with release date: 02/03/2012 09:20. That file contains a description of the terms and the semantic links amongst them. Accordingly, the system under investigation consists of 12 564 terms and 18 992 genes. The backbone of the GO is composed by 13 788 *is_a* links among terms. Other commonly used relationships in GO are *part_of* (1962 links) and *regulates* (1881 links). Based on the meaning of the *is_a* and *part_of* relationships, we associate to a term all genes directly annotated in its *is_a* or *part_of* children terms. To the purpose of this work, the *regulates* relationships have been disregarded.

2.2. Statistically validated networks

Many complex systems present an intrinsic bipartite nature and are often described and modeled in terms of networks. Bipartite networks are composed by two disjoint sets of nodes, say set **A** and set **B**, such that every link connects a node in the first set with a node of the second set. The bipartite network is often transformed by one-mode projecting, i.e. one creates a network of nodes belonging to one of the two sets and two nodes are connected when they have at least one common neighboring node of the other set. Bipartite networks are often very heterogeneous in the number of relationships that the elements of one set establish with the elements of the other set. A new methodology to statistically validate each link of the projected network against a null hypothesis taking into account the heterogeneity of the system has been recently introduced [6].

Let us consider two nodes A_1 and A_2 both belonging to **A**. Let N_1 be the number of elements in set **B** linked to node A_1 and N_2 the number of elements in set **B** linked to node A_2. The total number elements in set **B** is N_B and the observed number of elements in set **B** both linked to A_1 and A_2 is N_{12}. Under the null hypothesis of random co-occurrence, the probability of observing X co-occurrences of links both in A_1 and A_2 is given by the hypergeometric distribution [15]:

$$H(X|N_B, N_1, N_2) = \frac{\binom{N_2}{X}\binom{N_B-N_2}{N_1-X}}{\binom{N_B}{N_1}}. \tag{1}$$

We can therefore associate a p-value to the observed N_{12} as $p(N_{12}) = 1 - \sum_{X=0}^{N_{12}-1} H(X|N_B, N_1, N_2)$. It is therefore possible to associate a p-value to each link in the projected network. After fixing a threshold one is thus able to select those links whose p-value is below the threshold. These links constitute the statistically validated network. The selection of the appropriate threshold is a key point. In fact, since the null hypothesis is tested for all links of the original projected network, we are in the typical situation when multiple test correction procedures must be applied. There are two possible correction procedure: the Bonferroni correction [16] and the less restrictive FDR correction [17]. We refer to the network obtained by using the FDR correction for multiple comparisons as the FDR network. We refer to the network obtained by using the Bonferroni correction for multiple comparisons as the Bonferroni network. A software to compute the Bonferroni and FDR network is available at the following web-sites: http://ocs.unipa.it/validate.html.

When the heterogeneity in one of the two sets is large, the above approach must be modified. Suppose one wants to validate the links in the set **A** projected network and the heterogeneity in the elements of set **B** is high. One can construct bipartite subsystems S_k of the original bipartite system S consisting of all the N_B^k elements of set **B** with a given degree k and of all the elements from set **A** linked to them. By construction, a subsystem S_k is homogeneous with respect to set **B**. The methodology sketched above can therefore be applied to each subsystem S_k, thus obtaining a collection of statistically validated networks. We then aggregate all validations by generating a statistically validated network where a link between node pairs is established whenever it has been validated in at least one subsystem. Such link can be given a weight equal to the total number of subsystems in which the link itself has been statistically validated. A software to compute the Bonferroni and FDR network is available at the following web-sites: http://ocs.unipa.it/validate-k.html.

2.3. Community detection and characterization

Given a network, a first step in the understanding of the represented system is the identification of communities within the network. Communities are sets of nodes that are linked amongst them to a degree which is higher than the one expected on the basis of a null hypothesis of randomness [18]. Communities will be detected by using the Infomap algorithm [19]. We perform 100 runs of the algorithm, which performs 10 independent searches each run. We select the partition with the lowest *code length*. Another set of popular methods of community detection are based on modularity optimization, such as the one introduced in Ref. [20]. However, in all the networks considered in the present paper, communities obtained by using the modularity based methodology were larger than those obtained with Infomap. Therefore we decided to consider the

Infomap partition of the system, along the line of thinking that smaller communities are more homogeneous, and, therefore, their characterization in terms of attributes would be more effective, as illustrated below.

Indeed, when communities are detected, it is important to characterize them, i.e. to understand what are the main features that explain why nodes are grouped together in a community. A statistically robust methodology for the community characterization has been given in Ref. [21]. The main idea is to use attributes specific of the nodes involved in the community in order to see which attribute is most represented in the community. Suppose to have a community of K nodes. Suppose X out of K nodes are characterized by having a certain attribute A. Suppose that in a network of N elements the attribute A can be associated to M out of N nodes. Then the probability that X is observed by chance is given by the hypergeometric distribution:

$$H(X|N, M, K) = \frac{\binom{M}{X}\binom{N-M}{K-X}}{\binom{N}{K}}. \tag{2}$$

For each attribute present in a community and for each community in the network we can therefore have a p-value. By considering the multiple hypothesis testing corrections illustrated above we can thus investigate what are the attributes that result to be over-expressed in a community. A software to perform the statistical characterization of communities within a network is available at the following web-site: http://ocs.unipa.it/characterize.html.

One of the sets of attributes we used to perform the community characterization is the 3-words glossary described next. For each term in the whole semantic network we first consider the words that define each term. After eliminating articles and prepositions we have 12 855 distinct words. We then construct the 3-words obtained by concatenating together three consecutive words in the terms definition. For example, in the case of the GO term *GO:0048518 (positive regulation of biological process)* we get the two 3-words: *positive_regulation_biological* and *regulation_biological_process*. We have 34 309 distinct 3-words in the whole network. We also delete the 3-words that appear no more than two times within the whole network. We therefore have 10 254 distinct 3-words. We use these 3-words as attributes to the GO terms and statistically validate the over-expression of them in GO term communities. We have therefore devised a simple methodology that allows to have a first glance insight about 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 together. The procedure described above helps in this direction by providing the most relevant "words" of the "story".

3. Results and discussion

3.1. The semantic network

The semantic structure of the Gene Ontology can be described in terms of an adjacency network where the nodes are the GO terms and the links between terms are provided by the semantic links of the *gene_ontology_edit.obo* file. When restricting to the human case, we get a network with $N = 12\,564$ nodes and $L_s = 116\,422$ links.

The semantic network is naturally partitioned in three large communities of size $N_1 = 8118$, $N_2 = 3336$ and $N_3 = 1110$ that correspond to the three main *branches* of GO: *GO:0008150 (Biological Processes)*, *GO:0005575 (Cellular Component)* and *GO:0003674 (Molecular Function)*, named, respectively, BP, CC and MF. The number of links connecting terms inside the three branches are 85 447, 18 489 and 12 486, respectively. The network has been further partitioned by using the community detection algorithm Infomap [19]. We thus obtained a partition of the adjacency semantic network in 163 communities of size ranging from 1096 to 2, as shown in Fig. 1 (right panel). When using Infomap, we performed 100 partitions generated starting from different seeds and considered the one with the smallest value of *code length*. However, the 100 obtained partitions were nevertheless quite similar to each other. In fact, the average value of the Jaccard coefficient $\langle J_i \rangle$ between the partition with lowest code-length and the ith partition was 0.79 with a standard deviation of 0.08. The Jaccard coefficient is defined as in Ref. [22]. Similarly, the mutual information [23] between the partition with lowest code-length and the ith partition was 0.955 with a standard deviation of 0.013.

The partitioned network is shown in Fig. 1 (left panel), which clearly shows that communities of terms tend to be homogeneous with respect to the three GO branches, namely BP, CC, and MF. The full list with the association between the GO terms and their community is given in Table SM1 of the Supplementary material. For example the 50th largest community, of size 54, contains terms like *GO:0001504 (neurotransmitter uptake)* or *GO:0007268 (synaptic transmission)* which are homogeneous from a biological point of view. We used the 3-word glossary to characterize the semantic network communities, as described in Section 2.3. The application of this methodology shows that indeed the clusters of terms detected in the semantic networks can be characterized by triplets with homogeneous meaning thus indicating that these clusters are meaningful from a biological point of view. The full list of 3-words characterizing the above communities is reported in Table SM2 of the Supplementary material. For the 50th largest community mentioned above this approach gets the following over-represented 3-words: *regulation_synapse_assembly* and *positive_regulation_synapse*, thus confirming that the 50th largest community groups together GO terms dealing with the regulation of the biological processes active at the level of synaptic transmission. As further examples, in Table 1 we show the results for the first two largest communities.

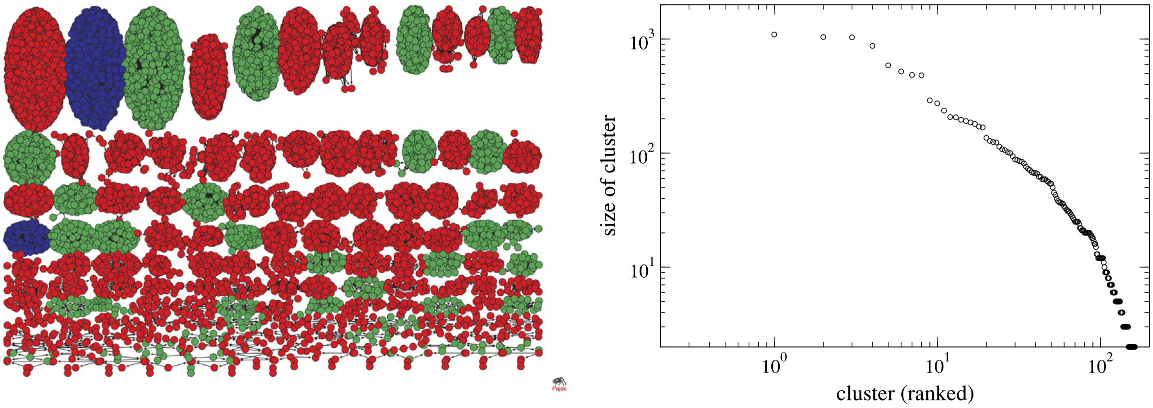

Fig. 1. Partition of the semantic adjacency network by using the Infomap algorithm [19]. The left panel shows the different communities observed. Nodes' color is based on the branch the node belongs (red: BP, blue: CC, green: MF). The right panel shows a rank plot of the size of communities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 1
Statistical characterization of the communities of the semantic adjacency network. Communities have been obtained by using the Infomap algorithm [19]. The characterization has been done by using the methodology of Ref. [21]. The attributes of each term are the 3-words obtained by concatenating together three consecutive words in the terms definition. Only the results for the first two communities are shown. The last column reports the GO ancestors common to at least 80% of the terms belonging to the community. The full list of characterizations is given in Table SM2 of the Supplementary material.

Community	Size	Overexpressed 3-words attribute	Common GO ancestors
1	1096	*(SSU-rRNA_5.8S_rRNA* *5.8S_rRNA_LSU-rRNA)* *sulfate_proteoglycan_biosynthetic* *proteoglycan_biosynthetic_process* *acid_catabolic_process* *mRNA_catabolic_process* *acid_metabolic_process* *compound_metabolic_process* *acid_biosynthetic_process*	Biological process Metabolic process Cellular process Cellular metabolic process
2	1043	*transport_vesicle_membrane* *endoplasmic_reticulum_membrane* *side_plasma_membrane* *ubiquitin_ligase_complex*	Cellular component Cell Cell part

The first community seems to aggregate terms involved in the metabolic processes that break down acids into smaller units and the second community groups terms involved in the transport and protein targeting processes.

Additional information about the communities composition can be obtained by detecting the common ancestors of the terms within each community. Operatively, common ancestors are defined as follows: for each term in a community we consider its ancestors and then we select those ancestors that are shared by at least 80% of the terms in a community. In many cases, the ancestor common to all the members of a community is only one of the top terms of the main three branches, i.e. BP, CC and MF. Then, in order to obtain more information about the community, we retrieved the ancestors common to at least the 80% of the terms in the community. With this approach, we noticed that the communities are composed by terms belonging to different sub-branches of the semantic network, each of them addressing specific biological meaning. In Table SM3 we report the significant 3-words and the common ancestors of the terms of each community. It is worth noticing that the two analytic approaches provide a different kind of information, the first giving an insight about the more specific GO terms belonging to the communities, the second providing their most general biological meaning. For the 50th largest community we observe that the common ancestors are: *synaptic transmission*, *cell–cell signaling*, *transmission of nerve impulse*, *multicellular organismal signaling*, *system process*, *cell communication*, *cellular process*, *signaling*, *multicellular organismal process*, *neurological system process*, *biological process* which are compatible with the 3-words characterization reported above. In Table 1 we show in the last column the results for the first two largest communities.

3.2. The gene-based projected networks

Let us consider the GO structure that emerges when considering the gene content of the terms. For each term of the Gene Ontology we consider all genes annotated in it. We assign a gene to a certain term whenever the gene is either directly annotated in it or in any of its children. We thus have a bipartite system of GO terms and genes. In Fig. 2 (left panel) we show

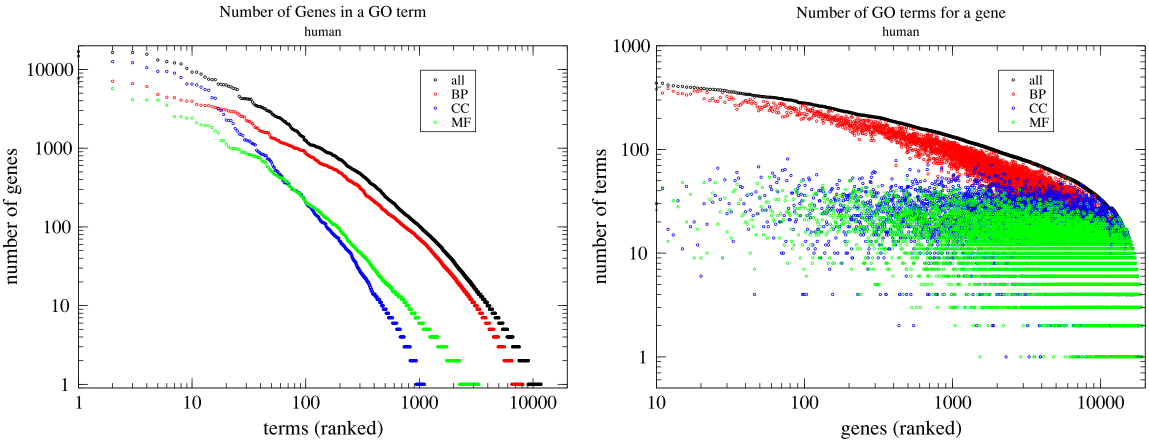

Fig. 2. For the GO terms/genes bipartite system, in the left panel we show the number of genes assigned to any term and in the right panel we show the number of terms a gene belongs to. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the number of genes assigned to any term and in Fig. 2 (right panel) we show the number of terms a gene belongs to. The heterogeneity is quite large in both cases. There are terms containing over 10 000 genes as well as terms containing only one gene. Analogously, there are genes present in over 500 GO terms as well as there are genes present in just two terms.

As mentioned in Section 2.2, starting from the bipartite system of GO terms and genes it is possible to construct two projected networks: the one of GO terms and the one of genes. The projected network of genes would be a network where nodes are genes and any two genes are connected if they belong to the same GO term. This would provide information about the links between genes based on their membership to biological processes, molecular functions and cellular components. For our purposes we will here consider the projected adjacency network of GO terms, i.e. we can generate the gene-based network where any two terms are connected by a link whenever there exists at least one gene assigned to both terms. Such adjacency network involves $N = 12\,564$ nodes and $L_g = 5\,142\,743$ links. It is worth noticing that the number of links in this network is much larger than the number L_s of semantic links.

In Fig. 3 (left panel) we show the degree distribution for the semantic (circles) and gene-based (triangles) adjacency networks. The profile of the two distributions is quite different, mainly due to the fact that terms in the gene-based network are much more linked to each other. This might be an indication of the fact that genes perform different tasks within different biological processes, molecular functions and cellular components.[1] In Fig. 3 (right panel) we show the scatter-plot describing the relationships between degree and betweenness for each term in the semantic (circles) and gene-based (triangles) networks. Contrary to the degree distribution, this panel shows that the two scatter-plots are quite similar. The solid line represents a reference curve $y \propto x^2$ showing the both scatter-plots are compatible with a quadratic dependence of the betweenness from the degree. These results suggest that the main difference between the two networks is in the number of links and not in the relative importance of the terms within the network. In this respect, it is worth mentioning that the average path length of the semantic adjacency network is 1.997 while the average path length of the adjacency gene-based network is 1.935 that is very close.

The number of links in the gene-based network is larger than in the semantic network, and this is maintained when one restricts the analysis to the three main branches, see Table 2. This probably explains why the average path length (APL) is on average slightly smaller in the gene-based network. It is perhaps surprising the fact that the difference in the average path lengths is so small, despite the fact that the difference in the number of links is so large between the two networks. This might be a confirmation of the existence of redundancy.

The gene-based network is fully connected. It cannot be partitioned in communities either by using the Infomap algorithm [19] or the Radatool [20] community detection algorithm. Therefore even terms belonging to different branches of the GO, i.e. BP, CC and MF, can be linked together when looking at their gene-content. Table 2 reports the number of links within and between the three branches. The total number of links connecting terms of different branches is 2 515 902, i.e. ∼50% of the total number of links. The majority of links therefore connects terms of the BP branch although a relevant number of links also connects the BP branch with the others.

The high number of links in the gene-based network might be a signature of redundancy: the relationships between biological processes mediated by genes might go through many different channels in order to ensure that a link between the terms is always active, despite possible impairments of some of the channels.

[1] Moreover, it should also be noted that this might be due to the fact that when generating the gene-based network a gene annotated in term T is also assigned to all terms that are parents of T.

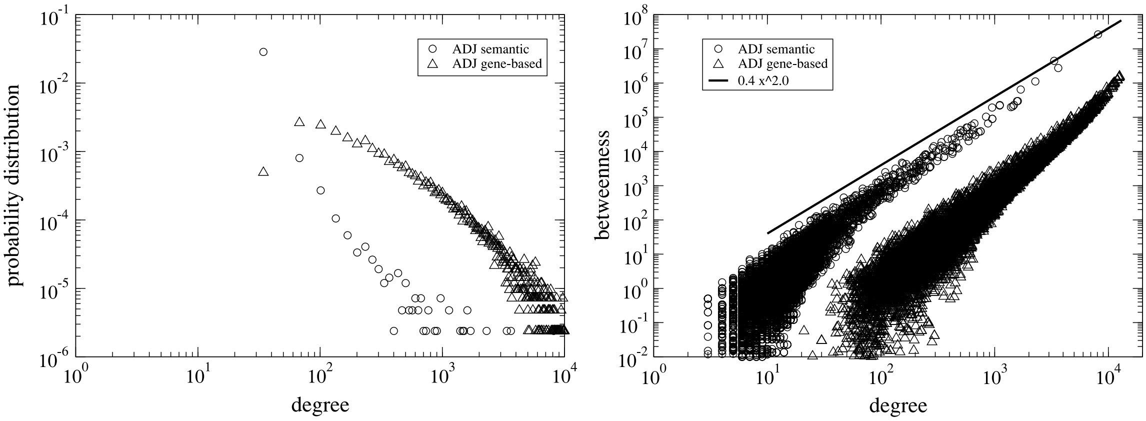

Fig. 3. The left panel shows a comparison between the degree distribution for the semantic (circles) and gene-based (triangles) networks. The right panel shows the scatter-plot describing the relationships between degree and betweenness for each term in the adjacency semantic (circles) and gene-based (triangles) networks. The solid line represents a reference curve $y \propto x^2$ showing the both scatter-plots are compatible with a quadratic dependence of the betweenness from the degree. Degree, betweenness and average path length have been computed by using the *igraph* library available for R (The R Project for Statistical Computing) http://cran.r-project.org/web/packages/igraph/index.html.

Table 2
Basic metrics for the semantic, gene-based and Bonferroni networks. In the table we also show the results for the three main in GO branches, namely *GO:0008150 (BP)*, *GO:0005575 (CC)* and *GO:0003674 (MF)*. The rows characterized by the notation X ↔ Y explicit the total number of links between the branches X and Y. The three branches are disconnected in the semantic network, while they are linked in the gene-based and Bonferroni networks. APL stands for Average Path Length. The second, third and fourth columns are relative to the semantic network. The last three columns are relative to the gene-based network.

Network	Nodes	Semantic network		Gene-based network		Bonferroni network		
Branch		Links	APL	Links	APL	Nodes	Links	s. links
ALL	12 564	116 422	1.997	5 142 743	1.935	533	3804	1338
BP	8 118	85 447	1.997	2 295 058	1.907	296	1515	752
MF	3 336	18 489	1.997	281 423	1.981	137	462	329
CC	1 110	12 486	1.980	50 360	1.899	100	355	257
BP ↔ CC	–	0	–	675 329	–	–	568	0
BP ↔ MF	–	0	–	1 600 604	–	–	658	0
CC ↔ MF	–	0	–	239 969	–	–	246	0

3.3. The Bonferroni gene-based network

As we mentioned above, the adjacency gene-based network is fully connected and shows a high level of redundancy. One might therefore ask whether there are preferential links amongst terms. The answer can be given by considering statistically validated networks. Such networks only involve links that are statistically validated against a null hypothesis of randomness that takes into account the natural heterogeneity of the system. Therefore they can be considered as a tool for filtering relevant information out of the system, based on the terms gene content.

Let us consider here the Bonferroni network of GO terms. Given the large heterogeneity in the genes set, we have constructed the validated network by adopting the validation procedure that involves the construction of subsystems where all elements have the same degree [24]. We had a total number of 346 possible subsets S_k, $k = 1, \ldots, 346$. The subsets with smallest degree was subsets S_1 composed of $N_B^1 = 285$ genes with degree 2. Such subset involved 14 GO terms. The subsets with highest degree was subsets S_{346} composed of $N_B^{346} = 1$ genes with degree 592. Such subset clearly involved 592 GO terms.

The Bonferroni network of GO terms thus obtained involves 533 nodes and 3804 links. It then involves the 4.75% of terms and the 0.08% of links with respect to the adjacency network. These numbers testify how large the reduction of both nodes and links can be when the statistical significance of a link is assessed by using a null hypothesis that properly takes into account the heterogeneity of the system. A less restrictive filtering is done when considering the FDR network. In this case one would have 915 nodes and 7670 links. A suggestive explanation of such large reduction is redundancy: the fact that the adjacency gene-based network contains so many links compatible with a null hypothesis of randomness might be due to the need of creating as many channels of communication amongst terms as possible so that impairments have negligible impact on the system. In this respect, the Bonferroni network provides the core of the system. Fig. 4 shows that the selected terms are generally the more populated and the closer to the root, but some term with low number of genes and high distance from the root is selected too. All the neglected terms are more specific than the terms selected in the Bonferroni network, and the vast majority of them are within four semantic links from the closest Bonferroni network term, as shown in Fig. 5.

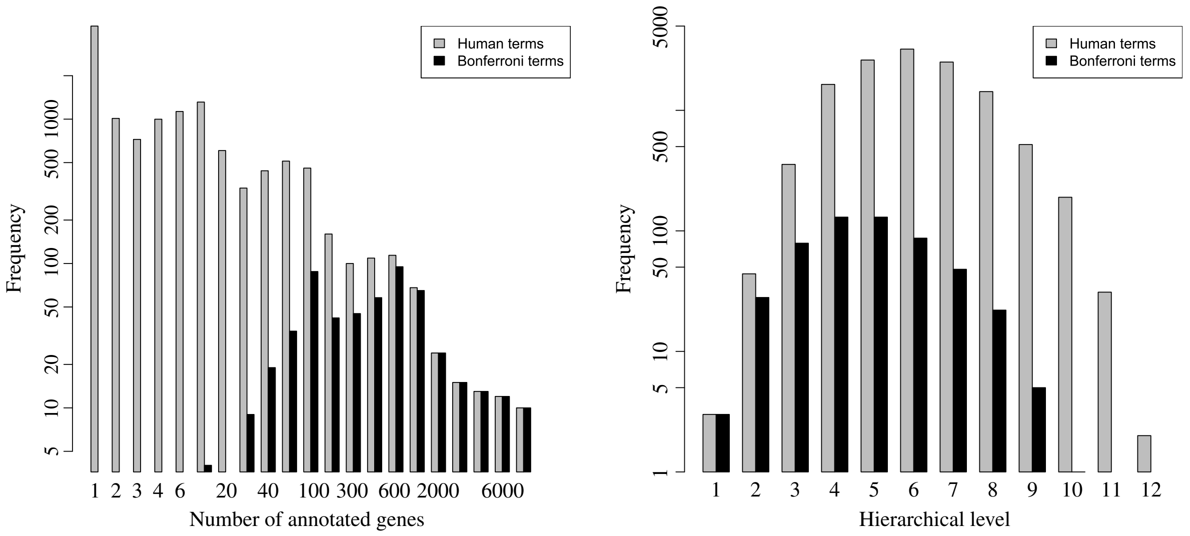

Fig. 4. Characterization of the Bonferroni network terms. Left: We plot the histogram of the number of genes annotated in GO terms (gray: All GO terms, black: Bonferroni network terms). Right: We plot the histogram of the hierarchical level of all GO terms (gray) and Bonferroni network terms (black). The hierarchical level of a term is computed as minimum number of semantic links from the term to the root.

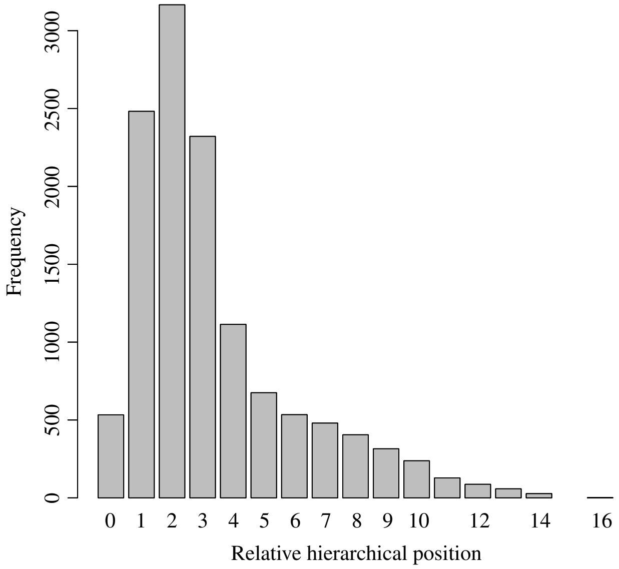

Fig. 5. Frequency plot of the minimum number of links (relative hierarchical position) between all the GO terms and the Bonferroni network terms.

This observations provide a clear image of the pruning performed by the Bonferroni network, which keeps terms closer to the root, although their distance from the root is not fixed *a priori*.

In Fig. 6 we show the degree of the nodes present in the Bonferroni network. The most connected nodes have degree values over 50. This means that despite the large reduction of nodes with respect to the adjacency network, there are still nodes that behave like hubs. Table 2 shows the number of links among the three GO branches found in the Bonferroni network. Among them, we counted the number of semantic links. For instance, we counted 1515 links within the BP branch, and only 752 of them are semantic links. The remaining 763 links are between unrelated sub-branches on the BP Ontology. These links, together with the links connecting the different GO branches, might reveal interesting relationships among biological processes, cellular components and molecular functions, that do not appear in the original semantic network.

3.4. *Community characterization of the Bonferroni gene-based network*

The Bonferroni network of GO terms is naturally partitioned in 30 communities. The largest community has size 467. The second largest has size 8, thus indicating the presence of a giant component. The network can be further partitioned by

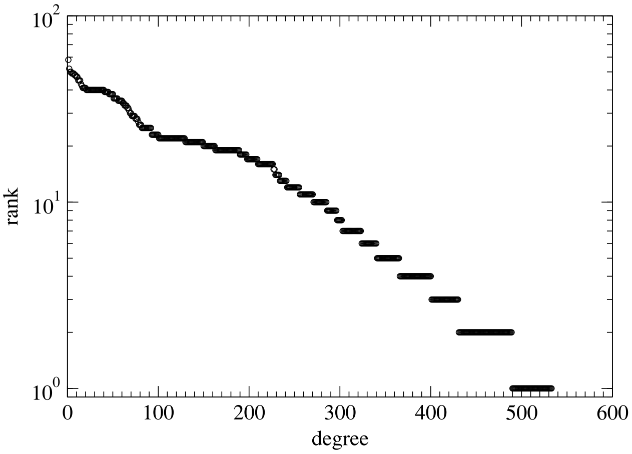

Fig. 6. The figure shows the degree rank plot for the Bonferroni gene-based network.

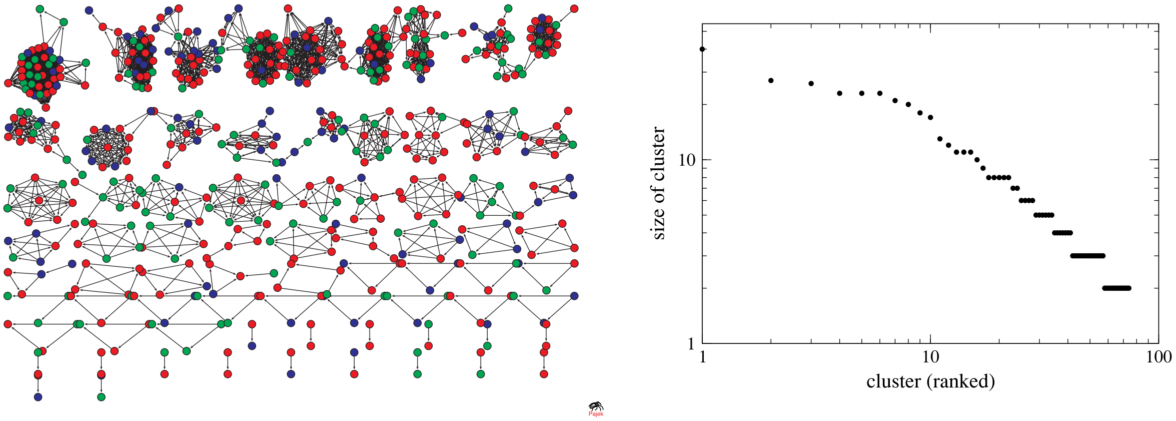

Fig. 7. Partition of the gene-based Bonferroni network by using the Infomap algorithm [19]. The left panel shows the different communities observed. Nodes' color is based on the branch the node belongs (red: BP, blue: CC, green: MF). The right panel shows a rank plot of the size of communities. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

using the community detection algorithm Infomap [19]. As a result, we get a partition involving 74 communities whose size ranges from 2 to 40, as shown in Fig. 7. Also in this case, when using Infomap, we performed 100 partitions generated starting from different seeds and considered the one with the smallest value of *code length*. However, all 100 obtained partitions had the same value of *code length*, thus indicating that in all cases we obtained the optimal partition of the system.

We can compare the information encoded in the semantic GO structure with the one associated to the genes annotated in the terms at the level of network communities. In fact, each of the above communities groups together GO terms on the bases of their gene content. However, any GO term T in a gene-based community C carries its semantic information inherited from the general GO semantic structure. We want to bring together these two levels of information by using the semantic information of any GO term T to characterize the gene-based community C. The idea is to use the semantic information of T obtained in the Infomap partitioning of Section 3.1 as attribute for the characterization of the gene-based community C.

In Section 3.1 we obtained a partitioning of the adjacency semantic network in communities that are homogeneous from a biological point of view. Thus, we here considered as attributes the membership of T to any of the communities A_i, $i = 1, \dots, 163$ of Fig. 1. Each community C of GO terms in the gene-based Bonferroni network can be therefore characterized in terms of 163 attributes. In Table 3 we report the over-expressions for the five largest communities. The full list of characterizations is given in Table SM4 of the Supplementary material.

Table 3 shows that in the same communities of the gene-based Bonferroni network one can have terms that belong to two different GO branches, thus confirming what we had already seen in Section 3.2. For example, the 7th gene-based community, of size 27, is characterized by two attributes: the semantic community 8, involving terms of the BP branch and the semantic community 12, involving terms in the CC branch. From Table SM2 in the Supplementary material we can see that the semantic communities 8 and 12 both involve terms dealing with signaling pathways and the related receptors. Thus the Bonferroni network communities provide a way to put together terms involved in similar biological function whereas disconnected from a semantic point of view. In other words, these examples show how there exist GO terms that have no semantic link between each other and can nevertheless be put in connection when the gene content of their children terms

Table 3
Statistical characterization of the communities of the Bonferroni gene-based network. Communities have been obtained by using the Infomap algorithm [19]. The characterization has been done by using the methodology of Ref. [21]. For each term T we have considered as attribute the membership of T to one of the communities of the adjacency semantic network. Only the results for the five largest communities are shown. The full list of characterizations is given in Table SM4 of the Supplementary material.

Bonferroni gene-based community	Size	Overexpressed attribute
4	40	*Community 8 (BP terms)*
7	27	*Community 8 (BP terms)*
7	27	*Community 12 (MF terms)*
17	26	*Community 42 (BP terms)*
17	26	*Community 39 (MF terms)*
15	23	*Community 44 (BP terms)*
15	23	*Community 17 (BP terms)*
15	23	*Community 14 (BP terms)*
13	23	*Community 60 (BP terms)*

is considered. If we think to the attributes as defining homogeneous and distinct subsets of the GO, such subsets might be disconnected from a semantic point of view and connected when the gene content of their terms is taken into account. This is 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 by the GO semantics.

3.5. The modular structure of GO: an analysis of the spectral properties of the correlation matrix of GO-terms

The fact that a considerable part of significant links in the Bonferroni network connects terms belonging to the same branch, as pointed out in Table 2, drove us to investigate the modular structure of the correlation matrix between GO-terms. While in the previous section we had to focus the analysis on the SVN, because the community analysis approach is not feasible for the whole gene-based network, here we are able to analyze the correlation matrix of the entire gene-based network.

A correlation coefficient between two terms can be calculated according to the number of common genes [24]. Specifically, we can describe both terms, A and B, as binary vectors of length equal to the number of genes classified in the considered release of the Gene Ontology, that is, $N_g = 18{,}922$. Any component of a vector is equal to 1 if the gene corresponding to that component belongs to the term that the vector describes, and it is equal to 0 otherwise. The correlation coefficient between the two terms is then calculated as the sample correlation coefficient between the corresponding binary vectors, and it can be calculated through a very simple equation. In particular, let us indicate the total number of genes classified in terms of A (B) with n_A (n_B), and the number of genes classified in both terms with n_{AB}. The correlation coefficient between the two terms can then be written as

$$\rho_{AB} = \frac{n_{AB} - \frac{n_A\, n_B}{N_g}}{\sqrt{n_A\left(1 - \frac{n_A}{N_g}\right) n_B\left(1 - \frac{n_B}{N_g}\right)}} \tag{3}$$

which is a straightforward adaptation of the Pearson correlation coefficient to the case of binary variables. The main advantage of using Eq. (3) to measure the similarity between two terms is that the correlation coefficient removes the spurious influence of the degree of terms in the measure of similarity. For instance, if we consider a simpler measure of similarity, such as the Jaccard coefficient defined as

$$J_{AB} = \frac{n_{AB}}{n_A\, n_B}, \tag{4}$$

we immediately note that such a quantity is always positive, while the correlation coefficient of Eq. (3) can also be negative when the actual number of common genes, n_{AB}, is lower than the number of common genes expected in the random case, i.e. $\frac{n_a\, n_b}{N_g}$. Such an expected number is obtained by assuming that the two vectors, A and B, are binary random vectors with n_A and n_B components different from 0, respectively. Furthermore, the correlation coefficient given in Eq. (3) can be interpreted in light of the null hypothesis of randomness that we use throughout the paper. Indeed, it can be shown that

$$\rho_{AB} = \frac{1}{\sqrt{N_g - 1}}\, z_{AB}, \tag{5}$$

where z_{AB} is the standard score associated with observation n_{AB} according to the hypergeometric distribution [25,26]. Therefore, the correlation coefficient provided in Eq. (3) measures the (standardized) deviation of the observed number of common genes, n_{AB}, from expected number of common genes, according to the null hypothesis described by the hypergeometric distribution.

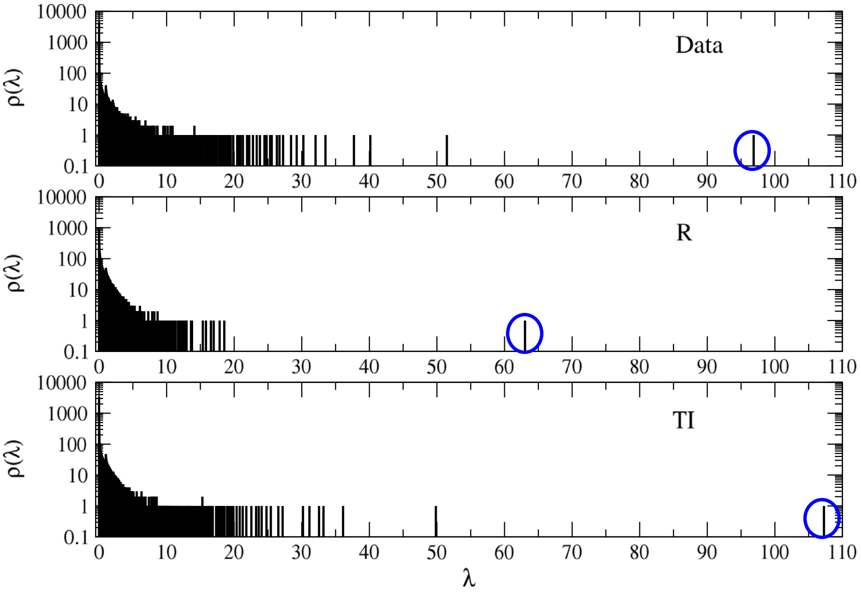

Fig. 8. Spectral density of the correlation matrix of terms for three systems, (i) the original data of GO (top panel), (ii) a random rewiring of the original system (middle panel), and (iii) an intra-branch target random rewiring (bottom panel).

The $T \times T$ ($T = 12{,}564$) matrix of correlation coefficients, Γ, is the correlation matrix of GO-terms. In this section, we shall investigate the modular structure of matrix Γ according to its spectral properties. In the top panel of Fig. 8, the spectral density of matrix Γ is reported, and the largest eigenvalue is highlighted. Such a density should be compared with the density of the correlation matrix of a system that fully preserves the heterogeneity of terms and nodes observed in real data and does not present any modular structure. This can be obtained by performing a random rewiring[2] of the original bipartite network of GO-terms and genes, and then calculating the associated correlation matrix of terms, Γ_R. The spectral density of Γ_R is reported in the second panel from the top of Fig. 8. It is apparent from the figure that the spectral densities of Γ and Γ_R are rather different. Specifically, the largest eigenvalue of Γ_R is smaller than the largest eigenvalue of Γ, and, more important, the shape of the bulk of the two densities is significantly different.

It is also useful to compare the spectral density of the correlation matrix of real data with the spectral density of a system that, again, preserves the heterogeneity of both terms and genes, but in which genes are mostly classified in terms that belong to only one of the three main branches of GO. Such a system, which should display an enhanced modular structure with respect to real data, has been obtained by first associating a target branch (MF, CC, BP) with each gene, according to the branch that contained the maximum number of terms in which the gene is classified in the real data,[3] and then performing a random rewiring where a single edge-switch is accepted if and only if, after the switch, one of the two involved genes has increased the number of terms belonging to the target branch in which it is classified, while the switch is at least neutral for the other gene. For instance, let us suppose that link $g_{i,BP} - T_{j,CC}$ is selected at a certain time step of the rewiring process. That link indicates that gene $g_{i,BP}$, with target branch BP, belongs to term $T_{j,CC}$, which is a term of branch CC, at that time. At the same time step, another link is randomly selected, say, $g_{m,CC} - T_{n,MF}$ between gene $g_{m,CC}$ with target branch CC and term $T_{n,MF}$ of branch MF. The edge switch action would require one to replace those two links with $g_{m,CC} - T_{j,CC}$ and $g_{i,BP} - T_{n,MF}$. In this example, the edge switch outcomes would increase the number of CC terms in which $g_{m,CC}$ is classified, which is positive, because CC is the target branch of $g_{m,CC}$. On the other hand, the status of gene $g_{i,BP}$ remains the same, with respect to its target branch, because neither $T_{n,MF}$ nor $T_{j,CC}$ are terms of its target branch, BP. Therefore the edge switch is accepted in this case. The correlation matrix associated with such a target intra-branch random rewiring is indicated with Γ_{TI}, and its spectral density is reported in the bottom panel of Fig. 8. The figure shows that the spectral density of Γ_{TI} is very similar to the one observed in real data, suggesting that the actual structure of GO is rather modular.

The conclusion about the modular structure of GO is also supported by an analysis of the components of the eigenvector associated with the largest eigenvalue of correlation matrices Γ, Γ_R, and Γ_{TI}. In Fig. 9 the components of the largest eigenvector, for the three systems, are reported after ordering terms according to the number of genes that they include, as shown in the bottom panel of the figure, extracted for convenience from Fig. 2. The color of vector components in the figure depends on the GO-branch corresponding to each term, namely, the component is *black* if the term belongs to BP, *red* if it

[2] We obtained a random rewired network by performing 10^7 local edge-switches.

[3] In case of degeneracy, a random selection has been performed.

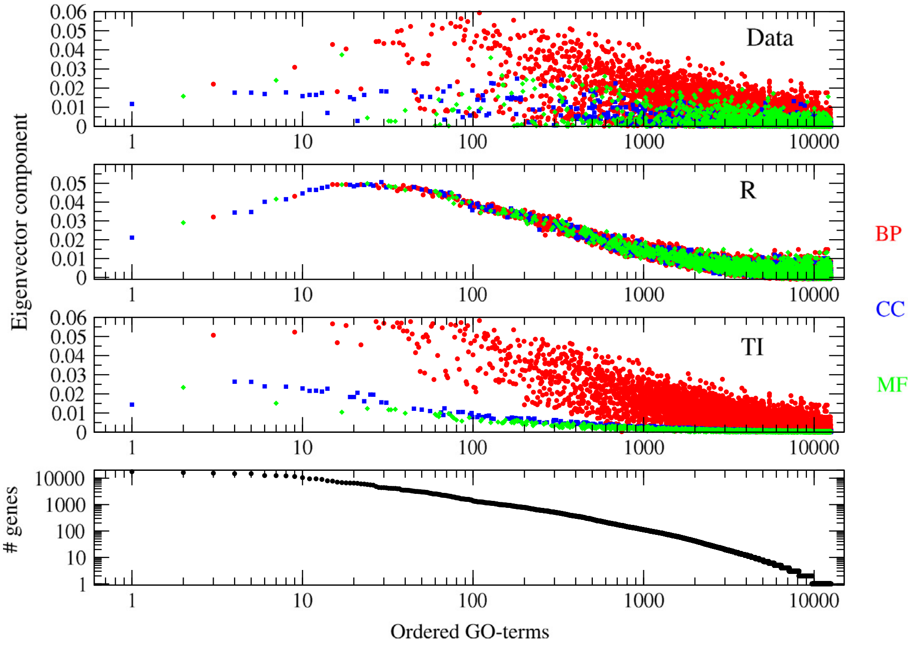

Fig. 9. Components of the eigenvector associated with the largest eigenvalue of correlation matrices Γ (top panel), Γ_R (second panel from the top), and Γ_{TI} (third panel from the top). Bottom panel: number of genes classified in each GO-term. Terms are ordered in decreasing order with respect to the number of genes they include (bottom panel), and the same ordering of terms is used in the top three panels.

belongs to CC, and *green* if it does belong to MF. The figure shows that the largest contribution to the largest eigenvalue comes from BP terms in both real data (matrix Γ) and target intra-branch rewiring (matrix Γ_{TI}), while no significant difference between the three branches is observed in the correlation matrix Γ_R associated with the random rewiring.

To better understand the modular structure of GO, we focused our attention on the Bonferroni network of terms, as they represent the core of the gene-based network structure. Fig. 10 shows the heatmap of the correlation matrix restricted to the terms belonging to the Bonferroni network. Terms are ordered with respect to the dendrogram obtained with average linkage hierarchical cluster analysis of the gene content profile of the terms. The ordered list of GO terms is reported in Supplementary Table SM5, together with the information about the GO branch and the community the terms belong to. It is noticeable that terms belonging to the same GO branch tend to cluster together. Nevertheless, the discernible blocks along the diagonal can be put in relation with the communities found in Section 3.4. This is an expected result, since both the analysis seek for groups of similar terms by considering the gene content of the terms, even if from different perspectives. Indeed, we observe again that, despite the structure modular in BP, CC and MF, the blocks/communities are generally composed by terms from different GO branches.

In conclusion, the analysis presented in this section indicates that the structure of GO is rather modular in terms of the three main branches, BP, CC, and MF. However, such a modular structure merges with a coexistent structure revealed by analyzing the similarities between the gene content of the terms. While the first structure is merely a drawback of the classification process inherent in the Gene Ontology, we believe that the second one might reflect genuine properties of the biological systems.

3.6. Over/under representation of genes in the three GO branches

One of the main results obtained so far is that the three GO branches (BP, MF, CC), which are semantically disjoint, can be put the one in relation with another when considering the gene content of each GO term.

Since the null hypothesis we used so far, in order to test the statistical significance of the joint gene content of terms, relies upon the assumption that genes are recorded randomly across the three branches of GO, that is, BP, CC, and MF, in this section, we investigate whether and at which extent such an analysis might be affected by the asymmetry in the recording of genes across these three GO branches. Specifically, we tried to assess whether each gene was more present in BP, in MF or in CC in a statistically significant way. Let us consider a specific gene g, and count the number of terms gene g belongs to, separately in the BP, CC, and MF branches. We indicate such three numbers as $n_g(BP)$, $n_g(CC)$, and $n_g(MF)$. The sum of such three quantities, $n_g = n_g(BP) + n_g(CC) + n_g(MF)$, represents the total number of terms where gene g is annotated. Other relevant quantities are the total number of terms in each branch, $N(BP)$, $N(CC)$, and $N(MF)$, and their sum, N, which represents the total number of terms. If we assume that gene g is recorded in n_g terms randomly selected among the N terms then the probability to observe $n_g(BP)$, $n_g(CC)$ and $n_g(MF)$ is given by a multivariate hypergeometric

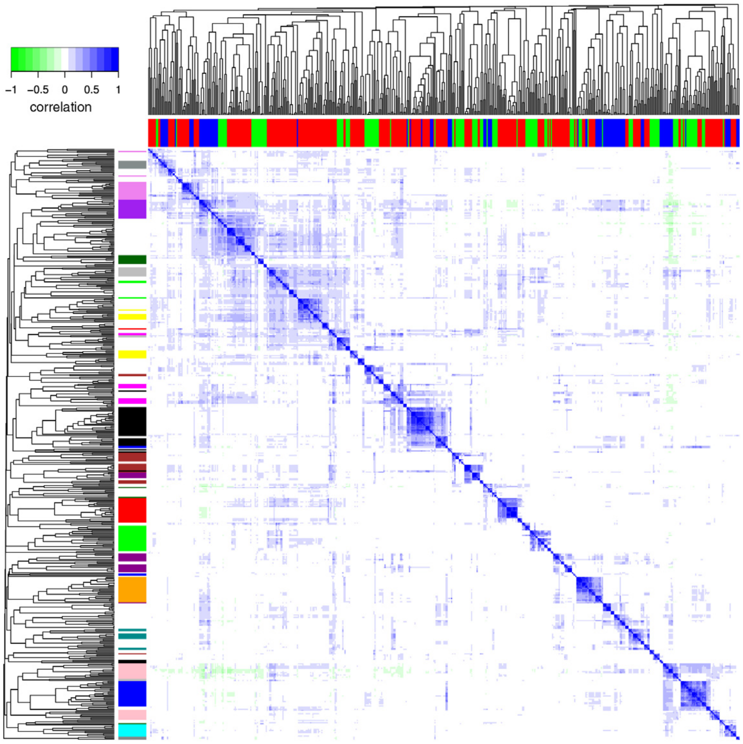

Fig. 10. Heat map of the symmetric cross correlation matrix among Bonferroni network terms. Each cell of the matrix represents the correlation between the respective two terms, with a color legend shown in the top left corner (green: negative correlation, blue: positive correlation). Terms are classified with respect to the GO branch they belong, with the top horizontal colored bar (red: BP, blue: CC, green: MF). With the left vertical colored bar we identify terms belonging to the same community, as each color represents a different community. Only the top 17 most populated communities have been highlighted here. The two dendrograms are identical, and are obtained with average linkage hierarchical cluster analysis of the terms' gene content profiles. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

distribution:

$$P(n_g(BP), n_g(CC), n_g(MF)) = \frac{\binom{N(BP)}{n_g(BP)}\binom{N(CC)}{n_g(CC)}\binom{N(MF)}{n_g(CC)}}{\binom{N}{n_g}}. \tag{6}$$

We used such a null hypothesis to test whether a gene is over-represented or under-represented in each one of the three branches of GO. Calculated p-values have been corrected for multiple hypothesis testing by using the Bonferroni correction. In Fig. 11 we show the distribution of genes inside the three GO branches. With black numbers we indicate the number of all annotated genes. With red numbers we indicate the statistically significant genes when the annotations takes into account the genes inherited because of the semantic relations among terms. With green numbers we indicate only the genes that are annotated in the most specific term.

If we consider all the gene-to-term annotations, by including the not explicit ones which are inherited because of the semantic relations among terms, we find a huge amount of statistically significant genes (red numbers in Fig. 11), ∼40% of all the genes (black numbers in Fig. 11), which is a result definitely against our null hypothesis.

However, we will show that this analysis is strongly biased by the different hierarchical structure of BP, CC, and MF branches. In Fig. 12, the fraction of terms with a given hierarchical level is plotted, separately, for the terms of each of the GO branches. The hierarchical levels of a term T_i are measured here in the following way. We consider all the possible paths that connect term T_i to the root node (GO), and take the length of all these paths, that is, the number of ancestor terms $M_A(T_i)$ in each one of the different paths that connect T_i and the root node. Therefore several values of $M_A(T_i)$ can be associated with T_i, as a consequence of the existence of paths with different length that connect T_i to the root. For instance, the hierarchical level of terms BP, CC, and MF is just 1, and they are the only terms with such an indentation level. Fig. 12 reveals the different hierarchical structure of the three branches. In CC, each term is associated in average with 5.51 indentation levels and more than 70% of terms is simultaneously associated with indentation levels 5, 6, 7 and 8. When we consider only the maximum of the indentation levels associated to each term, see Fig. 12 right panel, CC terms show a wide distribution of levels, ranging from 5 to 14. This behavior is clearly explained by the observation that the semantic links in the CC branch connect terms with very different maximum indentation levels, see Fig. 13 first panel. On the other end, both the distributions of indentation levels in MF have one narrow peak on level 6 and on average only 1.31 indentation levels are associated with each term.

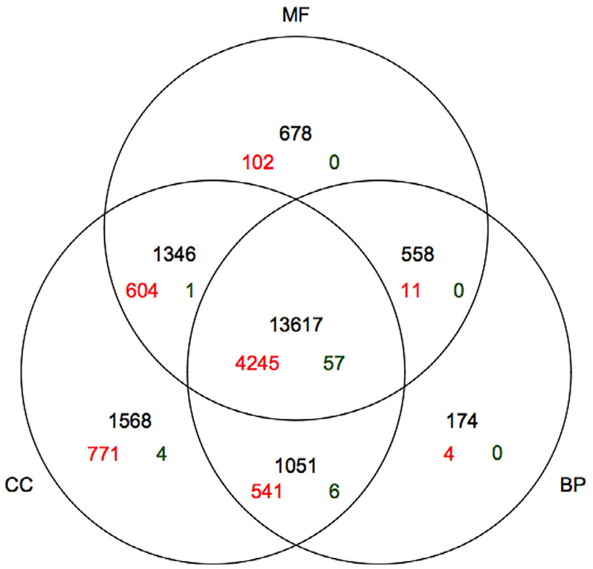

Fig. 11. Graphical representation of the number of genes in the three GO branches. Black: all genes; Red: statistical significant genes; Green: statistically significant genes when only the most specific term annotations are considered. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

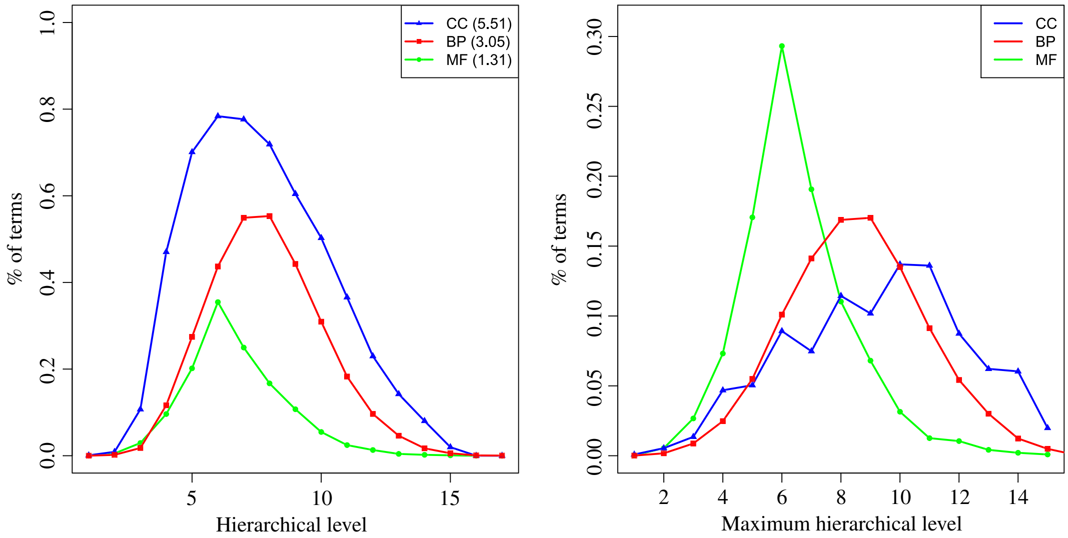

Fig. 12. Fraction of terms per hierarchical level in BP, CC and MF. The left panel considers all the possible hierarchical levels associated with the terms. The numbers in parenthesis represent the average number of hierarchical levels associated with the terms of each branch. The right panel considers only one hierarchical level per term, i.e. its maximum hierarchical level.

In this case, Fig. 13 shows that semantic links connect prevalently terms with a difference in maximum indentation level equal to 1, that is what would be expected for a simple tree-like classification. The BP branch shows a behavior halfway the CC and the MF branches.

To tackle the impact of the heterogeneous hierarchical structure of BP, CC, and MF on the analysis of the over-representation of a gene in one or more of the three branches of GO, we have performed the analysis by considering for each gene only annotations in terms that do not belong to the same path toward the root. When a gene is annotated in two related terms (i.e. one of them is an ancestor of the other) we choose to consider only the most specific one, i.e. the term with highest indentation level. In this way, we avoid any effect due to redundant annotation and hierarchical structure differences. Surprisingly, only 68 genes, among 18 992, came up statistically significant with a p-value lower than 1%, see green numbers in Fig. 11. The majority of the significant genes are over represented in CC terms. This result can be related to the fact that the percentage of genes with annotations in CC is higher with respect to the other branches. Nevertheless, these results support the hypothesis that the genes are recorded randomly across the GO branches, but then, when we add

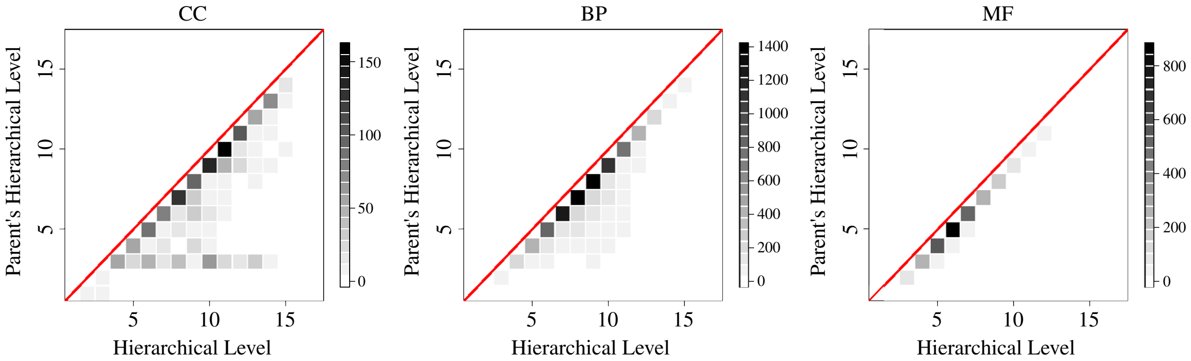

Fig. 13. Heatmap of the number of semantic links between couples of terms characterized by their maximum hierarchical levels in CC, BP and MF. *X* and *Y* axes represent the hierarchical level of the child and the parent term respectively. Each cell is colored based on the number of semantic links connecting two terms with the correspondent *X* and *Y* indentation levels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

the semantic links information to transfer the annotations to all the parent terms, we unbalance the annotations because of the differences in GO branches structure.

4. Conclusion

In this paper we have considered the Gene Ontology from a complex network perspective. We have considered two types of GO networks: the one associated to the semantic structure of the Gene Ontology and the one associated to the gene content of each GO term. We have first considered some basic network metrics showing that the main differences between the two networks are in the number of links and not in the relative importance of the terms within the network. We have then compared the two networks at the level of network communities. In the adjacency semantic network we detected 163 communities that are homogeneous from a biological point of view. This has been confirmed by a statistical analysis able to provide the 3-words that are over-expressed in the obtained communities. In the Bonferroni gene-based network we detected 74 small communities showing that there exist GO terms that have no semantic link between each other and can nevertheless be put in connection when the gene content of their children terms is considered, see Section 3.4. The level of modularity of the gene-based network has also been addressed in Section 3.5 by considering the spectral properties of the whole gene-based network and not only its statistically validated counterpart. We also investigated in Section 3.6 the role of the gene annotations in the GO terms in determining 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.

We have therefore shown how a deeper analysis of GO terms, based on their gene content, might reveal relationships between terms that are missed by looking at the semantic structure of GO. This has a practical importance for the evolution and maintenance of GO. In fact this kind of analysis can be useful to capture the relations amongst genes to be profitably transferred at the level of terms. However, this is also 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 by the GO semantics.

As a by-product, we have devised a simple methodology, based on the detection of the statistical significant 3-words, that allows to have a first glance insight about the biological meaning of groups of GO terms. This might become a routinary tool able to provide a first-glance biological interpretation of groups of GO terms.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.physa.2016.03.062.

References

[1] M. Ashburner, C.A. Ball, J.A. Blake, et al., Gene ontology: tool for the unification of biology. The gene ontology consortium, Nature Genet. 25 (2000) 25–29.
[2] S. Leonelli, A.D. Diehl, K.R. Christie, M.A. Harris, J. Lomax, How the gene ontology evolves, BMC Bioinformatics 12 (2011) 325–331.
[3] G. Alterowitz, M. Xiang, D.P. Hill, J. Lomax, J. Liu, M. Cherkassky, J. Dreyfuss, C. Mungall, M.A. Harris, M.E. Dolan, J.A. Blake, M.F. Ramoni, Ontology engineering, Nature Biotechnol. 28 (2010) 128–130.
[4] A.D. Diehl, J.A. Lee, R.H. Scheuermann, J.A. Blake, Ontology development for biological systems: immunology, Bioinformatics 23 (2007) 913–915.
[5] C. Coronnello, M. Tumminello, S. Miccichè, R.N. Mantegna, Networks in biological systems: An investigation of the gene ontology as an evolving network, Nuovo Cimento C 32 (2009) 157–160.
[6] M. Tumminello, S. Miccichè, F. Lillo, J. Piilo, R.N. Mantegna, Statistically validated networks in bipartite complex systems, PLoS One 6 (2011) e17994.
[7] P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in: IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence, ISBN:1-55860-363-8, 1995, pp. 448–453.

[8] S.G. Lee, J.U. Hur, Y.S. Kim, A graph theoretic modeling on GO space for biological interpretation of gene communities, Bioinformatics 20 (2004) 381–388.
[9] N. Speer, C. Spieth, A. Zell, Spectral communitying gene ontology terms to group genes by function, in: R. Casadio, G. Myers (Eds.), WABI 2005, in: LNBI, vol. 3692, 2005, pp. 1–12.
[10] J.Z. Wang, Z. Du, R. Payattakool, P.S. Yu, C.-F. Chen, A new method to measure the semantic similarity of GO terms, Bioinformatics 23 (2007) 1274–1281.
[11] H. Frohlich, N. Speer, A. Poustka, T. Beissbarth, GOsim—An R-package for computation if information theoretic GO similarities between terms and gene products, BMC Bioinformatics 8 (2007) 166.
[12] R.M. Othman, S. Deris, R.M. Illias, Z. Zakaria, S. Mohamad, Automatic communitying of gene ontology by genetic algorithm, World Acad. Sci. Eng. Technol. 31 (2007) 37–46.
[13] Z. Du, L. Li, C.-F. Chen, P.S. Yu, J.Z. Wang, G-SESAME: web tools for GO term based gene similarity analysis and knowledge discovery, Nucleic Acids Res. 37 (2009) W345–W349.
[14] D.W. Huang, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 41 (2009) 44.
[15] W. Feller, An Introduction to Probability Theory and its Applications, Vol. 1, third ed., Wiley, New York, 1968.
[16] R.G. Miller, Simultaneous Statistical Inference, second ed., Springer-Verlag, New York, 1981.
[17] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, J. R. Stat. Soc. B 57 (1995) 289–300.
[18] S. Fortunato, Community detection in graphs, Phys. Rep. 486 (2010) 75–174.
[19] M. Rosvall, C.T. Bergstrom, Maps of random walks on complex networks reveal community structure, Proc. Natl. Acad. Sci. USA 105 (2008) 1118–1123.
[20] J. Duch, A. Arenas, Community detection in complex networks using extremal optimization, Phys. Rev. E 72 (2005) 027104.
[21] M. Tumminello, S. Miccichè, F. Lillo, J. Varho, J. Piilo, R.N. Mantegna, Community characterization of heterogeneous complex systems, J. Stat. Mech. 2011 (2011) P01019.
[22] G. Saporta, G. Youness, Comparing two partitions: Some proposals and experiments, in: Compstat, Springer-Verlag HD, 2002, pp. 243–248.
[23] L. Danon, A. Diaz-Guilera, J. Duch, A. Arenas, Comparing community structure identification, J. Stat. Mech. 2005 (2005) P09008.
[24] M. Tumminello, C. Edling, F. Liljeros, R.N. Mantegna, J. Sarnecki, The phenomenology of specialization of criminal suspects, PLoS ONE 8 (5) (2013) e64703. http://dx.doi.org/10.1371/journal.pone.0064703.
[25] V. Hatzopoulos, G. Iori, R.N. Mantegna, S. Miccichè, M. Tumminello, Quantifying preferential trading in the e-MID interbank market, Quant. Finance 15 (2015) 693.
[26] A. Fiasconaro, M. Tumminello, V. Nicosia, V. Latora, R.N. Mantegna, Hybrid recommendation methods in complex networks, Phys. Rev. E 92 (2015) 012811.