33rd Annual International Conference of the IEEE EMBS
Boston, Massachusetts USA, August 30 - September 3, 2011

Topology Aware Functional Similarity of Protein Interaction Networks Based on Gene Ontology

Fei Li, Xiuliang Cui, Dafei Xie, Xiaochen Bo and Shengqi Wang

Abstract—**Functional comparison and alignment of Protein Interaction Networks (PINs) will enable a better understanding of cellular organization and processes. Gene Ontology (GO) provides a structured standard vocabulary of functional terms of gene products, and has been used to measure the functional similarity between proteins. In this study, we propose an algorithm to measure the functional similarity between PINs based on GO. The algorithm simultaneously takes the PIN's network topology and semantic similarity of the component proteins into account. We employ the algorithm to measure the similarity between pathways present in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and cluster the pathways according to similarity. The results show great consistency with the function of these pathways.**

I. Introduction

It is increasingly clear that most biological characteristics arise from complex interactions between cell's numerous constituents [1]. Spurred on by advances in high-throughput experimental techniques (e.g., yeast two-hybrid [2], [3], co-immunoprecipitation [4]) and new computational approaches for protein-protein interaction prediction, data on molecular interactions are increasing exponentially. In systems biology, a group of protein-protein interactions are often modeled as complex networks. Analyzing and understanding protein interaction networks (PINs) have become a key challenge for biology and attract much interest in recent years.

As most biological knowledge is derived from comparison and classification, measuring the functional similarity of PINs is a powerful method to address the challenge. Network comparison aims to contrast two or more interaction networks, which represent different species or different conditions. Just like sequence alignment has pushed our understanding of evolution, biology and disease forward greatly, network comparison and alignment will have a similar impact. However, and although sequence comparison has long been a staple of biological research, the development of a similar toolbox for comparing biological networks is still in its infancy [5].

The existing network comparison methods can mainly be classified into three types: network alignment, network integration and network query [4]. All of these are based on sequence similarity. Singh et al. introduce an algorithm, IsoRank, for global alignment of multiple PPI networks to get conserved subgraphs [5]. It simultaneously uses sequence similarity and network data. Kelley et al. propose a strategy for aligning two PPI networks that combines interaction topology and protein sequence similarity to identify high-scoring common paths and complexes [6]. However, they are all focus on detecting subnetworks that are sequence conserved or topology conserved, according to compare two or more networks across species.

Here, we propose a topology-aware algorithm based on GO to measure the functional similarity between different PINs within species, such as condition-responsive subnetworks or functional modules. The algorithm simultaneously takes the PIN networks topology and semantic similarity of the component proteins into account. We employ the algorithm to measure the similarity between pathways present in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database and cluster the pathways according to similarity. The results show great consistency with the function of these pathways.

II. Methord

A. GO-based Functional Similarity Measures

Gene Ontology provides a structured standard vocabulary of functional terms and allows for coherent annotation of gene products [7]. The GO ontologies are presented as directed acyclic graphs (DAG) in which the terms form nodes and the two kinds of semantic relations ('is-a' and 'part-of') form edges. And they are similar to hierarchies but differ in that a child, or more specialized, term can have many parents, or less specialized, terms. The GO is divided into three orthogonal ontologies, biological process (BP), molecular function (MF), and cellular component (CC). The cellular component terms characterize the location of gene products in the cell. The molecular function terms represent the molecular level activities of proteins. The biological process terms describe a series of events accomplished by one or more proteins. GO annotations capture the available functional information of a gene product and can be used as a basis for defining a measure of functional similarity between gene products.

The existing GO based measures for the functional similarity mainly focused on the similarity of GO terms and genes, which were proposed to measure the semantic

Manuscript received March 22, 2011. This work was supported by grant from the national nature science foundation of China (No. 30530650).

Fei Li is with Beijing Institute of Radiation Medicine, Beijing, 100850 China (phone: 086-010-66931422; e-mail: lifei@bmi.ac.cn).

978-1-4244-4122-8/11/$26.00 ©2011 IEEE

similarity of GO terms on the basis of information content (IC) or structure information of the Ontological DAG. Most of these methods utilize IC to measure the semantic similarity of GO terms, which were originally developed for the WordNet in natural language processing. Resnik defines similarity between two terms as the IC of their most informative common ancestor (MICA) [8]. Lin et al. take the distance between the terms and their common ancestor into account, and propose new methods based on Resnik's method [9]. As a method using structure information, Wang et al. proposed a method to encode a GO term's semantics into a numeric value by aggregating the semantic contributions of their ancestor terms in the DAG [10].

Similarity measures for genes or gene products have also been developed on the basis of the above semantic similarity measures of GO terms. The most common methods of measuring gene product functional similarity have been pairwise approaches based on IC. Lord et al. were the first to propose employing GO annotations to measure semantic similarity of gene products [11]. They defined the semantic similarity between annotated proteins as the average similarity of all the GO terms which can be calculated on the basis of IC measure. Sevilla et al. used the maximum of the pairwise terms similarity instead [12]. Couto et al. [13] opted for a composite average in which only the best-matching term pairs are considered. Schlicher et al defined a new measure which combining Lin's and Resnik's similarity measures to calculate to similarity between GO terms, and employ the maximum combination strategy [14].

As is discussed above, there are various methods to measure the functional similarity between gene products. As there is no direct way to ascertain the true functional similarity between two gene products, there is no clear best measure for comparing terms or gene products. Several efforts have been made to evaluate these semantic similarity measures, and the results show that Resnik's method performs better than other methods. Here, we employ Resnik's measure S_{Resnik} to calculate the semantic similarity of GO terms, and take the average similarity of the best-match terms as protein similarity.

B. Topology-aware Similarity Measure

Calculating the semantic similarity between the PPI networks relies on two aspects of information encoded in the nodes and edges, in other words the similarity between the sets of proteins contained in the networks and the topology message.

The proteins are typically annotated with more than one GO term, and they may perform different functions in different environmental conditions. Methods predicting physical interactions between proteins and functional modules based on their similarity rely on direct observation that interacting proteins often function in the same biological process or locate closely in the cell, so two proteins acting in the same biological process or co-localization are more likely to interact with each other [15-17]. We assume that, firstly a protein prefers to choose self-similar proteins to be its neighbors, and secondly the neighbors will affect the function of the protein. Based on these two hypotheses, we introduce our topology-aware similarity measure. Our algorithm works in two stages: Firstly it calculates a functional similarity score for each possible match between nodes of the two networks. Then it revise the similarity score by taking the neighbors into consideration, and a new similarity score is derived. The two stages are done recursively.

DEFINITION 1 (Network annotation). A powerful way of representing and analyzing the PINs is a graph $G = (V, E)$ where V is the set of nodes and E is the set of edges. Each node corresponds to a protein and an edge indicates a direct physical interaction between the proteins. $|V|$ is the size of this set, and $|E|$ is the number of edges. Let $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ denote the PINs, $\mathrm{S}(V_1, V_2)$ denotes the similarity matrix whose elements $\mathrm{S}(v_1, v_2)$ are the similarity value between node $v_1 \in V_1$ and node $v_2 \in V_2$. Various semantic similarity measures can be used in the calculation. Here we employ Resnik's measure, and denote the corresponding similarity matrix as $\mathrm{S}_{Resnik}(V_1, V_2)$

DEFINITION 2 (Aggregate function). Given a matrix $\mathrm{S}(V_1, V_2)$, we define an aggregate function $\mathrm{OA}\big(\mathrm{S}(V_1, V_2)\big)$ maps the matrix $\mathrm{S}(V_1, V_2)$ to a single value, which employs the maximum weighted bipartite graph matching algorithm to get a mapping between the nodes of G_1 and G_2, and take the average of $\mathrm{S}(v_1, v_2)$ over each mapping. The matching algorithm gives an optimal map of two sets V_1 and V_2, which is presented as a set of node pairs as M whose elements (v_i, v_j) means the matching between node v_i in G_1 and node v_j in G_2 is preferable. So we define the aggregate function $\mathrm{OA}\big(\mathrm{S}(V_1, V_2)\big)$ as follows.

$$\mathrm{OA}\big(\mathrm{S}(V_1, V_2)\big) = \frac{1}{|M|} \sum_{(v_i, v_j) \in M} \mathrm{S}(v_i, v_j)$$

DEFINITION 3 (Topology-aware similarity). Let $N(v)$ denote the set of proteins interacting directly with v, and $\mathrm{S}(N(v_1), N(v_2))$ denotes the similarity matrix of the neighbors of v_1 and v_2 respectively. We define the Topology-aware similarity $\mathrm{S}^*(v_1, v_2)$ as follows.

$$\mathrm{S}^*(v_1, v_2) = \frac{1}{2}\Big(\mathrm{S}_{Resnik}(v_1, v_2) + OA\left(\mathrm{S}_{Resnik}\big(N(v_1), N(v_2)\big)\right)\Big)$$

And the final network similarity is defined as follows.

$$\mathrm{S}(G_1, G_2) = OA\big(\, \mathrm{S}^*(V_1, V_2)\big).$$

III. Result

KEGG (Kyoto Encyclopedia of Genes and Genomes)[18] is a bioinformatics resource for understanding higher-order functional meanings and utilities of the cell or the organism from its genome information. The data at http://www.genome.ad.jp/kegg/ integrates current knowledge on molecular interaction networks such as pathways and complexes (PATHWAY database), information about genes and proteins generated by genome projects (GENES/SSDB/KO databases) and reactions (COMPOUND/GLYCAN/REACTION databases). These three types of database actually represent three graph objects, called the protein network, the gene universe and the chemical universe. The protein network, which is the most unique data object in KEGG, is stored as a collection of pathway maps in the PATHWAY database, representing wiring diagrams of proteins and other gene products responsible for various cellular functions [19]. And in this paper, we employ the human-related pathway data in KEGG pathway data. The KEGG pathway maps are hierarchically classified reflecting the map resolution and functional modules at different levels. We can employ the hierarchically classified maps as the artificial classification sample set to assess the method measuring the similarity of PPI networks.

GO-based semantic similarity can be used to compare gene products from the three aspects: BP, MF and CC. We employ BP to compute the similarity of the PPI networks, and take cellular processes related pathways in KEGG as test data set to assess the method measuring the similarity between PPI networks.

TABLE I
SIMILARITY BETWEEN HUMAN CELLULAR PROCESS NETWORKS

N	Pathway	1	2	3	4	5	6	7	8
1	Cell cycle	1.00	0.62	0.35	0.33	0.34	0.24	0.28	0.32
2	p53 signaling pathway	0.62	1.00	0.46	0.46	0.33	0.38	0.34	0.40
3	Apoptosis	0.35	0.46	1.00	0.53	0.34	0.40	0.37	0.54
4	Focal adhesion	0.33	0.46	0.53	1.00	0.52	0.62	0.48	0.64
5	Adherens junction	0.34	0.33	0.34	0.52	1.00	0.54	0.39	0.56
6	Tight junction	0.24	0.38	0.40	0.62	0.54	1.00	0.42	0.48
7	Gap junction	0.28	0.34	0.37	0.48	0.39	0.42	1.00	0.59
8	Regulation of actin cytoskeleton	0.32	0.40	0.54	0.64	0.56	0.48	0.59	1.00

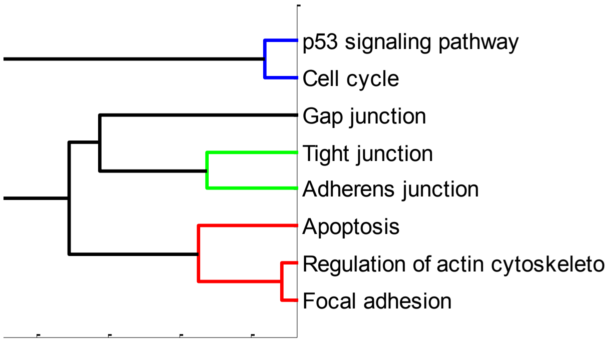

Fig. 1. The clustering result based on the network similarity measure of human cellular process pathways

There are 17 cellular process related pathways which contain four subcategories: Transport and Catabolism, Cell Motility, Cell Growth and Death and Cell Communication, and 8 of which are human related. There are 86 to 1869 interactions, 75 to 206 proteins, 95% of which have GO annotations in each pathway, as show in table. We convert these pathways to PINs, and employ our algorithm to calculate the similarity between them. The results are shown in Table 1. The similarity between cell cycle and p53 signaling pathway is the maximum score 0.617, and the similarity between cell cycle and tight junction is the minimum score 0.237. There are 7 pairs of pathways are high similar (similarity score >0.5). The average similarity (except the diagonal ones) is 0.437, which is high relatively because these pathways are all belong to human cellular process.

We cluster the human cellular related pathways based on the similarity matrix shown in Table 1. The hierarchical cluster analysis dendrogram is draw to examine the ability of our algorithm to distinguish different pathways. We use $1 - S(G_1, G_2)$ as dissimilarity matrix, and employ the furthest distance method. The result is show as Fig. 1. According to this clustering result base on network similarity, The cellular processes related pathways can be divided into three groups: 1) p53 pathway Cell cycle pathway and; 2) Adherens junctions channels, Tight junction channels and Gap junction channels; 3) Actin cytoskeleton regulatory pathways, Focal adhesion pathway and Apoptosis pathways. This division indicates that the Topology-aware similarity measure makes a good distinction between these pathways.

IV. Conclusion

In the complex network research area, such as the Internet networks, social networks, the network similarity measure has been widely applied, and in the field of biological research, functional similarity measure of protein interaction network is still in the preliminary stage. Our topology-aware similarity measure, which integrates the topology information of PINs, gives a more reasonable measure of functionality similarity of protein interaction network. Further clustering analysis on KEGG pathways shows that our similarity measure can effectively distinguish similar functional pathways and give a quite reasonable clustering result.

References

[1] A. Barabasi and Z. N. Oltvai, “Network biology: understanding the cell's functional organization,” *Nat Rev Genet*, vol. 5, no. 2, pp. 101-113, Feb. 2004.

[2] P. Uetz et al., “A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae,” *Nature*, vol. 403, no. 6770, pp. 623-627, Feb. 2000.

[3] T. Ito, T. Chiba, R. Ozawa, M. Yoshida, M. Hattori, and Y. Sakaki, “A comprehensive two-hybrid analysis to explore the yeast protein interactome,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 8, pp. 4569-4574, Apr. 2001.

[4] N. J. Krogan et al., “Global landscape of protein complexes in the yeast Saccharomyces cerevisiae,” *Nature*, vol. 440, no. 7084, pp. 637-643, Mar. 2006.

[5] R. Sharan and T. Ideker, “Modeling cellular machinery through biological network comparison,” *Nat Biotech*, vol. 24, no. 4, pp. 427-433, Apr. 2006.

[6] B. P. Kelley, B. Yuan, F. Lewitter, R. Sharan, B. R. Stockwell, and T. Ideker, “PathBLAST: a tool for alignment of protein

interaction networks," *Nucleic Acids Research*, vol. 32, pp. W83-W88, Jul. 2004.

[7] M. Ashburner et al., "Gene Ontology: tool for the unification of biology," *Nat Genet*, vol. 25, no. 1, pp. 25-29, May. 2000.

[8] P. Resnik, "Using Information Content to Evaluate Semantic Similarity in a Taxonomy," *Proceedings of IJCAI-95*, Nov. 1995.

[9] D. Lin, "An Information-Theoretic Definition of Similarity," in *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 296-304, 1998.

[10] J. Z. Wang, Z. Du, R. Payattakool, P. S. Yu, and C. Chen, "A new method to measure the semantic similarity of GO terms," *Bioinformatics*, vol. 23, no. 10, pp. 1274-1281, May. 2007.

[11] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble, "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation," *Bioinformatics*, vol. 19, no. 10, pp. 1275-1283, Jul. 2003.

[12] J. L. Sevilla et al., "Correlation between Gene Expression and GO Semantic Similarity," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 2, no. 4, pp. 330-338, 2005.

[13] F. M. Couto, M. J. Silva, and P. M. Coutinho, "Measuring semantic similarity between Gene Ontology terms," *Data & Knowledge Engineering*, vol. 61, no. 1, pp. 137-152, Apr. 2007.

[14] A. Schlicker, F. Domingues, J. Rahnenfuhrer, and T. Lengauer, "A new measure for functional similarity of gene products based on Gene Ontology," *BMC Bioinformatics*, vol. 7, no. 1, p. 302, 2006.

[15] W. Huh et al., "Global analysis of protein localization in budding yeast," *Nature*, vol. 425, no. 6959, pp. 686-691, Oct. 2003.

[16] H. Wu, Z. Su, F. Mao, V. Olman, and Y. Xu, "Prediction of functional modules based on comparative genome analysis and Gene Ontology application," *Nucleic Acids Research*, vol. 33, no. 9, pp. 2822 -2837, 2005.

[17] X. Wu, L. Zhu, J. Guo, D. Zhang, and K. Lin, "Prediction of yeast protein–protein interaction network: insights from the Gene Ontology and annotations," *Nucleic Acids Research*, vol. 34, no. 7, pp. 2137 -2150, 2006.

[18] M. Kanehisa, S. Goto, M. Furumichi, M. Tanabe, and M. Hirakawa, "KEGG for representation and analysis of molecular networks involving diseases and drugs," *Nucl. Acids Res.*, vol. 38, no. 1, pp. D355-360, Jan. 2010.

[19] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG resource for deciphering the genome," *Nucl. Acids Res.*, vol. 32, no. 1, pp. D277-280, Jan. 2004.