The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 110–120

Detecting Community Structures in Gene Networks using Gene-Ontology Annotation

Liuhuan Dong[1] Winking Q. Yu[2]
Andreas M. W. Dress[1,*]

[1] CAS/MPG Partner Institute for Computational Biology,
SIBS, CAS, Shanghai, 200031, China.
[2] Center for Combinatorics, LPMC, Nankai University,
Tianjin, 300071, China.

Abstract In this note, we explore — by means of a specific case study — the potential of a new approach towards identifying gene functional modules in *Gene-Ontology* (or, for short, *GO*) networks, i.e., networks whose nodes are formed by the genes from a given collection of genes under consideration while two such nodes are connected by an edge if and only if they share — according to their *GO annotation* — at least one gene attribute.

We construct such a network for genes from the Zebra-Fish-genome data base. Then, to detect the implied GO-based gene communities, we apply the *FastCommunity* heuristics developed by Aaron Clauset *et al* and, as an alternative, a *Linear-Programming*-based method for community detection recently developed by William Chen *et al*, we discuss the biological significance of the gene communities identified by these methods and, finally, we discuss the potential of community-detection methods as general tools for (i) identifying gene functional modules on the basis of GO data as well as (ii) for deriving insights regarding such modules even if, instead of reliable and comparatively detailed and sophisticated GO annotations, only comparatively primitive "yes-no" data about functional attributes are available, corroborating that community-detection methods may actually be useful for constructing (or checking) GO annotations on the basis of much less sophisticated primary data.

1 Introduction

In this note, we explore the potential of a new approach towards identifying putative gene-function modules. The approach is based on applying methods for *community-structure detection* to *GO networks*, i.e., networks whose nodes are formed by the genes from a given collection of genes under consideration while two such nodes are connected by an edge if and only if they share — according to their *GO annotation* — at least one gene attribute.

*The first-named author initiated and did most of the actual work, but received some help including suggestions about how to run the CLPEX software package, how to set up and to interpret the GO-network and its community structures, and how to present the results from the other two authors.Liuhuan Dong: *dlh@picb.ac.cn*, Winking Q. YU: *yuqiang@mail.nankai.edu.cn*, Andreas M. W. Dress: *andreas@picb.ac.cn*.

As is well known, many systems can be represented in terms of networks. Over the last few years, various types of biological networks such as metabolic [12], protein-protein interaction [19], and food-web networks [4] have proved to be rather useful models for representing our knowledge about biological systems and investigating their properties — in particular, if this knowledge is rather rudimentary and no detailed information about the exact mechanisms of interaction between the various agents participating in the network's activity is available. Though in no way deterministic, many of these networks have been found to share certain statistical properties not common to standard random networks.

E.g, they have been found to exhibit surprisingly low average distances (the *small world effect*, [2, 20]), right-skewed degree distributions (suggesting some kind of intrinsic *scale freeness*, [1, 3]), and high transitivity (*friends of friends tend to be each other's friends, too*, cf. [13]). These findings motivated the proclamation of certain universally applicable laws regarding the nature of *real-world networks* and caused much speculation about the potential *semi-random* mechanisms that may lead to the formation of such networks.

Related to transitivity, another network feature has been emphasized recently: The *community structure* of networks. Until now, there is no universally accepted clear-cut definition of this concept. However, many scientists, like Girvan *et al* [11], Newman *et al*, [16], and Radicchi *et al* [17] have made important contributions to this area in recent years. According to their papers, a community structure of a network is a partition of the network's agents into disjoint *communities* consisting of agents that appear to strongly interact with each other — and not so strongly with those in the other communities. To detect such partitions, we shall apply the *FastCommunity* heuristics developed by Aaron Clauset *et al* [8] and, as an alternative, a *Linear-Programming*-based method for community detection recently developed by William Chen *et al* [6, 7].

The other important concept that we will employ is lined out in "*http://www.GeneOntology.org/*": According to this web site, the Gene Ontology (or, for short GO) project intends to provide a "controlled vocabulary to describe gene and gene-product attributes in any organism". In its edition from December 12, 2006, Gene Ontology [9] contains $21,909$ terms indexed by their GO IDs, 96.1% coming with explicit definitions, of which $12,549$ refer to biological processes, $1,847$ to cellular components, and $7,513$ to molecular function. These terms are organized, from the rather general to more and more specialized concepts, in the form of a directed acyclic graph, thus giving rise to a hierarchical order with increasing depth level from the root to up to 15 levels of specification.

With this unambiguous vocabulary at hand, it is possible to explore the relationships of genes based on their ontological annotation provided in terms of their biological attributes. Thus, it is not a surprise that more and more applications based on GO-database search have appeared in recent years — for example, the very recent publications by Cai *et al* [5] and by Steuer *et al* [18]. In [5], an approach is proposed to finding reliable differences between the genes in two genomes based on all GO

levels. In [18], the information-theoretic concept of *mutual information* is suggested as a tool to investigate the relationship between gene clusters and their respective functional categories in GO data.

Here, we will use GO data to construct gene networks by connecting any two genes from a given collection of genes under consideration by an edge if and only if they share — according to their *GO annotation* — at least one gene attribute. To identify putative gene-function from the resulting networks, we will then apply the above-mentioned methods for community detection.

On the basis of our results, we will then argue not only that this approach appears to yield reasonable proposals for gene functional modules on the basis of GO data, but also that, if the enormous reduction of complexity achieved by replacing the full content supplied by GO annotation by the simple and purely formal "yes-no" data regarding the existence of shared gene attributes still allows the identification of putative gene functional modules, community-detection methods may also be useful for constructing (or checking) GO annotations on the basis of networks constructed from much less sophisticated and detailed primary data.

2 The Construction of GO Networks

To construct the networks whose community structures we want to detect, we consider the Zebra-fish genome using the publicly available Ensembl database Zv6 Release 41 to extract the genomic data we need, including Gene Ensembl IDs, and GO attributes (Table 1). We find that the percentage of genes with GO annotation varies among the different chromosomes from 21% to 52% (Table 2). The only chromosome with a ratio above 50% is Chromosome 20. Thus, we will focus on this chromosome here.

There are several methods to measure the relationship between genes using GO data, like *semantic analysis* [10,15] and *pathway covering* [14]. As the GO data are organized in terms of a directed acyclic graph, we adopt a rather simple and straightforward method to construct our gene networks: We associate, to each gene in question, the set of all of its GO attributes considered as subset of the set of all possible (currently $21,909$) attributes and construct the corresponding *subset-intersection* graph that is defined as follows:

Given a finite set X and a finite family $\mathscr{X}$ of subsets X_i $(i = 1,2,...,n)$ of X, we let

$$G_{\mathscr{X}} := (\{1,2,...,n\}, E_{\mathscr{X}})$$

denote the graph with vertex set $\{1,2,...,n\}$ and edge set

$$E_{\mathscr{X}} := \left\{\{i,j\} \in \binom{\{1,2,...,n\}}{2} \mid X_i \cap X_j \neq \emptyset\right\}.$$

Using GO data, we let X denote the set of all GO attributes identified in terms of their respective ID and, for each gene $i = 1,2,...,n$, we let X_i denote the set of all attributes of gene i.

Table 1: The example of Genes and GO IDs.

Gene ID	Gene Ensembl ID	GO ID
2	ENSDARG00000059215	GO:0005922 GO:0016021 ...
5	ENSDARG00000043963	GO:0005515
12	ENSDARG00000059187	GO:0005922 GO:0016021 ...
14	ENSDARG00000059183	GO:0005874 GO:0043234 ...
16	ENSDARG00000059168	GO:0016021 GO:0016020 ...
17	ENSDARG00000059164	GO:0016021 GO:0016020 ...
19	ENSDARG00000043982	GO:0005874 GO:0043234 ...
20	ENSDARG00000020785	GO:0005102 GO:0030903 ...
21	ENSDARG00000043973	GO:0005634 GO:0046872 ...
26	ENSDARG00000044013	GO:0005634 GO:0003677 ...
29	ENSDARG00000044016	GO:0016459 GO:0005524 ...
31	ENSDARG00000059041	GO:0000166 GO:0046872 ...
32	ENSDARG00000018477	GO:0000166 GO:0046872 ...
33	ENSDARG00000006332	GO:0007242 GO:0035091 ...
34	ENSDARG00000009020	GO:0004930 GO:0001584 ...
35	ENSDARG00000032261	GO:0004484 GO:0003676 ...
37	ENSDARG00000043902	GO:0016021 GO:0045211 ...
38	ENSDARG00000014057	GO:0016021 GO:0045211 ...
39	ENSDARG00000033489	GO:0006512 GO:0006464 ...
42	ENSDARG00000003751	GO:0016740 GO:0004674 ...
43	ENSDARG00000019747	GO:0006694 GO:0004769 ...
44	ENSDARG00000034076	GO:0004872
45	ENSDARG00000015201	GO:0006464 GO:0016740 ...
48	ENSDARG00000021509	GO:0008270 GO:0046872 ...
49	ENSDARG00000010425	GO:0005737 GO:0016020 ...
57	ENSDARG00000043858	GO:0005524 GO:0004672 ...
59	ENSDARG00000043854	GO:0006457 GO:0003676 ...
60	ENSDARG00000005416	GO:0005524 GO:0000166 ...
...	...	...

Although the GO graph represents a hierarchical structure, the resulting subset-intersection graph appears to be still informative enough because, down from the root, all the GO items in our genomic data are below the fourth level of the whole graph which implies that distinct genes have distinct sets of gene attributes.

We first generate the network of Chromosome 20 (Figure 1). It contains 645 vertices indexed by the gene IDs, and $33,459$ edges resulting from the subset-intersection graph construction. There are two components in the graph. In the bigger component, we can discover several communities even by eye that consist of IDs from a small range. This implies that quite a few genes gather closely together and form easily identifiable subgroups that are involved in similar biological processes, exhibit closely related molecular functions, or occur in the same cell components. To investigate the network in greater detail, we restrict our attention to the first 28 genes on Chromosome 20 with GO annotation. These are located quite closely to each other. The associated small gene network is shown in Figure 2.

3 Detecting Communities in the Gene Network

As mentioned already above, one network feature that has been emphasized recently is that they may give rise to easily detectable community structures. According

Table 2: Number of the Genes with GO annotations in Zebra Fish chromosomes.

Chr	Number of the genes with GO annotation	Total gene number	Ratio
1	331	1019	0.324828
2	353	997	0.354062
3	437	1234	0.354133
4	331	1191	0.277918
5	390	1414	0.275813
6	338	958	0.352818
7	437	1231	0.354996
8	331	1010	0.327723
9	282	777	0.362934
10	315	915	0.344262
11	261	754	0.346154
12	272	840	0.32381
13	321	890	0.360674
14	341	1624	0.209975
15	278	822	0.3382
16	296	882	0.335601
17	287	891	0.32211
18	219	699	0.313305
19	264	706	0.373938
20	**668**	**1277**	**0.523101**
21	330	870	0.37931
22	207	975	0.212308
23	289	892	0.323991
24	198	588	0.336735
25	208	643	0.323484

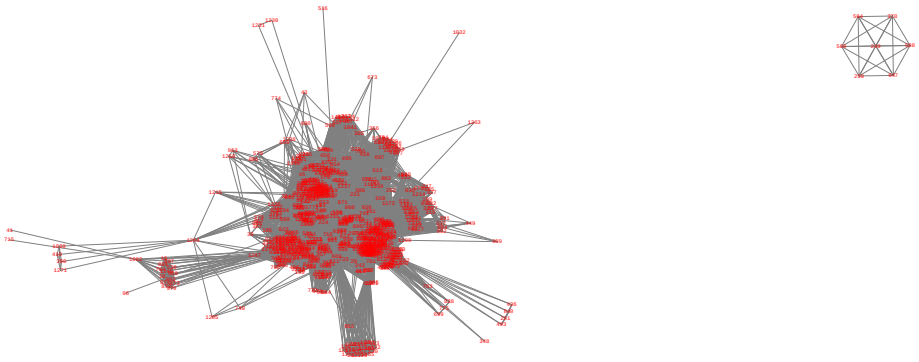

Figure 1: The gene network of Zebrafish chromosome 20 defined by the O data. The number of every node is the ID in Table 1.

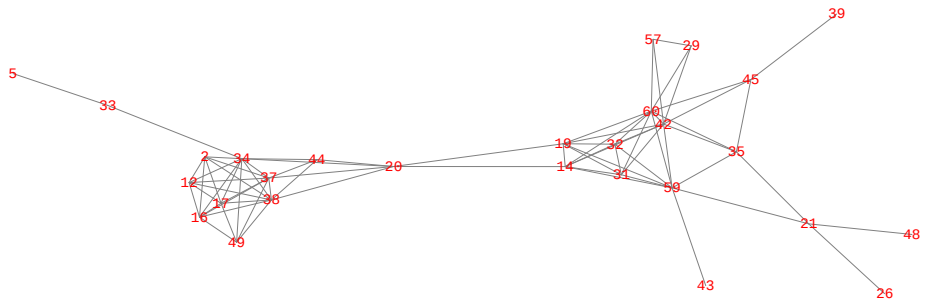

Figure 2: The network of the first 30 genes that have GO nnotations on the chromosome 20.

to [8, 11, 17] a community structure of a network is a partition of the network's vertex set into disjoint groups such that there is a higher density of edges within rather than across them.

In recent years, a lot of effort has been spent on developing algorithms for detecting communities in networks.

For example, M.Newman and M.Girwan defined a quantitative measure called "modularity" in [16] to measure the *goodness of fit* of an arbitrary community structure relative to a given network, and Clauset *et al* [8] proposed an amazingly fast greedy approach dubbed "FastCommunity" to find community structures with comparatively high modularity in any given network.

A very different Linear-Programming (LP) based approach to community-structure detection was recently proposed in [6, 7]. For the reader's convenience, we briefly introduce that approach: Define a graph $T = (V,F)$ with vertex set V and edge set $F \subset \binom{V}{2}$ to be a community graph if it is a disjoint union of cliques. Clearly, such community graphs are in a canonical one-to-one correspondence with the community structures defined on V that we want to detect.

Moreover, a graph $T = (V,F)$ is a community graph if and only if one has

$$\chi_T(uv) + \chi_T(vw) - \chi_T(uw) \leqslant 1,$$

for any three distinct elemants $u,v,w \in V$, for the associated indicator function:

$$\chi_T : \binom{V}{2} \mapsto \{0,1\} : \{u,v\} \mapsto \chi_T(uv) := \begin{cases} 1 & \text{if } u,v \in F, \\ 0 & \text{else.} \end{cases}$$

Consequently, given network $G = (V,E)$, to find — by modifying (via edge deletion and insertion) the network G — a community graph T that optimally approximates G, one may use Linear Programming to determine, within the set of *feasible maps* $\chi : \binom{V}{2} \to \mathbb{R}$, i.e., the set of maps $\chi : \binom{V}{2} \to [0,1]$ for which the above *constraints* $\chi_T(uv) + \chi_T(vw) - \chi_T(uw) \leqslant 1$ are satisfied for any three distinct elements u,v,w in

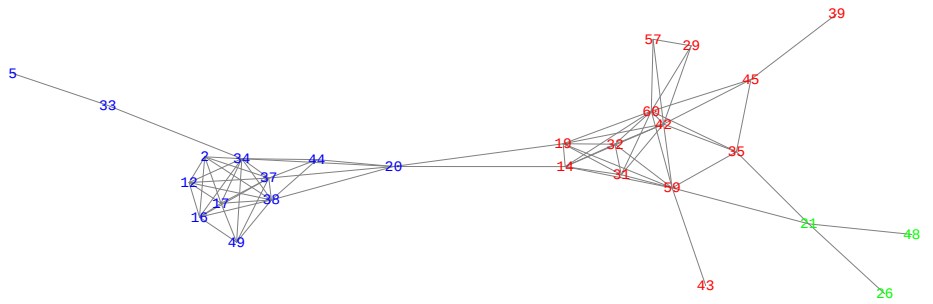

Figure 3: Communities detected by FastCommunity. The communities 1, 2, and 3 are in red, blue and green.

V, that map $\chi = \chi^*_G$ that maximizes a *penalty function* $L = L_{E,s}$ of the form

$$L_{E,s} : \mathbb{R}^{\binom{V}{2}} \to \mathbb{R} : \chi \mapsto \mathbf{s} \sum_{\{u,v\} \in E} \chi_T(uv) - \sum_{\{u,v\} \in \binom{V}{2} - E} \chi_T(uv)$$

where the *control parameter* **s** is used for automatically calibrating, for our family of penalty functions, the specific penalty one has to pay for deleting one single edge from the edge set E.

To do so, an amazingly simple and effective strategy was proposed in [6, 7]: Starting with $s := 1$, one increases s step by step until, at some value $s = s^*$, the relaxed LP problem has, for the first time, an integer-valued solution $\chi_{E,s^*} : \binom{V}{2} \mapsto \{0,1\}$, thus necessarily corresponding to a community graph.

4 Results

We applied both, the FastCommunity algorithm and the LP-based approach, to our GO network (cf. Figure 2). Even though the first method yields three (cf. Figure 3) and the second one four communities (cf. Figure 4), the partitions detected by the two approaches are — except for how they treat the nodes 5 and 33 — exactly the same. According to GO annotation, the gene products of Node 5 and 33 are involved in amino acid binding, in other words, they interact with other protein or protein complex, while the product of Node 33 is also involved in cell adhesion which is related to signal transduction. The LP-based approach regards these two nodes are as one community while FastCommunity regards them as members of Community 2.

Thus, we investigated the detailed GO annotation of every gene and found that both methods give reasonable partitions: The nodes 21, 35 and 59 in Community 1 are related to nucleic-acid binding, 14, 19, 29, 31, 32, 42, 57, 59, and 60 are related to nucleotide binding, Node 45 is involved in transferase activity. Except for node 45, all genes in Community 1 are related to either nucleic-acid binding or nucleotide binding. Thus, this community forms a *module* of molecules that selectively — often even stoichiometrically — interact with one or more specific sites on another molecule. Sometimes, the products of these genes can also be considered as ligands.

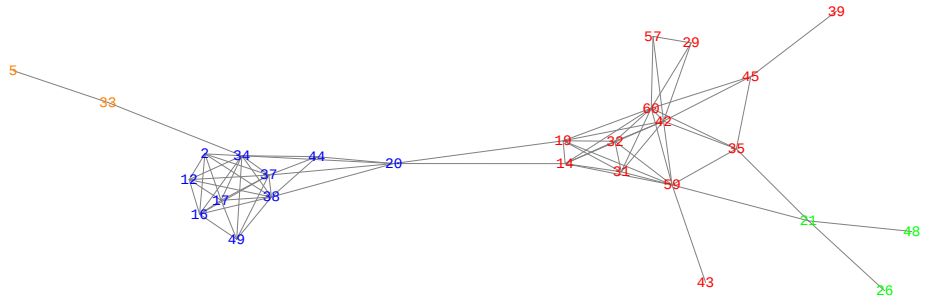

Figure 4: Communities detected by Linear-Programming based approach. The communities 1, 2, 3, and 4 are in red, blue, green and orange.

Table 3: Communities detected by Linear-Programming based approach.

	Gene ID.
Community 1	14,19,29,31,32,35,39,42,43,45,57,59,60
Community 2	2,12,16,17,20,34,37,38,44,49
Community 3	21,26,48
Community 4	5,33

The nodes $2, 12, 16, 17, 34, 37$, and 38 in community 2 are components of membranes that are single or double, lipid- and protein-associated layers that enclose cells. The nodes $20, 34, 37, 38$ and 44 are involved in receptor activities. According to the definitions in GO, receptor activity means that they combine with an extracellular or intracellular messenger to initiate a change in cell activity. Thus, we can infer that the products of all the genes in Community 2 play an important role in transducting signals between the outside and inside of cells.

The genes in Community 3 are a bit more complicated (even though they were detected by both methods). The products of the nodes 21 and 26 are proteins predominantly found in the nucleus. Nodes 21 and 48 are related to zinc-ion binding and Node 21 is also related to nucleic-acid binding. However, zinc-ion binding proteins are always involved in DNA activities, e.g., transcription-activator proteins. Thus, the genes of this community appears to be involved in nucleic-acid binding activities in the nucleus, and it can be considered as a sub-community of the community 1.

As described above, Community 4 (Nodes 5 and 33) is related to signal transduction and cell adhesion, so, it can be considered as a sub-community of the community 2. That's why it also makes sense for the FastCommunity algorithm to integrate the two nodes into Community 2.

5 Conclusions

As every other scientific enterprise in the natural or social sciences or the humanities, Mathematics deals with specific aspects of reality.

Table 4: Communities detected by FastCommunity.

	Gene ID.
Community 1	14,19,29,31,32,35,39,42,43,45,57,59,60
Community 2	2,5,12,16,17,20,33,34,37,38,44,49
Community 3	21,26,48

More exactly, it deals with certain rather formal aspects of reality, that is, not with what an object under investigation really *is*, i.e., with objects per se, but with the abstract form of relationships that may persist between various objects under investigation (being larger or smaller, similar or dissimilar, influencing one another in one or another way, ...).

Its ability to identify and conceptualize relevant abstract forms of relationships is a great strength of Mathematics, and simultaneously a limiting factor: As such, it will never deal with “real" content, but only with constructs based on purely formal relationships identified by utter abstraction.

The current work regarding “Networks" is a case in point. Its stunning success is based exclusively on the simple, yet rather surprising observation that, just recording in terms of a simple undirected graph whether or not two objects u and v in given collection V of objects are in some way related to each other or not and otherwise completely ignoring the very nature of these objects and their relationship can reveal important insights into the structure of the collection of objects under investigation, independently of whether one investigates the *World-Wide Web*, scientific-collaboration or citation networks, or ecological, genetic, regulatory, protein-protein interaction or metabolic networks.

The universally acclaimed proclamation of “scale-free" and “small-world" networks as constituting important new and universally applicable paradigms of interaction schemes observed in real-world systems, suggesting fundamentally new basic laws governing important processes addressed in the natural and the social sciences clearly underlines this fact.

In this note, we report on recent work regarding the usage of *community-structure detection* in networks for identifying putative gene functional modules. We construct a network comprising 28 genes from the Zebra-Fish genome and employ the community-detection method to identify biological meaningful communities within this network. GO-analysis allows to give reasonable explanations of the communities that were found based on common biological attributes of the genes in one community and elucidates the relations between different communities. This indicates that the idea of constructing gene networks using GO data, and detecting communities in the network using Linear-Programming and FastCommunity can facilitate comprehending the genes' relationships and discovering gene functional modules.

6 Acknowledgments

The authors would like to thank Chaofeng Wang, Zhongshan Li for their help on the work. This work was granted financial support from China Postdoctoral Science Foundation, was supported by the 973 Project on Mathematical Mechanization, the Ministry of Education, the Ministry of Science and Technology, the National Science Foundation of China, and the Max Planck Society. The work was done partially while the author was visiting the Institute for Mathematical Sciences, National University of Singapore in 2006. The visit was supported by the Institute.

References

[1] L A Adamic and B A Huberman, Power-law distribution of the world wide web, Science 287 (2000), 2115.

[2] L A N Amaral, A Scala, M Barthelemy, and H E Stanley, Classes of small world networks, Proc. Natl. Acad. Sci. USA 97 (2000), 11149 – 52.

[3] A L Barabasi and R Albert, Emergence of scaling in random networks, Science 286 (1999), 509 – 12.

[4] J Bascompte, C Melian, and E Sala, Interaction strength combinations and the overfishing of a marine food web, Proc. Natl. Acad. Sci. USA. 102 (2005), 5443 – 7.

[5] Z Cai, X Mao, S Li, and L Wei, Genome comparison using gene ontology (go) with statistical testing, BMC Bioinformatics. 7 (2006), 374.

[6] Y C Chen, A W D Dress, and Q Yu, Community structures of networks, Internatinal Conference on Mathematical Aspects of Computer and Information Sciences, Beijing, China, July 24-26, 2006.

[7] Y C Chen, AW D Dress, and Q Yu, Checking the reliability of a new approach towards detecting community structures in networks using linear programming, IET Systems Biology (2006), submited.

[8] A Clauset, J Newman, and C Moore, Finding community structure in very large networks, Phys. Rev. E 70 (2004), 066111.

[9] The Gene Ontology Consortium, Gene ontology: tool for the unification of biology, Nature Genet. 25 (2000), 25 – 9.

[10] F Couto, M Silva, and P Coutinho, Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors, Conference in Information and Knowledge Management, ACM CIKM, Oct, 2005.

[11] M Girvan and M E J Newman, Community structure in social and biological networks., Proc. Natl. Acad. Sci. USA 99 (2002), 7821 – 26.

[12] H Jeong, B Tombor, R Albert, Z N Oltvai, and A L. Barabasi, The large-scale organization of metabolic networks, Nature 5 (2000), no. 407, 651 – 4.

[13] P L Krapivsky, S Render, and F Leyvraz, Connectivity of growing random networks, Phys. Rev. Lett. 85 (2000), 4629 – 32.

[14] R Li, S Cao, Y Li, H Tan, Y Zhu, Y Zhong, and Y Li, A measure of semantic similarity between gene ontology terms based on semantic pathway covering, Progress in Natural Science 16 (2006), no. 7, 721 – 26.

[15] P W Lord, R Stevens, A Brass, and C A Goble, Semantic similarity measures as tools for exploring the gene ontology, Pacific Smposium on Biocomputing, Oct, 2003.

[16] M E J Newman and M Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69 (2004), 026113.

[17] F Radicchi, C Castellano, F Cecconi, V Loreto, and D Parisi, Defining and identifying communities in networks, Proc. Natl. Acad. Sci. USA 101 (2004), 2658 – 63.

[18] R Steuer, P Humburg, and J Selbig, Validation and functional annotation of expression-based clusters based on gene ontology, BMC Bioinformatics. 7 (2006), 380.

[19] A Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, Mol. Biol. Evol. 18 (2001), 1283 – 92.

[20] D Watts, Small worlds, Princeton University Press, Princeton, 1999.