A New Path Length Measure Based on GO for Gene Similarity with Evaluation Using SGD Pathways

Anurag Nagar
University of Houston-Clear Lake
Houston, TX, 77058, USA

Hisham Al-Mubaid
University of Houston-Clear Lake
Houston, TX, 77058, USA

Abstract

We propose a new method for measuring the semantic similarity of genes based on path length between their annotation terms in the Gene Ontology. Our method applies an exponential transfer function to the average path length between two genes to compute their similarity. The non-linear measure ensures that the semantic similarity decreases with distance and proves to be quite competitive when compared to other measures. The advantage of the proposed measure is its simplicity and ease of implementation which gives it a great appeal in this domain. The measure uses only one feature (path length) for computing the similarity between genes. For validation purposes, we computed the similarity of genes from the Saccharomyces Genome Database (SGD) taking part in various cellular pathways. We analyzed 152 pathways from SGD and compared our similarity results with two of the leading measures. The proposed measure proved to be very competitive in all cases and the clustering results showed that our method is able to surpass the leading methods in certain cases.

Keywords: Gene similarity, GO term similarity, Gene similarity in SGD.

1. Introduction

One of the greatest projects in bioinformatics is the Gene Ontology (GO) [2]. GO is a controlled and structured taxonomy designed mainly to describe the molecular functions, biological processes and cellular components of gene products independent of the organisms. The gene information terms in GO are presented in a structured format to make the study and comparison of gene properties easier. Gene Ontology is a Directed Acyclic Graph (DAG) in which terms may have multiple parents and thus two GO nodes can have multiple different paths between them. Computing the similarities between genes is an important and necessary task in bioinformatics [1, 3, 13, 16]. For example, comparing similarities between genes with known molecular functions with those with unknown functions would reveal the functions of the unknown genes to certain accuracy [13]. Gene Ontology annotations capture the available functional information of gene products, in an organism, and can be used as a basis for defining a measure of similarities between genes and gene products [13, 15].

In this paper, we propose a method for measuring the similarity between genes using the GO annotations terms of these genes. The proposed method measures semantic similarity of genes based on path length between their GO terms in the GO graph. To evaluate the method, we measured the semantic similarity of *yeast* genes (from SGD database http://www.yeastgenome.org) for various SGD pathways and compared our results with two of the leading measures (Resnik [11] and Wang et al. [15]). Our method showed impressive accuracy with results better than [11] and with very high agreement and competitive with [15]. The contribution of this paper is a simple yet elegant method with a competitive performance which gives it great appeal in the GO related research. Gene annotation data are represented in scientific natural language which is easier to be modeled and is more readable to human as compared to other bioinformatics data that exist, for example, in the form of sequences. The GO project is collaboration between 35 model organism databases; among them FlyBase (Drosophila melanogaster), SGD (Saccharomyces Genome Database) and MGD (Mouse Genome Database) were the first group of databases that started the collaboration and after that other databases have joined them.

2. Related Work

Ontology-based semantic similarity measures have been investigated for long time in the general English domain. For example, Resnik [11], Jiang and Conrath [5] and Lin [6] proposed information-content (IC) based measures for semantic similarity between terms, and these measures were designed mainly for WordNet [8]. These measures are proven to be useful in natural language processing (NLP) tasks [1, 3, 9]. Resnik's measure calculates the semantic similarity between two terms [t_1, t_2] in a given ontology (*e.g.,* WordNet) as the information content (IC) of the least common ancestor (*LCA*) of t_1, t_2. The IC of a term t can be quantified in terms of the likelihood (probability) of its occurrence $p(t)$. The probability assigned to a term is defined as its relative frequency of occurrence. Jiang and Conrath [5] proposed a different approach by combining the edge based measure with information content calculation of node based techniques. Lin [6] in 1998 developed a measure that considered how close the terms are to their least common ancestor (LCA) in the ontology. However, it disregards the level of detail of the lowest common ancestor.

In the Biomedical domain, Rada et al. [10] proposed the first semantic similarity measure in the biomedical domain by using path length between biomedical terms in the MeSH ontology (Medical Subject Heading www.nlm.nih.gov/mesh/) as a measure of semantic similarity. Several other biomedical ontologies, within the framework of UMLS (Unified Medical Language System http://www.nlm.nih.gov/research/umls/), have also been used for measuring semantic similarity in bioinformatics [1], *e.g.* Snomed-CT (www.nlm.nih.gov/snomed/) and ICD9CM (http://icd9cm.chrisendres.com/).

Lord et al. (2003) [7] were the first to apply a measure of semantic similarity to GO. They proposed a technique for calculating the semantic similarity of protein pairs based on Resnik's measure [11]. The semantic similarity between two proteins is defined as the average similarity of all GO terms with which these proteins are annotated. Speer et al. (2004) [14] used a distance measure based on Lin's similarity for clustering genes on a microarray according to their function. Chang et al. (2001) and MacCallum et al. (2000) [4] showed that similarity between annotation and literature will augment sequence similarity searches [9]. Sevilla et al. (2005) [12] analyzed the correlation between gene expression and Resnik's, Jiang and Conraths' and Lin's measures of semantic similarity [11, 5, 6]. They concluded that Resnik's measure correlates well with gene expression. More recently, Schlicker et al. (2006) [13] introduced an information content (IC) based measure for measuring the similarity between GO terms in Gene Ontology. It is based on a combination of Lin's and Resnik's techniques. Their result shows that those proteins with the highest sequence similarities tend to have similar molecular functions. However there are lots of cases that the functional similarity is not correlated (directly proportional) with the sequence similarity. Wang et al. (2007) [15] proposed a measure to calculate the similarity of GO terms based on term's semantics (*S value*) which is an aggregate of the contributions of the term's ancestors in the GO graph. In the evaluation, they found that their method produces results closer to human perception when compared with the results of Resnik's measure on the same genes [15].

3. The Proposed Measure

The length of the shortest path (PL) between two terms in a given ontology has been proved to be a good indicator of the semantic distance (*semantic distance is the inverse of semantic similarity*) between them [3, 10]. GO is considered the most comprehensive resource for gene functional information. The PL has not been extensively investigated in GO as a potential measure of similarity between GO terms leading to a similarity measure between genes. In our method, we compute path length (PL) between GO terms (Eq.1) and between genes (Eq.2). Then we measure the similarity between two genes by using a transfer function for mapping the PL distance into similarity value (Eq.3). We define the path length function between two GO terms go_x and go_y as follows:

$$\mathrm{PL}(go_x, go_y) = \textit{the minimum path length in the GO graph between the two GO terms } go_x \textit{ and } go_y \quad \ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Notice that the minimum PL has to go through the LCA; that is, we do not count the paths that pass via the greatest common descendant.

3.1. Path Length between Genes

Given two genes G_p and G_q such that gene G_p is annotated with a set of *n* different GO terms, we call it the set GO_p: $GO_p = \{go_p^1, go_p^2, \ldots., go_p^n\}$, and similarly, the annotation set for gene $G_q = GO_q = \{go_q^1, go_q^2, \ldots., go_q^m\}$; that is, gene G_q is annotated with *m* distinct GO terms. From these two sets, GO_p and GO_q, we compute an $n \ x \ m$ matrix of *PL* values between GO term pairs $PL(go_p^i\ , go_q^j)$ *for all* i = 1, .., n and j = 1, …, m. Then we calculate the average of all *PL* values in the matrix which will be the *PL* for the two genes, that is:

$$PL\ (\mathrm{G_p}, \mathrm{G_q}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} PL\ \ (go_p^i, go_q^j)}{n \times m} \ldots\ldots\ldots(2)$$

Example: Consider the following example from SGD: The two genes ABF1 and IFH1 are annotated with the following Go-terms:

GO_{ABF1} ={3682, 8301, 3677, 3688, 16563, 16564}
GO_{IFH1} = {3700, 3704}

The 6x2 matrix containing the pair-wise path length (PL) is shown in Table 1. Using these values, the (average) *PL* between IFH1 and ABF1 is computed as follows:

PL(IFH1, ABF1) =

$$\frac{4+5+2+7+1+6+3+8+6+5+6+5}{6x2} = 4.833$$

3.2. Similarity between Genes

We derive the similarity between two genes as an exponent function of the negated average path length between their GO terms. Li et al. (2003) [16] proposed using exponent function for transferring path length into similarity value using ontology. They applied and tested their method using WordNet 1.6 ontology in the general English domain [16]. We propose the similarity between G_p and G_q as follows:

$$sim(Gp, Gq) = e^{-f*PL(Gp,Gq)} \ldots\ldots\ldots(3)$$

where f is a factor for tuning the contribution of the PL into the similarity function, *sim()*, between the two genes. This transfer function converts the PL into similarity value such that the similarity is a monotonically decreasing function of the path length.

		IFH1	
		GO:0003700	GO:0003704
ABF1	GO:0003682	4	5
	GO:0008301	2	7
	GO:0003677	1	6
	GO:0003688	3	8
	GO:0016563	6	5
	GO:0016564	6	5

Table 1. PL (path length) values between GO terms of two SGD genes (ABF1 and IFH1).

The function ensures that the similarity is maximum when path length is zero (Eq .3); that is; when the two genes are annotated with the same GO function term. The function, moreover, guarantees the sim() value to range between 0 and 1. The similarity is thus a decreasing function of the path length. In our experiments, we tested with parameter *f* values between 0.10 and 0.50.

4. Experimental Results and Evaluation

In general, there are few techniques for evaluating the accuracy of a given similarity measure. In NLP, for example, the two common approaches for evaluating the computed semantic similarity values of a given measure is (a) by computing correlation coefficient with human scores using a dataset of term pairs scored for similarity by human evaluators; (b) by using the measure in an application like information retrieval (IR) system or text categorization [3]. In the scope of this paper, i.e., within the context of functional similarity of genes using GO annotations, the evaluation methodologies include: –comparing the similarity values computed by the measure with gene sequence similarity [1, 3, 5, 13], –comparing with gene expression profiles [12], or –using gene pathways and clusters information to validate the results [15]. In this paper we followed the third approach, and we compare our method with two of the best performing measures: Resnik (we refer to it as *M-I* in this paper) [11] and Wang et al. (we call it *M-II*) [15]. As what Sevilla et al. (2005) [12] found from the analysis of the correlation between gene expression and other IC based measures (Resnik, 1995; Jiang and Conrath, 1997; Lin 1998) [5, 6, 11], Resnik's measure turned out to be more accurate than the others. We used the SGD (*yeast*) database (www.yeastgenome.org) in the evaluation. We used on the GO annotation terms of MF (molecular function) ontology from SGD database. We analyzed the results for the pathways retrieved from http://pathway.yeastgenome.org/. Like in [15], we analyzed all the pathways containing 3 or more genes and compared our results with M-I and M-II [11, 15]. The results of our proposed measure were quite impressive and competitive. In the rest of this evaluation, we report and discuss few example pathways: pathways #5: *allantoin degradation* and #6: *arginine biosynthesis,* containing 4 and 7 genes respectively (*the first four pathways contain 3 or less genes*); pathway #54: *glycolysis* (14 genes); and pathway # 93: *phospholipid biosynthesis* (8 genes). We chose these pathways with various numbers of genes as examples to discuss our experiments and evaluation. Wang et al. (2007), in [15], conducted a comprehensive analysis of their measure versus Resnik's measure for SGD pathways having 3 or more genes. They concluded that their measure performs similar or better than Resnik's in all tested SGD pathways. Initially, we ran our measure on large number of gene sets from SGD and compared our results with M-II while varying the value of parameter *f* in Equation (3); Table 2 shows some of the results. We chose *f*=0.20 as it produces the best performance according to our evaluation. The result in Table 2 shows the correlation coefficient (agreement) between our method and M-II. These results demonstrate that our measure produces extremely similar results to measure M-II even though our method is much simpler than M-II [15]. The similarity values among the gene pairs of pathways 5 & 6 are shown in Table 3 for our proposed measure, M-I, and M-II. We notice from this table that our measure is extremely well correlated with the other two measures (Table 3). This also is shown in Table 2 as well; our measure (with *f=0.2*) has correlation coefficients of 0.998 and 0.985 for pathways 5 and 6 respectively. The similarity results of our measure along with M-I and M-II using genes in pathways 54 and 93are shown in Tables 4 and 5 respectively. These Results (Tables 2 – 5) indicate that our measure, with its simplicity, is competitive and compares favorably with M-I and M-II. For example, in pathway 5, Table 3, our measure gave the gene pair {DAL2, DAL3} the highest similarity (0.67) whereas the 3 pairs {DAL1, DUR1,2}, {DAL2, DUR1,2}{DAL3, DUR1,2} received the lowest similarity; and this is in full agreement with both M-I and M-II.

Pathway#	Number of genes	*f* =								
		0.10	0.15	0.20	0.25	0.3	0.35	0.40	0.45	0.50
5	4	0.994	0.996	0.998	0.999	1.000	1.000	0.999	0.998	0.996
6	7	0.953	0.973	0.985	0.990	0.991	0.988	0.984	0.980	0.976
54	14	0.993	0.995	0.993	0.989	0.983	0.974	0.965	0.954	0.944
93	8	0.998	0.996	0.990	0.979	0.965	0.947	0.927	0.906	0.884
141	12	0.993	0.989	0.984	0.978	0.971	0.963	0.955	0.946	0.938

Table 2. Correlation values of our method with M-II [15] for a number of pathways from SGD for different values of *f*.

	Gene1	Gene2	M-I	M-II	Proposed
Pathway 5	DAL1	DAL2	2.469	0.512	0.449
	DAL1	DAL3	2.469	0.512	0.449
	DAL1	DUR1,2	1.740	0.419	0.333
	DAL2	DAL3	5.221	0.728	0.670
	DAL2	DUR1,2	1.740	0.419	0.333
	DAL3	DUR1,2	1.740	0.419	0.333
Pathway 6	ARG1	ARG2	0.281	0.155	0.135
	ARG1	ARG3	0.281	0.235	0.247
	ARG1	ARG4	0.281	0.235	0.247
	ARG1	ARG5,6	0.281	0.227	0.247
	ARG1	ARG8	0.281	0.235	0.247
	ARG1	ECM40	0.281	0.155	0.135
	ARG2	ARG3	1.378	0.218	0.165
	ARG2	ARG4	0.281	0.128	0.111
	ARG2	ARG5,6	1.013	0.176	0.135
	ARG2	ARG8	1.378	0.218	0.165
	ARG2	ECM40	5.755	0.932	0.819
	ARG3	ARG4	0.281	0.199	0.202
	ARG3	ARG5,6	1.013	0.270	0.247
	ARG3	ARG8	1.378	0.338	0.301
	ARG3	ECM40	1.378	0.218	0.165
	ARG4	ARG5,6	0.281	0.193	0.202
	ARG4	ARG8	0.281	0.199	0.202
	ARG4	ECM40	0.281	0.128	0.111
	ARG5,6	ARG8	1.013	0.270	0.247
	ARG5,6	ECM40	1.104	0.181	0.135
	ARG8	ECM40	1.378	0.218	0.165

Table 3. Comparison of similarity results of M-I, M-II, and proposed measure in two pathways from SGD.

	Gene 1	Gene 2	M-I	M-II	Proposed
Pathway 54	CDC19	ENO1	0.281	0.18	0.202
	CDC19	FBA1	0.281	0.18	0.202
	CDC19	PGK1	3.143	0.599	0.670
	CDC19	PYK2	3.394	1.000	1.000
	CDC19	TDH2	0.281	0.157	0.165
	CDC19	TDH3	0.281	0.157	0.165
	ENO1	ENO2	5.529	1.000	1.000
	PFK1	PFK2	7.826	1.000	1.000
	TDH1	TDH2	7.935	1	1.000
	TDH2	TDH3	7.935	1	1.000
	TDH3	TPI1	0.281	0.173	0.165

Table 4. Excerpts from similarity results of genes from pathway 54 *glycolysis* using M-I, M-II, and proposed measure.

In pathway 54, the two genes ENO1 and ENO2 are annotated with the same function *phosphopyruvate hydratase activity* (GO:0004634); our measure gives max similarity (1.0) for this pair (Table 4). The same applies for the gene pair CDC19 and PYK2 which share the same GO term *pyruvate kinase activity* (GO:0004743) and this pair is assigned the max similarity value of 1.0 (Table 4). On the other hand, M-I gives similarity value 3.39 for {CDC19, PYK2} and similarity 5.53 for the pair {ENO1, ENO2} and both are not max similarity values as the max similarity (7.826) is given to the gene pair {PFK1, PFK2}; see Table 4. In Table 5, we notice that the two genes PSD1 and PSD2 have similarity of 1.0 (max) and they share the same function *phosphatidylserine decarboxylase activity* (GO:0004609). Pathway #6 (Table 3) demonstrated some differences in the similarity values produced by our measure and M-I. For example, if we compare the two pairs (ARG2, ARG3) and (ARG3, ARG5,6) we see that M-I gives higher similarity value for (ARG2, ARG3) than for (ARG3, ARG5,6), however, in GO tree, the distance between the terms annotating (ARG2, ARG3) and (ARG3, ARG5,6) are 9 and 6 respectively. That is, ARG3 & ARG5,6 are semantically closer to each other. Our measure gave higher similarity (0.25) for (ARG3, ARG5,6) than for the other pair (0.16) which is more consistent with the annotations in the GO tree. To look at the performance of our measure from a different perspective, [15] suggests that we cluster the genes according to the similarity values computed by similarity measure, and then we can evaluate the measure by examining these clusters with human perspective with the help of gene functional pathways. We conducted this evaluation and clustered the genes according to our measure as well as according to measure M-I and the clustering results are shown in Figure 1 through Figure 5. What makes our method more attractive is that it is much simpler and easier to implement. It uses only one information source (GO) and does not uses node counts nor term frequencies/probabilities. We tested our measure against a fairly large-sized SGD pathways –the *glycolysis* pathway contains 91 gene pairs; the results reported in this paper include more than 210 gene pairs. We compared our results with two measures on all SGD pathways.

The result of pathway 54 (*glycolysis*) analysis is shown in Table 4 and the clustering results of this pathway are shown in Figure 3. Figure 3 shows that our measure clusters the genes PYK2 & CDC19 together in the first clustering step whereas M-I put them in the same cluster in the 5th clustering step (clustering illustration for M-I not shown). This proves that our measure is more accurate as both PYK2 and CDC19 are annotated with the same GO function (GO:0004743) as mentioned earlier.

In Figures 4 and 5 we notice that our measure clusters the two genes PDS1 & PDS2 before clustering CHO1 & PGS1 together while M-I clusters this latter pair earlier. This indicates that our method is more accurate if we know that PDS1 & PDS2 share the same function *phosphatidylserine decarboxylase activity* (GO:0004609). On the other hand, CHO1 is annotated with the GO term GO:0003882 (CDP-diacylglycerol-serine O-phosphatidyltransferase activity) and PGS1 with the GO term GO:0008444 (CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase activity) which both in turn descend from the common parent GO:00017169

	Gene1	Gene2	M-I	M-II	Proposed
Pathway 93	CDS1	CHO1	2.53	0.445	0.368
	CDS1	CHO2	1.378	0.266	0.247
	CDS1	CRD1	2.53	0.445	0.368
	CDS1	OPI3	1.378	0.266	0.247
	CDS1	PGS1	2.53	0.445	0.368
	CDS1	PSD1	0.281	0.199	0.202
	CDS1	PSD2	0.281	0.199	0.202
	CHO1	CHO2	1.378	0.229	0.202
	CHO1	CRD1	3.143	0.544	0.449
	CHO1	OPI3	1.378	0.229	0.202
	CHO1	PGS1	6.904	0.746	0.670
	CHO1	PSD1	0.281	0.173	0.165
	CHO1	PSD2	0.281	0.173	0.165
	CHO2	CRD1	1.378	0.229	0.202
	CHO2	OPI3	4.977	0.789	0.670
	CHO2	PGS1	1.378	0.229	0.202
	CHO2	PSD1	0.281	0.157	0.165
	CHO2	PSD2	0.281	0.157	0.165
	CRD1	OPI3	1.378	0.229	0.202
	CRD1	PGS1	3.143	0.544	0.449
	CRD1	PSD1	0.281	0.173	0.165
	CRD1	PSD2	0.281	0.173	0.165
	OPI3	PGS1	1.378	0.229	0.202
	OPI3	PSD1	0.281	0.157	0.165
	OPI3	PSD2	0.281	0.157	0.165
	PGS1	PSD1	0.281	0.173	0.165
	PGS1	PSD2	0.281	0.173	0.165
	PSD1	PSD2	5.987	1.000	1.000

Table 5. Comparison of results of M-I, M-II, and proposed measure for pathway 93 *phospholipid biosynthesis*

(CDP-alcohol phosphatidyltransferase activity). In pathway 54, according to our measure, PGK1 clusters with CDC19 and PYK2 in the second clustering step (Figure 3) whereas M-I does not cluster them until the 6th step. PGK1 and CDC19 are assigned the two GO functions *phosphoglycerate kinase activity* (GO:0004618) & *pyruvate kinase activity* (GO:0004743) that share the same parent *kinase activity* (GO:0016301); and this validates further the accuracy of the proposed measure.

5. Conclusion

We presented a simple measure for semantic similarity of GO terms and then the functional similarity of genes. The measure is based strictly on the ontology structure of GO. Specifically, our measure estimates the semantic similarity between two GO terms using only the path lengths between them. Then we map the path length between GO terms using an exponential function into similarity between genes. The strength of our measure comes from its simplicity yet with competitive and impressive performance compared with the existing measures. We examined our measure with a large number of gene groups from SGD (*yeast*) pathways. The experimental results showed that the proposed measure performs better than the measure of Resnik in most cases or equal in the rest of the cases, and very competitive or sometimes better than Wang et al.'s measure. Since our measure is based solely on GO structure, the outcome of this research validates the accuracy and correctness of the GO as a controlled and structured ontology of gene functions developed and maintained by human expert curators.

Threshold	Initial	0.819	0.301	0.247	0.202	0.111
Clustering Result	ARG8	ARG8	ARG8	ARG8	ARG8	ARG8
			ARG3	ARG3	ARG3	ARG3
	ARG3	ARG3		ARG5,6	ARG5,6	ARG5,6
			ARG5,6		ARG4	ARG4
	ARG5,6	ARG5,6		ARG4	ARG1	ARG1
			ARG4	ARG1		ECM40
	ARG4	ARG4			ECM40	ARG2
			ARG1		ARG2	
	ARG1	ARG1		ECM40		
			ECM40	ARG2		
	ECM40	ECM40	ARG2			
		ARG2				
	ARG2					

Figure 1. Clustering genes in pathway 6 *arginine biosynthesis* according to our measure.

Threshold	Initial	5.755	1.355	1.055	0.255
Clustering Result	ARG8	ARG8	ARG5,6		
					ARG4
	ARG3	ARG3	ARG4	ARG4	ARG1
					ARG5,6
	ARG5,6	ARG5,6	ARG1	ARG1	ARG3
					ARG8
	ARG4	ARG4			ECM40
				ARG5,6	ARG2
	ARG1	ARG1	ARG3	ARG3	
			ARG8	ARG8	
	ECM40	ECM40	ECM40	ECM40	
		ARG2	ARG2	ARG2	
	ARG2				

Figure 2. Clustering genes in pathway 6 *arginine biosynthesis* according to measure M-I.

6. Reference

[1] Al-Mubaid H. and Nguyen H.A.(2007) "Similarity Computation Using Multiple UMLS Ontologies in a Unified Framework." Proceedings of 22nd ACM Symposium on Applied Computing SAC'07.

[2] Ashburner M. et al. (2000). "Gene ontology: tool for the unification of biology." The Gene Ontology Consortium. Nat Genet. 2000;25:25–9.

[3] Caviedes J.E, Cimino J.J. Towards the development of a conceptual distance metric for the UMLS. Journal of Biomedical Informatics, vol. 37, 2004

[4] Chang J., Raychaudhuri S. and Altman R. (2001) "Including biological literature improves homology search." Pac. Symp. Biocomput., 6, 374–383.

[5] Jiang J.J, and Conrath D.W. (1997). Semantic similarity based on corpus statistics and lexical ontology. In Proc. on International Conference on Research in Computational Linguistics, 1997.

[6] Lin, D. (1998). "An information-theoretic definition of similarity." In Proc. of the Int'l Conference on Machine Learning.

[7] Lord P.W., Stevens R.D., Brass A., Goble C.A. (2003) "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation." Bioinformatics vol.19 (10):1275-83.

[8] Miller G.A. (1995). "WordNet: A Lexical Database for English," Comm. ACM, vol. 38, no. 11.

[9] Pedersen T. et al (2006), "Measures of Semantic Similarity and relatedness in the biomedical domain" Journal of Biomedical Informatics.

[10] Rada R, Mili H, Bicknell E, Blettner M. (1989) "Development and application of a metric on semantic nets." IEEE transactions on systems, man and cybernetics, 1989;19(1): p. 17–30.

[11] Resnik, P. (1995). "Using Information Content to Evaluate Semantic Similarity in a Taxonomy." Proc 14th Int'l Joint Conf Artificial Intelligence.

[12] Sevilla J. L. et. al (2005). "Correlation between Gene Expression and GO Semantic Similarity" IEEE/ACM Transaction on computational biology and bioinformatics, vol.2, No. 4.

Threshold	Initial	1	0.67	0.449	0.30	0.20	0.165	0.135
Clustering Result	TPI1	TPI1	TPI1	TPI1	TPI1	TPI1	TPI1	TPI1
			PGI1	PGI1	PGI1	PGI1	PGI1	PGI1
	PGI1	PGI1			GPM1	GPM1	GPM1	GPM1
			GPM1	GPM1		ENO2	ENO2	ENO2
	GPM1	GPM1			ENO2	ENO1	ENO1	ENO1
			ENO2	ENO2	ENO1	FBA1	FBA1	FBA1
	ENO2	ENO2	ENO1	ENO1	FBA1		PFK2	PFK2
		ENO1				PFK2	PFK1	PFK1
	ENO1		FBA1	FBA1	PFK2	PFK1	PYK2	PYK2
		FBA1			PFK1	PYK2	CDC19	CDC19
	FBA1		PFK2	PFK2	PYK2	CDC19	PGK1	PGK1
		PFK2	PFK1	PFK1	CDC19	PGK1		TDH3
	PFK2	PFK1		PYK2	PGK1		TDH3	TDH2
			PYK2	CDC19		TDH3	TDH2	TDH1
	PFK1	PYK2	CDC19	PGK1	TDH3	TDH2	TDH1	
		CDC19	PGK1		TDH2	TDH1		
	PYK2			TDH3	TDH1			
		PGK1		TDH2				
	CDC19		TDH3	TDH1				
		TDH3	TDH2					
	PGK1	TDH2	TDH1					
		TDH1						
	TDH3							
	TDH2							
	TDH1							

Figure 3. Clustering genes in pathway 54 *glycolysis* according to our measure.

[13] Schlicker A. et al. (2006). "A new measure for functional similarity of gene products based on Gene Ontology." BMC Bioinformatics.

[14] Speer N, Spieth C., and Zell A. A. Memetic (2004) "Clustering Algorithm for the Functional Partition of Genes Based on the Gene Ontology." IEEE Symp on Computational Intelligence in Bioinformatics & Computational Biology CIBCB'2004.

[15] Wang J.Z. et al. (2007). "A new method to measure the semantic similarity of GO terms." Bioinformatics, 2007.

[16] Li Y., Bandar Z.A, and McLean D. "An Approach of measuring semantic similarity between words using multiple information sources". IEEE Trans. Knowledge and Data Engineering. Vol.15, no.4, 2003.

Threshold	Initial	1	0.67	0.449	0.368	0.20	0.16
Clustering Result	PGS1	PGS1	PGS1	PGS1	PGS1	PGS1	PGS1
			CHO1	CHO1	CHO1	CHO1	CHO1
	CHO1	CHO1		CRD1	CRD1	CRD1	CRD1
			CRD1		CDS1	CDS1	CDS1
	CRD1	CRD1		CDS1		OPI3	OPI3
			CDS1		OPI3	CHO2	CHO2
	CDS1	CDS1		OPI3	CHO2		PSD2
			OPI3	CHO2		PSD2	PSD1
	OPI3	OPI3	CHO2		PSD2	PSD1	
				PSD2	PSD1		
	CHO2	CHO2	PSD2	PSD1			
			PSD1				
	PSD2	PSD2					
		PSD1					
	PSD1						

Figure 4. Clustering genes in pathway 93 *phospholipid biosynthesis* according to our measure.

Threshold	Initial	6.90	5.90	4.90	3.10	2.50	1.30	0.20
Clustering Result	PGS1	PGS1	PGS1	PGS1	PGS1	PGS1	PGS1	PGS1
		CHO1	CHO1	CHO1	CHO1	CHO1	CHO1	CHO1
	CHO1				CRD1	CRD1	CRD1	CRD1
		CRD1	CRD1	CRD1		CDS1	CDS1	CDS1
	CRD1				CDS1		OPI3	OPI3
		CDS1	CDS1	CDS1		OPI3	CHO2	CHO2
	CDS1				OPI3	CHO2		PSD2
		OPI3	OPI3	OPI3	CHO2		PSD2	PSD1
	OPI3			CHO2		PSD2	PSD1	
		CHO2	CHO2		PSD2	PSD1		
	CHO2			PSD2	PSD1			
		PSD2	PSD2	PSD1				
	PSD2		PSD1					
		PSD1						
	PSD1							

Figure 5. Clustering genes in pathway 93 *phospholipid biosynthesis* according to measure M-I.