IL NUOVO CIMENTO Vol. 32 C, N. 2 Marzo-Aprile 2009
DOI 10.1393/ncc/i2009-10386-7

Colloquia: CSFI 2008

Networks in biological systems: An investigation of the Gene Ontology as an evolving network

C. Coronnello, M. Tumminello, S. Miccichè and R. N. Mantegna
Dipartimento di Fisica e Tecnologie Relative, Università di Palermo
Viale delle Scienze, Ed. 18, 90128 Palermo, Italy

(ricevuto il 22 Giugno 2009; pubblicato online il 4 Settembre 2009)

Summary. — Many biological systems can be described as networks where different elements interact, in order to perform biological processes. We introduce a network associated with the Gene Ontology. Specifically, we construct a correlation-based network where the vertices are the terms of the Gene Ontology and the link between each two terms is weighted on the basis of the number of genes that they have in common. We analyze a filtered network obtained from the correlation-based network and we characterize its evolution over different releases of the Gene Ontology.

PACS 87.18.-h – Biological complexity.

1. – Introduction

Biology and biomedical sciences have changed their status from disciplines with a low rate of experimental data production to disciplines with a huge rate of experimental data production. The explosion of available data has allowed researchers to move from a description of phenomena involving single units or processes to the description of much more complex systems where a large number of elements interact among them in a way that can be suitably described in terms of a network [1]. For illustrative reasons, we wish to refer to a very limited set of representative examples. Specifically, biological and biomedical networks have been detected in gene regulation [2], gene regulation of model organisms [3], internal organization of the cell [4, 5] and disease genes [6].

In some investigations devoted to the discovery or construction of biological and biomedical networks the relationship between networks of genes and biological tasks has been obtained by using the Gene Ontology (GO) database, which is a controlled vocabulary where sets of genes and gene attributes are organized. One of the most common uses of the GO consists in the detection of the ontology terms, which are enriched with respect to a list of genes of interest [7]. The statistical assessment of the enrichment of an ontology term is however affected by the fact that the Gene Ontology database is continuously updated. In fact, some ontology terms can disappear from a release to the successive, other terms can be added, and relations amongst terms can be modified. On the other hand, such process can also affect the classification of genes. In fact, some genes that are missing in a certain release of the GO can be annotated in successive releases, while other genes that are annotated in some terms of a certain release can move out from those terms, and possibly from the GO, in the next GO releases. The conclusion of such observations is that terms, which are assessed as enriched in a GO investigation performed by using a specific GO release, might turn out not to be enriched in successive releases and vice versa.

In the present paper we aim to present preliminary results on how the Gene Ontology database evolves over time. We consider the Gene Ontology as a weighted network in which the nodes are the terms and a link between each two terms is weighted according to the number of genes that are annotated in both terms. After the construction of the network, we filter the complete correlation-based network by using the average linkage clustering algorithm, in order to get quantitative information about the evolving structure of the GO. Specifically, we construct the average linkage minimum spanning tree (ALMST) [8]. This tree clearly shows that there is a general tendency of the GO to increase its compactness over time. Indeed, the diameter of these trees generally diminishes moving from one release of GO to the next. This tendency toward a more compact structure does not come together with a significant variation of the degree distribution of the trees, suggesting that a more compact structure is obtained by modular rearrangement of the network. In this paper we present only results obtained with the ALMST. We have also applied the single linkage clustering algorithm to the GO network and extracted its minimum spanning tree. The results obtained with this last approach are similar to the present ones [9].

2. – The GO correlation-based network

The Gene Ontology project [10] provides a controlled vocabulary as a way to organize gene and gene product attributes in a knowledge base. The easiest way to understand how GO works is to use the web service AmiGO, provided by the GO web site, `www.geneontology.org`. Terms, together with their relationships, are the skeleton of the ontology. In the GO each term has a unique numerical identifier of the form GO:nnnnnn, and a name consisting in phrase representing a key concept in molecular biology. Terms are organized following three principles: i) the biological process, ii) the molecular function, and iii) the cellular component. The three ontologies are structured as directed acyclic graphs, which are similar to hierarchies, but differ in that a more specialized term can be related to more than one less specialized term.

In order to show how the GO evolves in time we have considered the 18 GO releases described hereafter. The GO consortium provides a large archive containing all of the past releases of GO. In the archive there are two different kinds of files, both useful for our analysis. There are files related to the structure of the ontology, reporting terms and relationships and there are files in which all the gene annotations are reported. The two types of file are often non-synchronous. Since we needed to use both these files, we made a synchronization by choosing the files closest to the end of each month in a quarterly period from 29 March 2004 to 28 June 2008. We downloaded only annotation files related to the Homo Sapiens species. One problem we had to solve was to standardize the genes' ID used in the different releases. In fact, when the gene ID changes, in the successive releases the annotations will be written using the newest gene ID. We were able to find a table of synonyms for the UniProt symbols gene ID, directly from the UniProt web site [11]. The annotations characterized with a UniProt symbol ID are approximately 80% of the total.

The GO can be described as a network where the vertices are the terms and the links are the relationships, *e.g.*, either *is a* or *part of*, between terms. It is a knowledge-based network, constructed by the curators of the GO Consortium. We have instead tried to describe the properties of the GO by constructing a network of GO terms based on the annotations of genes. In fact, we consider a weighted network, where all the terms of the GO are linked with each other, with links that are weighted according to the number of genes the two terms have in common. Of course, all the relationships introduced in the GO by the curators have a corresponding link with a positive weight because if two terms are linked in the knowledge-based network they have in common all the genes annotated in the more specific term. However, there can also be a link with positive/negative weight between terms without any kind of relationship built up by the GO curators, either direct or inherited. We evaluated the weight of the links in the following way. We consider the list of $N = 35483$ genes annotated in at least one of the 18 releases. It should be noted that we consider only the genes named with a UniProt ID and if the same gene was named with different names in two different releases, we count this gene only once. Finally, the weight of the link between two terms is evaluated by using

$$\rho(i,j) = \frac{n_{ij} - \frac{n_i n_j}{N}}{\sqrt{n_i \left(1 - \frac{n_i}{N}\right) n_j \left(1 - \frac{n_j}{N}\right)}}\,, \tag{1}$$

where n_i is the number of genes annotated in the term i, n_j is the number of genes annotated in the term j and n_{ij} is the number of genes annotated in both the terms i and j.

We observe that, in all the 18 releases, more than 90% of the correlations are negative. It means that more than 90% of pairs of terms have a number of genes in common n_{ij} lower than the number of genes expected by assuming a random distribution of genes across the terms $\frac{n_i n_j}{N}$. We also observe that a large number of pairs of terms has a correlation equal to 1, meaning that such terms contain exactly the same genes. On one side the redundancy can be due to an effective multiplicity of the role of some genes. In fact, some genes can have more than one function and the same group of genes can possibly be annotated in more than one term with a different biological meaning. On the other hand, the redundancy can be also a characteristic of the GO. In fact the GO contains terms with close biological meaning that can share the same genes in a given release. We decided to unify all the terms with the same gene content in a single *meta-term*, in order to simplify the network representation. We filter the correlation-based network of the GO by generating the ALMST for each of the 18 available releases. Details about the filtering protocol used to generate the ALMST can be found in ref. [8]. In fig. 1 we show the ALMSTs of the first (panel A) and the last release (panel B). To quantify in a simple way the evolution of the network, we have calculated the percentage of the number of vertices present in the shortest path used to estimate the diameter of the filtered tree with respect to the total number of vertices in the network. This percentage is reported in panel C of fig. 1. We notice that this percentage is decreasing from a value of 12.7% observed in the first release to a value of 4.3% in the last release. In other words we observe that the diameter of the ALMST shows a significant relative decrease over time. When the relative diameter of a network decreases, the network appears more dense and compact as shown in fig. 1.

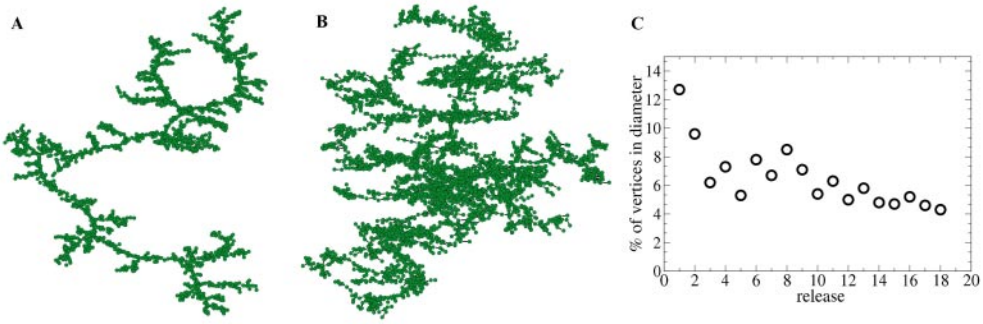

Fig. 1. – In panel A we report the ALMST obtained from the correlation-based network of the 1st release. In panel B we report the ALMST obtained from the correlation-based network of the 18th release. In panel C we plot the percentage of the number of vertices present in the shortest path used to estimate the diameter of the ALMST with respect to the total number of vertices in the network. All graphs refer to the biological process ontology.

3. – Conclusions

The GO is characterized by a dynamics over different releases, which is altering the topological organization of GO terms. Any study using the GO should consider this dynamical aspect especially when an enrichment of a specific term is statistically assessed starting from the selection of a list of genes of interest. This aspect can be especially relevant in the study of complex diseases where one can move from a search for genes whose alterations induce the disease to the search of biological processes, molecular functions and cellular components (*i.e.* ontology terms) that can induce the disease [12].

REFERENCES

[1] Albert R. and Barabasi A.-L., *Rev. Mod. Phys.*, **74** (2002) 47.
[2] Levine M. and Davidson E. H., *Proc. Natl. Acad. Sci. U.S.A.*, **102** (2005) 4936.
[3] Tong A. H. Y. *et al.*, *Science*, **303** (2004) 808.
[4] Jeong H. *et al.*, *Nature*, **411** (2001) 41.
[5] Barabasi A.-L. and Oltvai Z. N., *Nat. Rev. Genet.*, **5** (2004) 101.
[6] Goh K.-I. M. *et al.*, *Proc. Natl. Acad. Sci. U.S.A.*, **104** (2007) 8685.
[7] Draghici S., *Data Analysis Tools for DNA Microarrays* (Chapman & Hall) 2003.
[8] Tumminello M. *et al.*, *Int. J. Bifurcation Chaos*, **17** (2007) 2319.
[9] Coronnello C., *Gene Ontology statistical analysis of complex diseases: methodological issues and applications to Autism*, PhD Thesis, Palermo University 2009.
[10] The Gene Ontology Consortium, *Nature Genet.*, **25** (2000) 25.
[11] http://www.uniprot.org/help/uniprotkb.
[12] Romano V. *et al.*, *Autism Spectrum Disorders: from candidate genes to candidate ontology terms*, in *Causes and Risk Factors for Autism*, edited by Giordano A. C. and Lombardi V. A. (Nova Science Publishers, Inc, Hauppage NY) 2008, pp. 33-50.