

SEMANTIC SIMILARITY DEFINITION OVER GENE ONTOLOGY BY FURTHER MINING OF THE INFORMATION CONTENT

YUAN-PENG LI¹ and BAO-LIANG LU^{1,2*}

¹*Department of Computer Science and Engineering, Shanghai Jiao Tong University,*

²*Laboratory for Computational Biology, Shanghai Center for Systems Biomedicine,
800 Dong Chuan Road, Shanghai 200240, China*

E-mail: {yuanpengli, bllu}@sjtu.edu.cn

The similarity of two gene products can be used to solve many problems in information biology. Since one gene product corresponds to several GO (Gene Ontology) terms, one way to calculate the gene product similarity is to use the similarity of their GO terms. This GO term similarity can be defined as the semantic similarity on the GO graph. There are many kinds of similarity definitions of two GO terms, but the information of the GO graph is not used efficiently. This paper presents a new way to mine more information of the GO graph by regarding edge as information content and using the information of negation on the semantic graph. A simple experiment is conducted and, as a result, the accuracy increased by 8.3 percent in average, compared with the traditional method which uses node as information source.

Keywords: Gene Ontology; Semantic Similarity; Information Content.

1. Introduction

1.1. Gene Ontology

Gene Ontology (GO)¹ was created to describe the attributes of genes and gene products using a controlled vocabulary. It is a powerful tool to support the research related to gene products and functions. For example, it is widely used in solving the problems including identifying functionally similar genes, and the protein subcellular or subnuclear location prediction. GO has not been completed and the number of biological concepts in it is still increasing. As GO puts its primary focus on coordinating this increasing number of concepts, at the risk of losing the characteristics of formal ontology, it has some differences from the ontology in Philosophy or Computer Science.^{2,3} Gene Ontology Next Generation (GONG)⁴ was established to solve this problem and discuss the maintenance of the large-scale biological ontology. Recently, as the use of similarities on GO is increasing, some convenient databases and softwares⁵⁻⁸ are developed and freely available, which makes it easier to use GO semantic similarity.

*To whom correspondence should be addressed.

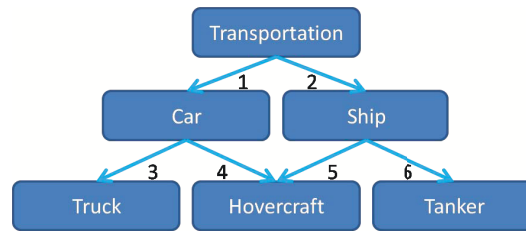


Fig. 1. Example of ontology.

The Gene Ontology⁹ is made up of three ontologies: Biological Process, Molecular Function and Cellular Component. On May 2007, there are 13,552 terms for Biological Process, 7,609 for Molecular Function and 1,966 for Cellular Component.

From the graph point of view, each of these ontologies is a connected directed acyclic graph (DAG), with only one root node in that ontology. It is also true that a special node can be set to combine these three ontologies into one, i.e., the special node has the three root nodes of each ontology as its children.

Each node represents a concept, or an ontology term. If two concepts have some relationship, an edge is drawn from one to the other. Gene Ontology only has “is_a” relationship and “part_of” relationship. “is_a” relationship indicates that the concept in the in-node of the edge contains the concept in the out-node. The example in Figure 1 is not Gene Ontology, but just an ordinary ontology for explanation. In the ontology, edge 3 means that “Truck” is a kind of “Car”. “is_a” relationship can also be regarded as a standard that distinguishes a concept from other concepts contained in the parent concept. Here, “Truck” is distinguished from “Hovercraft” by the standard the edge 3 provides. “part_of” relationship denotes that the in-node concept has the out-node concept as one of its parts.

If a concept is contained in another concept, then this information is considered positive information. On the other hand, when a concept is NOT contained in another concept, this information is considered negative information. In Figure 1, edge 4 is negative information for “Truck”.

1.2. GO and Similarity between Gene Products

The final aim of this research is to define the similarities between gene products using GO information. Since each gene product has several GO terms, the similarity of gene product can be calculated from the similarities of these GO terms. There are two steps in this process.

The first step is to obtain the similarity of two GO terms from the GO graph. This is the main focus of this paper.

The second step is to get the gene product similarity from the GO term similarities. Let g_1 and g_2 be the GO term vectors of two gene products A and B, in which 1 means the gene product has the GO term, while 0 means it does not. In

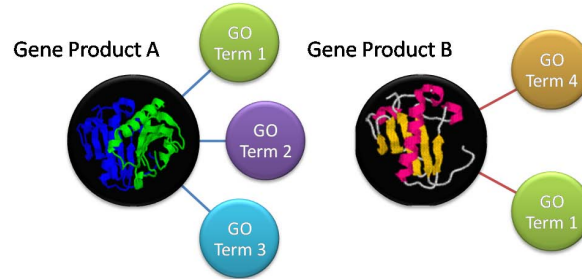


Fig. 2. Example of gene products and their corresponding GO terms.

the example of Figure 2^a, g_1 and g_2 will be as follows.

$$g_1 = (0, 1, 1, 1, 0, 0)^T \quad g_2 = (0, 1, 0, 0, 1, 0)^T$$

Also, let M be a square matrix, in which the value of the i th row and the j th column represents the similarity of the i th and the j th GO terms, obtained in the first step. Then the similarity of two gene products $Sim(A, B)$, or $Sim(g_1, g_2)$, can be defined as follows.

$$Sim(g_1, g_2) = g_1^T M g_2 \quad (1)$$

This research is conducted to fully mine the information in GO graph and define similarities between GO terms. In other words, to get a better similarity matrix M .

1.3. Related Work

There are many semantic similarity definitions of GO terms. Some representative ones can be classified by two kinds of standards (Table 1).

The first standard is to divide the definitions into probability-based and structure-based ones. The probability-based methods depend on the occurrence frequency of each GO term in some database. Resnik,¹⁰ Jiang and Conrath,¹¹ and Lin¹² provided their definitions from this point of view. Lord¹³ introduced these definitions into Gene Ontology. Later, Couto¹⁴ proposed a method to better apply them to DAGs rather than trees. This kind of methods is based on information theory, and seems to be reasonable. However, it relies on a particular database, SWISS-PROT. On the other hand, another idea is developed to define the similarity from the structure of ontology. The definitions proposed by Rada,¹⁵ Wu,¹⁶ Zhang,⁶ and Gentleman⁷ are examples of this idea. They made it possible to reasonably obtain the similarity of two GO terms in any database, even if the distribution of the data is highly unbalanced or the size of the database is quite small.

^aPicture source is [<http://lectures.molgen.mpg.de/ProteinStructure/Levels/index.html>].

Table 1. Similarity definition methods.

	Probability-based	Structure-based
Distance	Jiang and Conrath	Rada
Info content	Resnik	Zhang, Wu
Content ratio	Lin	Gentleman

The definition measures can also be classified by another standard into three groups. The first group is to define the similarity of two nodes by the distance between them. Rada¹⁵ proposed the original framework of this idea. Jiang and Conrath¹¹ investigated the weights of the edges to make it more reasonable. The second group of definitions is to calculate the shared information content of two nodes. Resnik¹⁰ first proposed the using of information content. Zhang⁶ and Gentleman⁷ provided similar definitions based on the structure of ontology. The third group of definitions is to compare the shared information of the two concepts and all the information needed to describe both of these concepts. Lin¹² and Gentleman⁷ did some work concerning this idea.

2. Method

2.1. Notations

c denotes a term, or a node, in an ontology graph. An edge e fluxes into c means that there exists a path from the root node to c which contains e . The induced graph $V(c)$ of c is the graph made up of all paths from the root node to c . $|V|_n$ and $|V|_e$ denote the number of nodes and the number of edges in V .

In Figure 1, for example, if c is “Hovercraft”, the edge $e = 4$ fluxes into c , because there exists a path $\{\{\text{“Transportation”, “Car”, “Hovercraft”}\}, \{1, 4\}\}$ from the root node “Transportation” to c , which contains e (Figure 3(a), left). The induced graph $V(c)$ is $\{\{\text{“Transportation”, “Car”, “Ship”, “Hovercraft”}\}, \{1, 2, 4, 5\}\}$ (Figure 3(b)). $|V(c)|_n = |\{\text{“Transportation”, “Car”, “Ship”, “Hovercraft”}\}| = 4$ and $|V(c)|_e = |\{1, 2, 4, 5\}| = 4$.

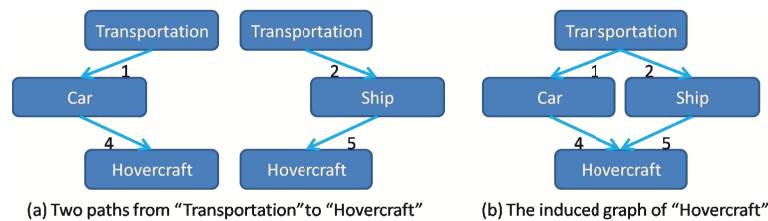


Fig. 3. The paths and the induced graph of “Hovercraft”.

2.2. Traditional Definition

The idea of Gentleman⁷ is used as a traditional definition. The similarity is defined as the number of nodes that the two induced graphs share in common, divided by the number of nodes contained in at least one of the two induced graphs.

$$SimUI(c_1, c_2) = \frac{|V(c_1) \cap V(c_2)|_n}{|V(c_1) \cup V(c_2)|_n} \quad (2)$$

In the example of Figure 1, the similarity of “Truck” and “Hovercraft” is 0.4 since they have 2 nodes in both induced graphs and 5 in at least one induced graph.

The basic idea is similar to that of Lin. Here, the information content of a node is regarded as being represented by its ancestor nodes. The shared information of two nodes is the intersection of their ancestor node sets. All information needed to describe the concepts of two nodes is the union of their ancestor node sets.

The ideas proposed in this paper can be considered as the counterparts of this method, and one of the differences is that the proposed ideas use edges, instead of nodes, to calculate information content. Therefore, SimUI should be chosen as a traditional method to be compared with the new ones.

2.3. Proposed Similarity Definitions

The first new method provides the positive similarity of two nodes c_1 and c_2 . It is similar to SimUI, but edges are used instead of nodes.

$$SimPE(c_1, c_2) = \frac{|V(c_1) \cap V(c_2)|_e}{|V(c_1) \cup V(c_2)|_e} \quad (3)$$

Since GO is a DAG, unlike tree, edges contain more information than nodes (SEE 4.1). In Figure 1, the induced graphs of “Truck” and “Hovercraft” have one edge in common and 5 different edges altogether. Therefore the similarity is 0.2.

On the other hand, for a node c and an edge e , if e has its in-node as an ancestor of c , but e does not flux into c , it means that the node c does not meet the standard provided by the edge e . To define the negative similarity, the negative edge set should be defined first. The negative edge set of c , $NES(c)$, denotes the set of edges that have in-nodes in the induced graph of c , but not their out-nodes. This consideration of out edges of each node can also be found in the local density introduced by Jiang and Conrath.¹¹

$$NES(c) = \{ \langle c_{in}, c_{out} \rangle \in E \mid c_{in} \in V(c), c_{out} \notin V(c) \} \quad (4)$$

Here, E is the set of all edges in the GO graph. Then the negative similarity can be defined as follows.

$$SimNE(c_1, c_2) = \frac{|NES(c_1) \cap NES(c_2)|}{|NES(c_1) \cup NES(c_2)|} \quad (5)$$

Here, the numerator means the size of shared negative information of both nodes, i.e., the number of the standards that c_1 and c_2 both do NOT meet. And the denominator indicates the number of standards that at least one of the nodes does NOT meet. In Figure 1, the similarity of “Truck” and “Hovercraft” is 0.

To combine these two similarities, the easiest way is to multiple them together.

$$SimEG(c_1, c_2) = SimPE(c_1, c_2) \cdot SimNE(c_1, c_2) \quad (6)$$

For an edge e that has both its in-edge and out-edge NOT in $V(c)$, whether c meets the standard provided by e is unknown, or meaningless. In Figure 1, the standard of edge 3 makes the concept “Truck” different from the concept “Car”. But this standard is meaningless when applied to the concept “Tanker”, since “Tanker” is not a “Car” at all. Therefore, such edge is not considered to contain either positive or negative information of c .

3. Results

To evaluate the methods UI, PE and EG, an experiment of protein subcellular location prediction was conducted. The experiment was composed of several steps. Firstly, the proteins were randomly chosen, and the corresponding GO terms were found. Secondly, the chosen proteins were divided into training and test samples. Thirdly, a classifier was used to predict the subcellular locations of test samples from the subcellular locations of the train samples, using their similarities.

3.1. Dataset

The Gene Ontology structural data are from the Gene Ontology.⁹ As the whole ontology contains 32,297 of “is_a” relationships, but only 4,759 of “part_of” relationships, all “part_of” relationships are ignored to make the problem simple.

The training and test data were obtained by choosing from the dataset created by Park and Kanehisa.¹⁸ The GO terms corresponding to these proteins were obtained through the InterPro. i.e., corresponding InterPros were first found from the protein, and then the GO terms of the InterPros were marked to the protein. If one protein was marked by more than one exactly the same GO terms, only one of them was left. In the experiment, several large classes (Table 2) of subcellular locations were used. To avoid the unbalance between the classes, 600 samples were randomly chosen for each of these classes. Each of these samples had at least one GO term so that the similarity of any two chosen proteins could be found via their GO term similarities.

3-fold cross validation was used to assess the performances of the definitions. Each class was divided into three sets of samples randomly. Then, two of these sets in each class were chosen and mixed as a training set and the one left over was used in a test set. Consequently, three groups of training and test sets were prepared^b.

^b<http://bcmi.sjtu.edu.cn/~liyuanpeng/APBC2008/{train,test}{1,2,3}.txt>.

Table 2. Number of samples in each class.

Class	Subcellular location	# of Samples
1	Chloroplast	600
2	Cytoplasmic	600
3	Extracellular	600
4	Mitochondrial	600
5	Nuclear	600
Total		3000

3.2. Classifier

k-Nearest Neighbor (k-NN) classifier was designed to predict the subcellular locations, or classes, of the test samples. The distance of two samples was defined as the minus value of their similarity, and majority voting method was used. If two classes appeared the same number of times in the k-nearest neighbors of a test sample, one of them was selected randomly as the predicted class of that test sample.

3.3. Tables and Graphs

The prediction accuracies of the experiments are listed on Table 3 as percentages, followed by the corresponding k values that brought the best results. The three graphs in Figure 5 demonstrate the accuracies for each group as the change of k values. In each of these graphs, the horizontal axis represents the value of k and the vertical axis represents the accuracy percentage. The accuracies of each class, corresponding to the best k values, are listed on Table 4, for each group and the average. Their increases are plotted in Figure 6. In all tables and graphs, "Increase" means the difference between the values of the EG and UI methods.

4. Discussion

4.1. The Use of Edges and Negative Information

From the results, it is obvious that PE has advantage over UI, and EG has advantage over PE. The reason can be found in information gain. Consider a small ontology example in Figure 4, $\text{SimUI}(B,D)$ will not change even if the edge from A to D is deleted. In other words, the information of the edge is ignored. $\text{SimPE}(B,D)$ can contain this information, but the information of the edge from A to C is not

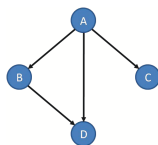
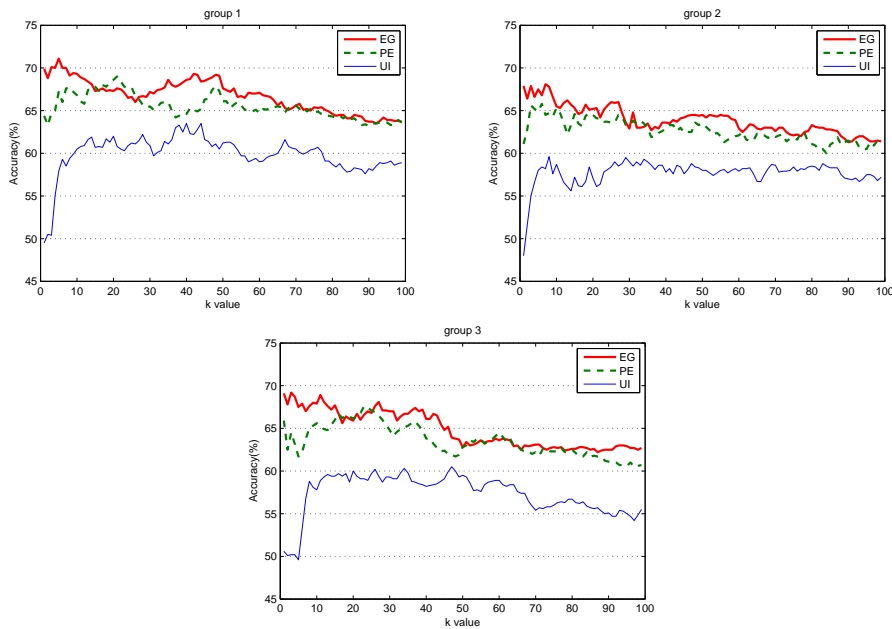


Fig. 4. Example of ontology structure.

Table 3. The accuracies of each group (%)

group	UI (k)	PE (k)	EG (k)	Increase
1	63.5 (40)	69.0 (21)	71.1 (5)	7.6
2	59.6 (8)	65.8 (6)	68.1 (7)	8.5
3	60.5 (47)	67.6 (23)	69.2 (3)	8.7
average	61.2	67.5	69.5	8.3

Fig. 5. The relationship of total accuracies and values of k for each group and method.

included. And when SimEG(B,D) is used, this edge information can also be included. Therefore, more information can be used in PE than in UI, and in EG than in PE.

4.2. The Difference among Classes

Table 4 and Figure 6 show that different classes prefer different methods of classification. For class 5, the accuracy was already close to 100% when the UI method was applied, and this could be the reason for the less change of the accuracies when the PE and EG methods were used.

4.3. More Comparison Results

An experiment, without cross validation, was conducted for each kind of structure-based methods. The results were 65.2% for method of Rada,¹⁵ 61.4% for Wu,¹⁶ 66.0% for Zhang⁶ and Gentleman,⁷ 64.5% for UI, 69.4% for PE, and 70.8% for EG.

Table 4. Class accuracies corresponding to the best k values (%).

group 1					group 2				
Class	UI	PE	EG	Increase	Class	UI	PE	EG	Increase
1	28.5	39.0	40.0	11.5	1	35.0	38.0	45.0	10.0
2	62.5	71.0	77.0	14.5	2	58.0	64.0	65.0	7.0
3	70.5	83.5	83.5	13.0	3	56.5	71.0	74.0	17.5
4	60.5	57.0	60.5	0.0	4	54.5	61.5	61.5	7.0
5	95.5	94.5	94.5	-1.0	5	94.0	94.5	95.0	1.0

group 3					average				
Class	UI	PE	EG	Increase	Class	UI	PE	EG	Increase
1	44.5	43.0	47.5	3.0	1	36.0	40.0	44.2	8.2
2	58.5	74.5	73.5	15.0	2	59.7	69.8	71.8	12.1
3	62.0	75.5	81.5	19.5	3	63.0	76.7	79.7	16.7
4	41.0	50.0	47.5	6.5	4	52.0	56.2	56.5	4.5
5	96.5	95.0	96.0	-0.5	5	95.3	94.7	95.2	-0.1

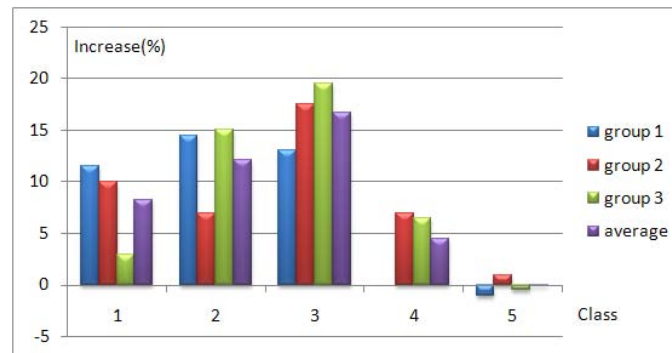


Fig. 6. Increases in each class and group.

5. Conclusions

From the experiment, it can be concluded that the use of edges as information carriers is better than the use of nodes, and that negative information, combined with positive information, provides further support for better predictability.

6. Acknowledgments

The authors thank Bo Yuan, Yang Yang and Wen-Yun Yang for their valuable comments and suggestions. This research is partially supported by the National Natural Science Foundation of China via the grant NSFC 60473040.

References

1. The Gene Ontology Consortium. Gene Ontology: tool for the unification of biology. *Nature Genet*, 25:25-29, 2000.
2. B. Smith, J. Williams, S. Schulze-Kremer. The Ontology of the Gene Ontology. *AMIA Symposium Proceedings*, 609-613, 2003.
3. M. E. Aranguren, S. Bechhofer, P. Lord, U. Sattler and R. Stevens. Understanding and using the meaning of statements in a bio-ontology: recasting the Gene Ontology in OWL. *BMC Bioinformatics*, 8:57, 2007.
4. Gene Ontology Next Generation [<http://gong.man.sc.uk>]
5. E. Camon, M. Magrane, D. Barrell, V. Lee, E. Dimmer, J. Maslen, D. Binns, N. Harte, R. Lopez, and R. Apweiler. The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research*, 32:D262-D266, 2004.
6. P. Zhang, J. Zhang, H. Sheng, J. J Russo, B. Osborne and K. Buetow. Gene functional similarity search tool (GFSST). *BMC Bioinformatics*, 7:135, 2006.
7. R. Gentleman. Visualizing and Distances Using GO. 2006.
8. H. Froehlich, N. Speer, A. Poustka, T. Beissbarth. GOSim - An R-package for computation of information theoretic GO similarities between terms and gene products. *BMC Bioinformatics*, 8:166, 2007.
9. the Gene Ontology [<http://www.geneontology.org>].
10. P. Resnik. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *Proc. the 14th International Joint Conference on Artificial Intelligence*, 448-453, 1995.
11. J. J. Jiang and D. W. Conrath. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. *Proc. International Conference Research on Computational Linguistics, ROCLING X*, 1997.
12. D. Lin. An Information-Theoretic Definition of Similarity. *Proc. of 15th International Conference on Machine Learning*, 296-304, 1998.
13. P. Lord, R. Stevens, A. Brass and C. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275-1283, 2003.
14. F. Couto, M. Silva, P. Coutinho. Semantic Similarity over the Gene Ontology: Family Correlation and Selecting Disjunctive Ancestors. *Conference in Information and Knowledge Management*, 2005.
15. R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and Application of a Metric on Semantic nets. *IEEE Transaction on Systems, Man and Cybernetics*, 19(1):17-30, 1989.
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*, 33(9):2822-2837, 2005.
17. J. L. Sevilla, V. Segura, A. Podhorski, E. Guruceaga, J. M. Mato, L. A. Martinez-Cruz, F. J. Corrales, and A. Rubio. Correlation between Gene Expression and GO Semantic Similarity. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(4), 2005.
18. K. J. Park and M. Kanehisa. Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs. *Bioinformatics*, 19(13):1656-1663, 2003.
19. Z. Lei, Y. Dai. Assessing protein similarity with Gene Ontology and its use in subnuclear localization prediction. *BMC Bioinformatics*, 7:491, 2006.