Understanding the Global Properties of Functionally-Related Gene Networks Using the Gene Ontology

L. Tari, C. Baral, and P. Dasgupta

Pacific Symposium on Biocomputing 10:209-220(2005)

UNDERSTANDING THE GLOBAL PROPERTIES OF FUNCTIONALLY-RELATED GENE NETWORKS USING THE GENE ONTOLOGY

L. TARI, C. BARAL AND P. DASGUPTA
Dept. of Computer Science and Engineering, Arizona State University,
Brickyard Suite 501, 699 South Mill Avenue
Tempe, AZ 85287,USA
Email: {luis.tari,chitta,partha}@asu.edu

The global behavior of interactions between genes can be investigated by forming the network of functionally-related genes using the annotations based on the Gene Ontology. We define two genes to be connected when the pair of genes is involved in the same biological process. There has been other work on the analysis of different kinds of cellular and metabolic networks, such as gene coexpression network, in which genes are paired when they are found to be coexpressed in the microarray experiments. We observe that our functionally-related gene networks among humans, fruit flies, worms and yeast exhibit the small-world property, but all except the network of worms show the existence of the scale-free property.

1. Introduction

Uncovering the underlying functions and interactions within a living cell is an important goal in the post-genomic era. Recent advances in technology, such as the development of microarrays and protein chips, allow biologists to study the functioning of the cell in many new ways. While it significantly speeds up the process of understanding bio-molecular interactions, modeling interactions of a cell in quantifiable terms is a major challenge for biologists. By modeling the interactions, the ultimate goal is to discover the fundamental properties that govern the behavior of a cell.

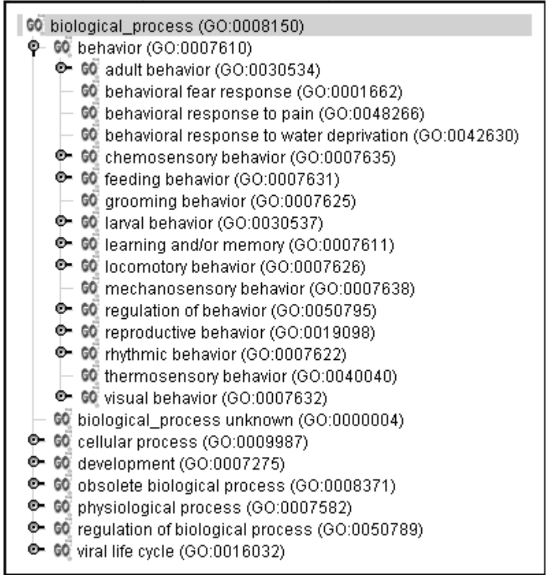
Work by Watts and Strogatz on the small world characteristic[1] and by Barabasi and Albert on the scale-free feature[2] of large, sparse and complex networks has been applied to various areas such as sociology and computer networks. The small-world phenomenon demonstrates the famous property of six degrees of separation between any two persons in the world[3]. The neural network of the worm *Caenorhabditis elegans*, the power grid of the western United States and the collaboration of film actors have been shown to exhibit small-world properties[4]. Some of the important outcomes of such small-world networks are the increase of signal-propagation speed, computational power and synchronizability. In biological domains, infectious diseases are found to be more easily spread in small-world networks than in regular networks[4].

Previously, complex networks have been thought to exhibit the property of classical random networks, in which the fundamental randomness of the model leads to the same number of edges in most nodes. Empirical studies on the structure of the World Wide Web[5] show that a few highly connected nodes dominate the structure. This phenomenon is known as scale-free. Scale-free networks are robust with respect to random attacks and component failures, since the chance of harming a highly connected node is low[6]. This is in contrast to a random network, in which the removal of several nodes can effectively disrupt the network. On the other hand, attacks on highly connected nodes in a scale-free network can catastrophically disrupt the network.

Most cellular functions are known to be carried out by groups of molecules within interacting functional modules[7]. It is essential to study interactions within a cell and the properties that govern the interactions. Recently, there has been a significant amount of interest in examining the universal property that oversees the complex molecular interactions between the cell components, by modeling the molecular interactions in the form of networks, equivalently known as graphs in mathematical terms. Numerous studies of various types of cellular and metabolic networks have shown the existence of small-world and scale-free properties. Some of these studies include the characterizations of physical interactions between protein-protein, protein-nucleic-acid, protein-metabolite molecule pairs[8,9,10]. Modeling of more complex functional interactions such as metabolites that are substrates or products in the same biochemical reaction[11,12], or chemical reactions that share at least one chemical component either as substrate or as product, also reveal such properties[13]. Further examples of small-world, scale-free organization includes the study of genetic-regulatory network such as protein domain interactions and coexpression of genes based on microarray data[14,15], in which coexpression of genes are paired to imply that the genes are involved in the same biological process.

The main focus of this work is on the characterization of the gene involvement in the same biological process in a large scale. Unlike the previous work[14,15] in which the study of the involvement of genes in the same biological process were based on coexpression of genes in microarray experiments, our work utilizes the Gene Ontology to study the global behavior of such networks.

The Gene Ontology is a hierarchy of controlled vocabulary that includes three independent ontologies for biological process, molecular function and cellular component. Standardized terms in the Gene Ontology describe roles of genes and gene products in any organism. Figure 1 illustrates the main terms in the biological process ontology. A gene product has one or more molecular functions, can be used in one or more biological processes, and can be associated with one or more cellular components[16]. As a way to share knowledge about functionalities of genes, the Gene Ontology itself does not contain gene products of any organisms. Rather, biologists annotate biological roles of gene products using the Gene Ontology, known as annotations.



- biological_process (GO:0008150)
  - behavior (GO:0007610)
    - adult behavior (GO:0030534)
    - behavioral fear response (GO:0001662)
    - behavioral response to pain (GO:0048266)
    - behavioral response to water deprivation (GO:0042630)
    - chemosensory behavior (GO:0007635)
    - feeding behavior (GO:0007631)
    - grooming behavior (GO:0007625)
    - larval behavior (GO:0030537)
    - learning and/or memory (GO:0007611)
    - locomotory behavior (GO:0007626)
    - mechanosensory behavior (GO:0007638)
    - regulation of behavior (GO:0050795)
    - reproductive behavior (GO:0019098)
    - rhythmic behavior (GO:0007622)
    - thermosensory behavior (GO:0040040)
    - visual behavior (GO:0007632)
  - biological_process unknown (GO:0000004)
  - cellular process (GO:0009987)
  - development (GO:0007275)
  - obsolete biological process (GO:0008371)
  - physiological process (GO:0007582)
  - regulation of biological process (GO:0050789)
  - viral life cycle (GO:0016032)

Figure 1 The hierarchy of main terms in the biological process ontology, with the term "behavior" expanded to show its children. Screenshot was taken from the TGen GOBrowser (to be published.)

Our model is a network composed of genes or proteins as nodes and an edge exists between two nodes if they are involved in the same biological process[a]. A biological process is defined as a biological objective to which the gene or gene product contributes[16]. We selected four evolutionarily conserved organisms: *Homo sapiens*, *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* to construct functionally-related gene networks from the Gene-Ontology-based annotations and study the global behavior that governs such network.

The remainder of the work is organized as follows. Section 2 provides the descriptions and definitions of the formal measures of the small-world and scale-free properties. The process of constructing our functionally-related gene networks is described in Section 3. Section 4 describes the existence of small-world and scale-free properties in our networks, and section 5 discusses some of the possible implications of such properties in the networks.

[a] Currently the generality or specificity of the Gene Ontology terms is not taken into consideration.

2. Preliminaries

The small-world property can be characterized by two statistical quantities: clustering coefficient C and characteristic path length L[1]. L is the average minimal distance between any two nodes in the network, while C is a measure of how clustered a graph is, which implies an average of interconnectivity among the neighbors of each node. More formally,

$$C = \frac{\sum_{i=1}^{n} C_i}{n} \text{ and } C_i = \frac{e_i}{k_i \times (k_i - 1)/2} \tag{1}$$

if node i has k_i immediate neighbors with e_i number of edges between i's neighbors in a graph of n nodes. It is easy to see that a fully connected graph has a clustering coefficient of 1. If a given node j has no neighbors or one neighbor, then we define $C_j = 1$. Regular networks have large C and L grows linearly with n, while random networks have small C and L only grows logarithmically with n[17]. In other words, regular networks have relatively large L, while random networks have relatively small L. By having the same configuration (i.e. the same number of nodes and edges) but with different probability p of rewiring edges, a collection of graphs between a regular network (p=0) and a random network (p=1) can be generated. Such random rewiring procedure shows that for intermediate values of p, the graph is a small-world network[4]. This phenomenon implies that small-world networks fall in between the two; small-world networks are highly clustered like regular networks, while the characteristic path length is as small as random networks[4]. With the relation among the three kinds of networks, showing a network exhibit the small-world property requires the comparison of the actual configuration of the network with the random configuration of itself. For random networks, the two quantities can be computed as

$$C \approx \bar{k}/(N-1),\, L \approx \ln N / \ln \bar{k} \tag{2}$$

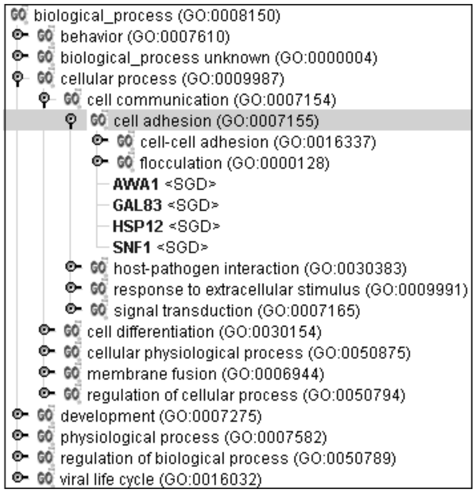
where N is the number of nodes in a network and $\bar{k}$ is the average number of edges per node[18].

The scale-free property is defined by an algebraic behavior in the probability of degree distribution $P(k)$, i.e. the probability that a selected node has exactly k edges. Scale-free networks are networks that have a degree distribution approximated as the power law, $P(k) \sim k^{-\gamma}$, where k is the number of edges and γ is the degree exponent[10]. The existence of the scale-free property in a network implies that there can be a few nodes with a significantly larger number of edges than the typical nodes.

3. Construction of functionally-related gene networks

The Gene Ontology is composed of three independent ontologies: molecular function, biological process and cellular component. Our functionally-related gene networks are constructed from annotations based on the biological process ontology of the Gene Ontology. The networks are composed of genes or proteins as nodes and two nodes are connected if they are involved in the same biological process based on the annotations. Figure 2 illustrates the idea.

The annotations that we used to construct the networks are curated by various highly recognized organizations and institutes. The annotation of *Homo sapiens* is obtained from the collection at the European Bioinformatics Institute, while the annotation of *Drosophila melanogaster* is from the FlyBase organization. The annotation for *Caenorhabditis elegans* is obtained from the WormBase organization, while the annotation for *Saccharomyces cerevisiae* is from the collection at Stanford University. In all four cases, the actual files used were the annotation files lodged with the Gene Ontology Consortium by the four currating organizations.



- biological_process (GO:0008150)
  - behavior (GO:0007610)
  - biological_process unknown (GO:0000004)
  - cellular process (GO:0009987)
    - cell communication (GO:0007154)
      - cell adhesion (GO:0007155)
        - cell-cell adhesion (GO:0016337)
        - flocculation (GO:0000128)
        - **AWA1** <SGD>
        - **GAL83** <SGD>
        - **HSP12** <SGD>
        - **SNF1** <SGD>
      - host-pathogen interaction (GO:0030383)
      - response to extracellular stimulus (GO:0009991)
      - signal transduction (GO:0007165)
    - cell differentiation (GO:0030154)
    - cellular physiological process (GO:0050875)
    - membrane fusion (GO:0006944)
    - regulation of cellular process (GO:0050794)
  - development (GO:0007275)
  - physiological process (GO:0007582)
  - regulation of biological process (GO:0050789)
  - viral life cycle (GO:0016032)

Figure 2 The genes products for yeast genes AWA1, GAL83, HSP12, SNF1 are all annotated as being involved in the biological process of cell adhesion. Pairs of the four genes are linked in the functionally-related gene network for yeast. Screenshot was taken from the TGen GOBrowser (to be published.)

Annotations of the four organisms were first preprocessed to remove genes that are mapped to unknown and obsolete Gene Ontology terms. Unknowns, referred to as the term "biological process unknown" in the biological process ontology, are used when annotation of gene products whose function are not known or cannot be inferred. Obsolete Gene Ontology terms are terms that have been removed from the active biological process ontology[19]. The Gene Ontology

Consortium defines a set of evidence codes to support the functional assignments of gene products. As part of the annotation process, curators are required to provide an evidence code when assigning a Gene Ontology term to a gene product. Reliability of annotations varies with different evidences. To further increase the reliability of our network, gene product annotations that are inferred from electronic annotation (IEA) were removed from our network. The evidence IEA is used when no curator has checked the annotation to verify its accuracy, and thus has the lowest quality among all evidences[20]. The resulting networks are composed of 7512 proteins for humans, 4641 genes for fruit flies, 3254 proteins for worms and 4660 genes for yeast.

4. Results

Our results show that the functionally-related gene networks of the four organisms exhibit the small-world property. All networks except the network for worms also demonstrate the scale-free property.

We first present the results regarding the existence of small-world property of the networks. As shown in Table 1 (referred to as the actual configuration of the networks), the clustering coefficients *C*, computed by equation 1, of the four networks are very high, while the characteristic path lengths *L* are surprisingly quite small. The low characteristic path lengths in the actual configuration are related to a high number of edges for each node. In particular, the human functionally-related gene network has an average number of 276.68 edges for each node and on average a node can be reached by another node within 2.58 links. To examine the existence of small-world property of the networks, the clustering coefficients *C* and characteristic path lengths *L* of the random configuration of the networks with the same parameters N and $\bar{k}$ were approximated by equation 2[18], as shown in Table 2. The results show that the networks with actual configuration have much higher clustering coefficients *C*, while the characteristic path lengths *L* are about the same as the random configurations. Table 3 describes the minimal paths between any two nodes among the four networks in the actual configuration, showing that in the worst case there can be 8 degrees of separation between two nodes in the human network, but with a very low probability of 6.38×10^{-7}. These results confirm the networks are highly clustered but with short characteristic path lengths. In other words, the functionally-related gene networks of the four organisms are highly clustered and at the same time have small path lengths, which coincide with the property of small-world network.

As for the scale-free property, our results shown in figure 3 illustrates convincingly that the functionally-related gene network of humans, fruit flies and yeast follow a power-law distribution. Figures 4, 5, 6, 7 show the exact degree distribution for each of the organism. However, in the case of worms, it does not seem to follow a power-law distribution. The property of the power-law distribution shows that the functionally-related gene network of humans, fruit flies and yeast can be modeled by scale-free networks, while we cannot make the same observation for worms.

Table 1 Results for the functionally-related gene network constructed from the Gene-Ontology-based annotations (actual configuration). *N* is the total number of nodes (genes or proteins), $\bar{k}$ is the average number of edges per node, *C* is the clustering coefficient and *L* is the average shortest path.

	N	$\bar{k}$	C	L
Humans	7512	276.68	0.87	2.58
Fruit Flies	4641	100.22	0.87	2.90
Worms	3254	1573.85	0.87	1.55
Yeast	4660	73.44	0.88	3.51

Table 2 Results for the functionally-related gene networks (random configuration) with the same parameters *N* and $\bar{k}$ as in Table 1.

	N	$\bar{k}$	C	L
Humans	7512	276.68	0.037	1.59
Fruit Flies	4641	100.22	0.022	1.83
Worms	3254	1573.85	0.48	1.10
Yeast	4660	73.44	0.016	1.97

Table 3 The length distribution of the minimal paths between two nodes of length L_n among the four organisms. No path exists between two nodes if $n = 0$.

n	L_n (Humans)	L_n (Fruit Flies)	L_n (Yeast)	L_n (Worms)
0	0.0834	0.1840	0.2004	0.0129
1	0.0368	0.0216	0.0158	0.4838
2	0.3958	0.2140	0.0929	0.4698
3	0.4100	0.4297	0.3032	0.0321
4	0.0676	0.1295	0.2721	0.0013
5	0.0058	0.0186	0.0923	1.927×10^{-5}
6	5.24×10^{-4}	0.0022	0.0203	-
7	2.3×10^{-5}	0.00036	0.0028	-
8	6.38×10^{-7}	2.158×10^{-5}	3.002×10^{-4}	-
9	-	5.57×10^{-7}	2.34×10^{-5}	-
10	-	-	1.29×10^{-6}	-

Table 4 Reachability of the network. *U* is the percentage of unconnected pairs in the networks, *L* is the average shortest path and *R* is the percentage of node pairs that can be reached within $\lceil L \rceil$ (ceiling of *L*) number of links.

	U	*L*	*R*
Humans	8.34%	2.58	91.93%
Fruit Flies	5.1%	2.90	94.51%
Worms	1.29%	1.55	96.61%
Yeast	20.03%	3.51	85.53%

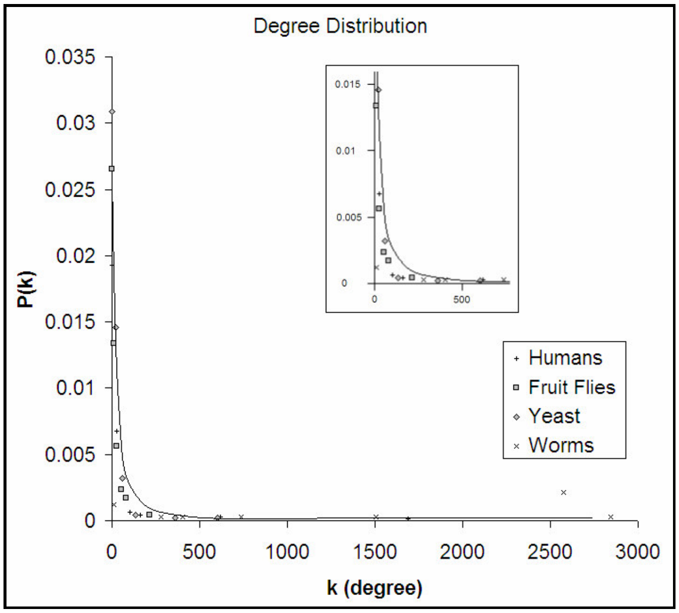


Figure 3 - Alegbraic scaling behavior of *P*(*k*) for humans, fruit flies and yeast, but not worms. *P*(*k*) is the probability that a selected node has *k* number of edges. The inset shows a clearer view of the curve.

We also examine the reachability of the four networks. Table 4 shows that all four networks have surprisingly small path lengths. This behavior can be explained with the presence of highly connected nodes in the networks, as illustrated in figure 3. In fact, less than 8% of the proteins in the human network can be reached by more than 3 proteins from any given protein. Similarly, the other 3 networks also exhibit such close connectivity between any two nodes. On the other hand, we also observe that there are nodes that cannot be reached by another given node in the networks. Specifically, out of all possible pairings in the human network, 8.34% of the pairs cannot be reached by each other. Among all of the four networks, the yeast network has the most number of unconnected pairs – about 20% of all pairs. The existence of unconnected pairs can be explained by the fact that not all functions of genes for each of the four organisms have been fully discovered.

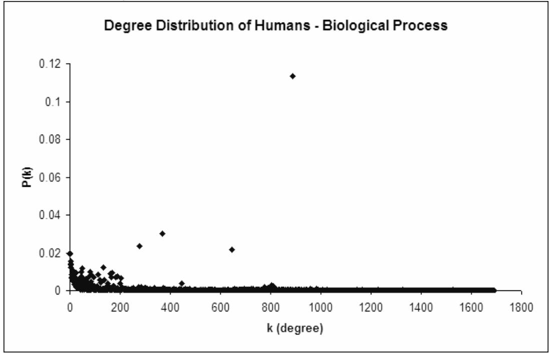


Figure 4 The exact degree distribution of the functionally-related gene network of Humans, where *P*(*k*) refers to the probability that a selected node has *k* edges.

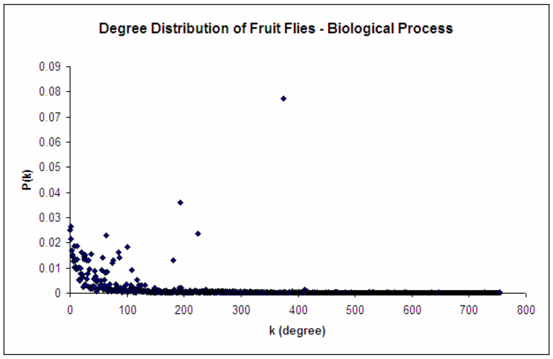


Figure 5 The exact degree distribution of the functionally-related gene network of Fruit Flies, where *P*(*k*) refers to the probability that a selected node has *k* edges.

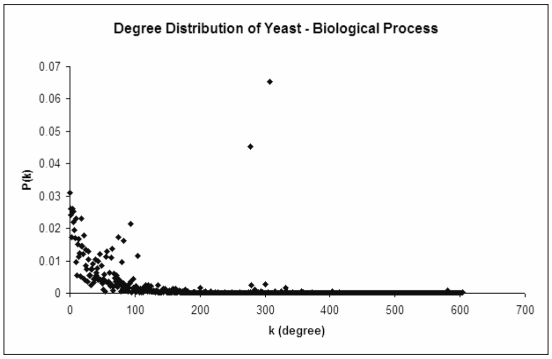


Figure 6 The exact degree distribution of the functionally-related gene network of Yeast, where *P*(*k*) refers to the probability that a selected node has *k* edges.

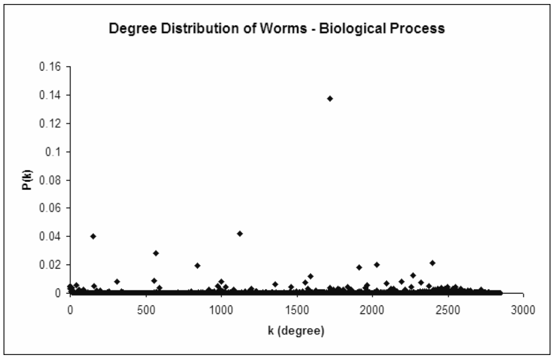


Figure 7 The exact degree distribution of the functionally-related gene network of Worms, where *P*(*k*) refers to the probability that a selected node has *k* edges.

5. Discussion

There have been various studies on the global property that governs the behavior of different aspects of metabolic and cellular networks. Our work differs from the others as we focus on a different perspective of biological network, specifically gene products that are involved in the same biological process. Our work is most closely related to the study of the gene coexpression network, in which coexpressed genes in the microarray experiments are connected to form the network. Coexpressed genes may imply that the genes are functionally related, i.e. genes are involved in the same biological process. However, coexpression of genes depends on the threshold of coexpression correlation and thus has an effect of the size and connectivity of such network. The implication of coexpression of genes to be functionally related genes can arguably be an assumption. Our use of the Gene-Ontology based annotations is independent of such assumption. In addition, our method does not have dependence on threshold values and experimental bias in the microarray data. It is inevitable that our method also introduces some potential bias by utilizing the annotations to construct our networks. Annotations are curated based on different evidences such as direct assay, sequence similarity and expression pattern, in which each has its own experimental bias. However such bias should be restrained to the minimum, as there are strict guidelines on the approval of the annotations, and the Gene-Ontology based annotations are widely accepted by the biomedical community. Among all of the evidence codes used for the annotations, the evidence code "inferred from electronic annotation" is applied to annotations that are yet to be verified for their accuracy by the curators. Such annotations are removed to ensure high quality and reliability of our networks.

Our work also goes in line with the common application of the Gene Ontology – interpreting microarray data from a biological point of view[21,22]. Microarray experiments allow biologists to find a set of differentially expressed genes between two or more conditions being studied, for instance among tissues treated with drugs and untreated tissues. Identifying which genes are differentially expressed is important, but it is also essential to interpret the biological roles of these genes. With the Gene Ontology, biologists can acquire a list of functionally-related genes from microarray experiments.

Our results of functionally-related gene networks are consistent with other studies of metabolic and cellular networks. We find that the network of the four organisms – human, fruit flies, yeast and worms have the property of small-world. Due to the fact that the studies of functionalities of genes in the organisms have not been completed, annotations are updated periodically. Even as the annotations evolve, the conclusion of the existence of small-world property still holds as the characteristic path lengths L in small-world networks grow only logarithmically with the number of nodes[4]. In other words, as more gene products are added to the annotations, L would not be changed by much. Other than the network of worms, all of them also exhibit the scale-free property. Such findings can be of huge implication for the evaluation of newly derived gene product interactions and the practice of medicine. As described in work[23] by Goldberg and Roth, a potential application is to exploit the neighborhood cohesiveness derived from the small-world property of our network to define measures of confidence. Such confidence can be applied to evaluate gene product interactions that are inferred from new data. Another possible scenario could be on searching for new targets for antibiotics, a pharmacologist can utilize the functionally-related gene network to find gene products that are involved in bacterial protein synthesis[b] and other known involvement of biological process in such gene products. Because of the small-world and scale-free properties of the networks, the number of genes that a pharmacologist needs to consider can be significantly reduced.

Acknowledgement

We would like to acknowledge John Pearson at Translational Genomics Research Institute (TGen) for introducing the Gene Ontology and his valuable comments on this paper. Luis Tari is grateful for the support of Edward Suh at TGen. Chitta Baral acknowledges the support of NSF through grant number 0412000. We also appreciate the helpful comments provided by the reviewers of this paper.

[b] Example taken from the Gene Ontology website.

References

1 D. Watts, Networks, dynamics and the small world phenomenon, *American Journal of Sociology*. **105**(**2**), 493-527 (1999).

2 A. Barabási, R. Albert, Emergence of scaling in random networks, *Science*. **286**, 509-512 (1999).

3 D. Watts, Six degrees of interconnection, *Wired Magazine*. **136** (2003).

4 D. Watts, S. Strogatz, Collective dynamics of 'small-world' networks, *Nature*. **393**, 440-442 (1998).

5 R. Albert, H. Jeong, A. Barabasi, Diameter of the World-Wide Web, *Nature.* **400**, 130-131 (1999).

6 S. Bornholdt, H. Schuster, Handbook of Graphs and Networks: from Biological Nets to the Internet and WWW, *Oxford University Press* (2003).

7 L. Hartwell, J. Hopfield, S. Leibler, A. Murray. From molecular to modular cell Biology. *Nature.* **402** supplement. 6761, C47-C52 (1999).

8 H. Jeong, S. Mason, A. Barabasi, Z. Oltva. Lethality and centrality in protein networks. *Nature.* **411**, 41-42 (2001).

9 A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Bio. Evol.* **18**, 1283-1292 (2001).

10 A. Barabási, Z. Oltvai, Network biology: understanding the cell's functional organization. *Nature Reviews Genetics.* **5**, 101 -113 (2004).

11 H. Jeong, B. Tombor, R. Albert, Z. Ottvai, A. Barabasi. The large scale of metabolic networks. *Nature,* **407**, 651-654 (2000).

12 D. Fell, A. Wagner, The small world of metabolism. *Nat Biotechnology.* **18,**112-122 (2000).

13 U. Alon, M. Surette, N. Barkai, S. Leibler, Robustness in bacterial chemotaxis. *Nature (London).* **397,** 168–171 (1999).

14 H. Agrawal, Extreme self-organization in networks constructed from gene expression data. *Phys. Rev. Lett.* **89**, 268702 (2002).

15 V. van Noort, B. Snel, M. Huynen, The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Reports.* **5,** 280-284 (2004).

16 M. Ashburner, C. Ball, J. Blake, D. Botstein, et al., Gene Ontology: tool for the unification of biology. *Nature Genetics.* **25**, 25 – 29 (2000).

17 B. Bollobas, Random Graphs. *Academic Press*, London (1985).

18 D. Watts, Small Worlds. *Princeton University Press*, Princeton (1999).

19 http://www.geneontology.org/ontology/GO.defs Gene Ontology definitions

20 http://www.geneontology.org/GO.evidence.html GO evidence codes

21 S. Draghici, P. Khatri, R. Martins, C. Ostermeier, S. Krawetz, Global functional profiling of gene expression. *Genomics.* **81**(**2**), 98-104 (2003).

22 B. Zeeberg et al., GoMiner: A Resource for Biological Interpretation of Genomic and Proteomic Data. *Genome Biology.* **4**(**4**), R28 (2003).

23 D. Goldberg, F. Roth, Assessing experimentally derived interactions in a small world. *Proc. of the Nat. Acad. of Sciences.* **100**(**8**), 4372-4376 (2003).