A Path Length Based Method for Finding Functional Gene Similarity Using Gene Ontology

Elham Khabiri
University of Houston - Clear Lake. Houston, TX 77058, USA.

Abstract

The size and volumes of genomic data resulting from the various genome projects are extremely huge and increasing in very high rates. Finding gene groups with similar functions is one of the most important tasks in bioinformatics. In this paper, we present a novel technique for estimating gene functional similarity using Gene Ontology (GO) annotations. GO is considered the most comprehensive resource for functional gene annotation. The proposed method is an ontology-structure-based method and relies only on path length (PL) between GO nodes. We evaluated the method based on the correlation between PL and gene sequence similarity using *Blast e-values.* We conducted experiments with two genome annotation databases: *SGD* and *Flybase* using molecular function (MF) terms. The experimental results proved that the method has fairly impressive agreement with Blast sequence similarity. Furthermore, the evaluations showed that PL can be used as a tool for determining the genes with similar functions within a genome.

1. Introduction

Computing the functional similarity between genes is an important and necessary task in bioinformatics. Comparing similarities between genes with known molecular functions with those with unknown ones would reveal the functions of the unknown genes to certain accuracy [2].Although the sequence similarity, in general, holds for most genes and proteins with the same molecular function, there are genes that are not evolved from a common ancestor and therefore the sequence similarity between them are not considerable, but still they have the similar molecular functions.

There are limitations on using the sequence similarity measures. Previous studies showed that up to 30% of the function annotations made through sequence similarity searches might be erroneous [17][2]. One of the greatest projects which have been done in this domain is Gene Ontology (GO) [1, 9]. GO is a controlled and structured vocabulary and taxonomy that is designed mainly to describe the molecular functions, biological process and cellular components of gene products independent of the organisms. The gene functional related terms in GO are presented in a controlled vocabulary format that makes the comparison of the genes easier. Gene Ontology is a Directed Acyclic Graph (DAG) in which terms may have multiple parents and thus, two nodes can have multiple different paths between them.

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 functional similarity between gene products [2]. Gene annotation data is 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), the SGD (Saccharomyces Genome Database) and the Mouse Genome Database (MGD) were the first group of databases that started the collaboration and after that other databases have joined them [1]. In this project each gene is annotated with one or more terms and saved in the annotation file of the related organism.

In this paper, we propose a novel method for measuring the functional similarity between genes using the GO annotations.

The method is based on calculating the average path length (PL) between GO annotation terms of the genes. We evaluated the method with a series of experiments based on the correlation between PL and gene sequence similarity using *Blast e-values.* The experimental results proved that the method has fairly impressive agreement with Blast sequence similarity. Furthermore, the evaluations showed that PL can be used as a tool for determining the genes with similar functions within a genome. We used in the evaluation two genome annotation datasets: SGD and Flybase [24, 33, 1, 21]. Each dataset is divided into a number of sequence similarity ranges based on the E-value in gene pairs. Then, we grouped the genes into genes with high sequence similarity

(HSS), low sequence similarity (LSS) and no sequence similarity (NSS) and each one of these three groups was tested separately.

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 [12] and Lin [13] proposed information-content (IC) based measures for semantic similarity between terms and these measures were designed mainly for WordNet [28]. WordNet is a freely available lexical database that represents an ontology of approximately 100,000 general English concepts. These measures are proven to be useful in natural language processing (NLP) tasks [10]. Resnik's measure calculates the semantic similarity between two terms [t_1, t_2] in Ontology (*e.g.,* WordNet) as the information content (IC) of the least common subsumer (*LCS*) 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 higher a term appears in the ontology means the lower is its information content because, simply, more general terms tend to occur more frequently in general than specialized terms. The probability assigned to a term is defined as its relative frequency of occurrence. Jiang and Conrath [12] proposed a different approach by combining the edge based measure with information content calculation of node based techniques. Lin [14] in 1998 developed a measure that considered how close the terms are to their least common subsumer (LCS) in the ontology. However, it disregards the level of detail of the lowest common ancestor.

In the Biomedical domain, measures of semantic similarity based on ontology were developed as early as 1989. Reda et al. [16] proposed the first semantic similarity measure in the biomedical domain by using path length between biomedical terms in the MeSH ontology [29] as a measure of semantic similarity. Path Length (PL) can be calculated easily for the tree structured Ontologies such as WordNet. But for DAG-type ontologies, like Gene Ontology, path length is more complicated, since each node may have multiple parents, and thus, two nodes can have several different paths between them. Several other biomedical ontologies, within the framework of UMLS (unified medical language system) [30], have also been used for measuring semantic similarity in bioinformatics. These include Snomed-ct [31] and ICD9CM [32].

Lord et al. (2003) [15] 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. Each protein pair receives three similarity values, one for each Ontology (Molecular Function, Biological Process and Cellular Component Ontologies) [15]. Speer et al. (2004) [19] used a distance measure based on Lin's similarity for clustering genes on a microarray according to their function. Chang et al. (2001) [34] and MacCallum et al. (2000) [35] showed that Similarity between annotation and literature will augment sequence similarity searches [15]. They improved PSIBLAST (Altschul *et al.*, 1997 [26]) with similarity scores calculated over the annotations and Medline [27] references. Sevilla et al. (2005) [18] analyzed the correlation between gene expression and Resnik's, Jiang and Conraths' and Lin's measures of semantic similarity [18]. They used microarray data analysis to determine expression levels of genes and compare them with those annotated in GO. They concluded that Resnik's measure correlates well with gene expression. More recently, Schlicker et al. (2006) [2] introduced a new measure of similarity between GO terms in Gene Ontology that is based on Lin's and Resnik's techniques. Their measure (sim_{Rel}) takes into account how close terms are to their least common subsumer as well as how detailed the LCA is, i.e., distinguishes between generic and specific terms. This sim_{Rel} score is the basis for a new measure, called *funSim*, to compute the functional relationship between two gene products. The score ranges from 0 to 1. A *funSim* score close to one indicates high functional similarity whereas a score close to zero indicates low similarity. The distribution of the *funSim* score analyzed and compared for four different categories of protein pairs corresponding to four levels of evolutionary relationship: no sequence similarity (NSS), low sequence similarity (LSS), high sequence similarity (HSS), and orthology[1] according to Inparanoid (IO) that have more sequences similarity than HSS. The result is that almost 60% of the protein pairs in the IO dataset

[1] Orthologs are genes in different species that originate from a single gene in the last common ancestor of these species. Such genes have often retained identical biological roles in the present-day organism [20].

have the score above 0.8. Those proteins with the highest sequence similarities tend to have similar molecular functions. However, some protein pairs in the IO set have scores below 0.2, indicating no functional similarity. The percentage of proteins with high functional similarity is highest for the IO category, and decreases for HSS and LSS, to almost no protein pairs without sequence similarity (NSS). These results confirm that functionally related proteins tend to have higher sequence similarity [2]. Although Path length measure has been applied and explored with several biomedical ontologies [16, 10] for biomedical term similarity, it has never been applied or investigated with the gene ontology. All gene functional similarity techniques that use GO are, thus far, based on IC of terms or node depth features [2, 12, 15, 16].

3. The Proposed Methods

Our proposed method is based on estimating the gene functional similarity based on calculating the semantic similarity between the GO terms annotated for genes. That is, we use the ontology structure (of GO) for estimating the similarity between pairs of genes based on their annotated terms. More specifically, we propose the path length between two terms in GO as an indicator of functional relatedness of the genes annotated with these terms. For example, suppose that two genes g_1 and g_2 are annotated with the GO terms t_1 and t_2, respectively, for their molecular functions. Then, the shortest path length between t_1 and t_2, $PL(t_1, t_2)$, in GO is a good measure of the functional similarity between g_1 and g_2.

3.1 Path Length Calculation

We developed an application for calculating the *shortest* path length between two genes (gene pair) based on their annotated terms. The method selects the gene pairs from an organism annotation file (*e.g.* SGD), then extracts the GO MF terms related to each gene and stores them in a link list; see Figure 2. Then it calculates the first common subsumer of the two genes. . We used the February 2007 release of GO from the gene ontology website [21]. The yeast gene annotations were downloaded from the SGD site (Dec.2006) [33], Flybase gene annotations were obtained from the GO website (Dec.2006) [1, 21]. – For each pair of genes $\{g_1, g_2\}$ in the annotation file, the terms related to each gene are extracted from the database.

– The path length between the GO terms are calculated from the GO DAG using edge counting.

– For the pair of genes $\{g_1, g_2\}$ such that g_1 is annotated (for its MF) with the terms $t_1, .., t_n$ while g_2 is annotated with terms $t_1,..,t_m$. We calculate all the possible short paths between the MF terms of g_1 and g_2. Let d_{ij} be the shortest path length between term t_i of g_1 and term t_j of g_2. The method computes the average of all paths:

$$avg\{ d_{ij} | i:1..n, j:1..m\}.$$

There were two ways for implementing our algorithm for calculating the shortest path length between two GO nodes: the mentioned algorithm:

a. Recording all the ancestors of each node till it reaches to the root and then comparing all paths to come up to the common ancestor.
b. Recording just the first level ancestors of each node and comparing them to see if they have anything in common or not.

The second approach uses less memory and also would be done in less time compared to the first approach.

3.2 Sequence Similarity

We used Blast tool [23] for computing sequence similarity between gene pairs. The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares gene sequences to sequence databases and calculates the statistical significance of matches. [25]

In some experiments, we used another tool, WU-BLAST2 [24], to find genes having high sequence similarity to a given gene. We changed the settings in this program so that more genes with less sequence similarities are shown in the result. Lower EXPECT thresholds in Blast settings causes more stringent selection that lessen the chance of matching sequences [25].

3.2.1 E-value

The Expect value (E-value) is a parameter that describes the number of hits one can "expect" to see just by chance when searching a database of a particular size [25]. In the gene sequence similarity results from Blast, the E-value of 0 means that the genes are totally similar, and as the E-value increases the sequence similarity decreases. This means that the lower the E-value, or the closer to 0 the more sequence similarity they have [25]. Bit-score is another metric of sequence similarity that

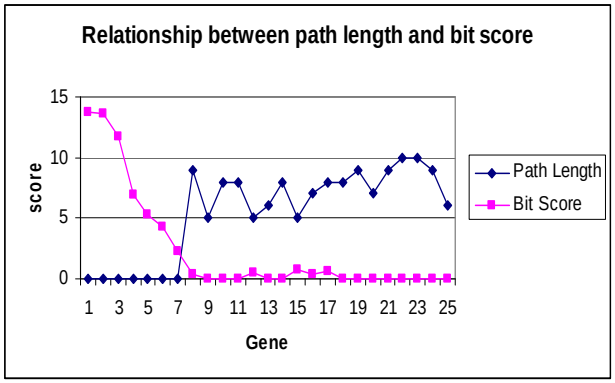

Figure 1. Relationship between path length and bit score

BLAST gives and that indicates how much alignment and sequence similarity two genes have. The higher the bit-score the better the alignment, and hence, higher sequence similarity.
The path length between two genes is inversely proportional with the bit score. When the path length between two genes increases, their Blast bit score decreases; this relation is shown in Figure 1.

4. Experiments and Results

To evaluate our method, we used three datasets of genes extracted from SGD (Saccharomyces cerevisiae) and one dataset from Flybase (Drosophila melanogaster) [22, 34]. – Firstly, we wanted to explore the distribution of path length between gene pairs in SGD genes. For that, 1000 gene pairs were selected randomly from SGD. The distribution of path length of these randomly selected gene pairs are shown in Figure 2. From this experiment (Figure 2) we notice that the majority of these gene pairs (64%) have path length between 3 and 7. Furthermore, 12% of these pairs have path length of at most 2 which indicate that these genes have somewhat significant semantic similarity (small path length) between their GO terms. Moreover, we found that 24% of these gene pairs have path length of 8 or greater which indicates that these pairs have no similarity in their GO annotation terms. This leads to the observation that there is no significant pattern or relation (by chance) of the path length feature between these SGD genes.
– Similarity we collected randomly 500 gene pairs from Flybase to examine the path length. The path length distribution is illustrated in Figure 3. Again, no pattern or relation exists between Flybase genes.

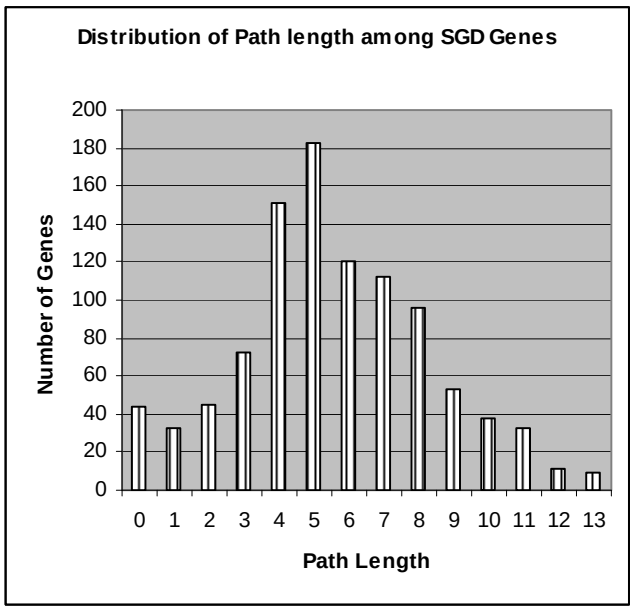

Figure 2. Distribution of path length among 1000 gene pairs randomly selected from SGD.

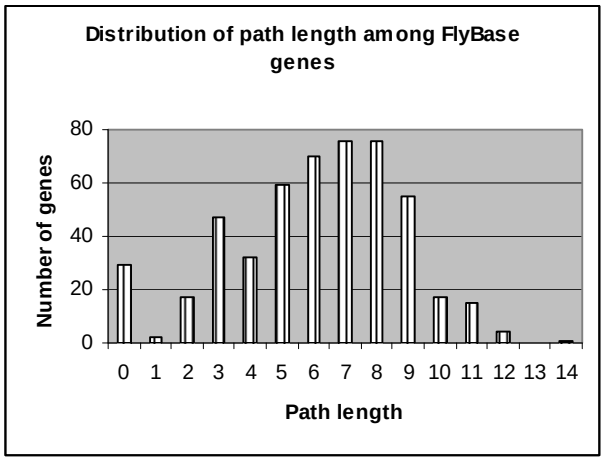

Figure 3. Distribution of path length among 500 gene pairs randomly selected from Flybase.

– Next, we examined our method to test the correlation between path length and sequence similarity of gene pairs. For that, we extracted three datasets of gene pairs from SGD: HSS, LSS, NSS. The high sequence similarity (HSS) gene pairs are those with the Blast E-value $\leq 10^{-5}$. The gene pairs with low sequence similarity (LSS) are those with the E-value $> 10^{-5}$ but less than one. The gene pairs with no sequence similarity (NSS) are those with the E-value = 1. Figure 4 shows a small part of the result for the HSS dataset.

Gene1	Gene2	PL	BitScore	EValue
ADP1	ADP1	0	4946	0
AFG2	AFG2	0	3805	0
AFT1	AFT1	0	3190	0
PMC1	SPF1	1	132	1.00E-12
ALD3	ALD4	2	1013	6.10E-104
ACS1	ACS1	0	3616	0
ACS2	ACS2	0	3614	0
PMC1	PCA1	3	168	3.90E-18
AAC3	SFC1	3	145	1.80E-09

Figure 4. Example from SGD gene pairs.

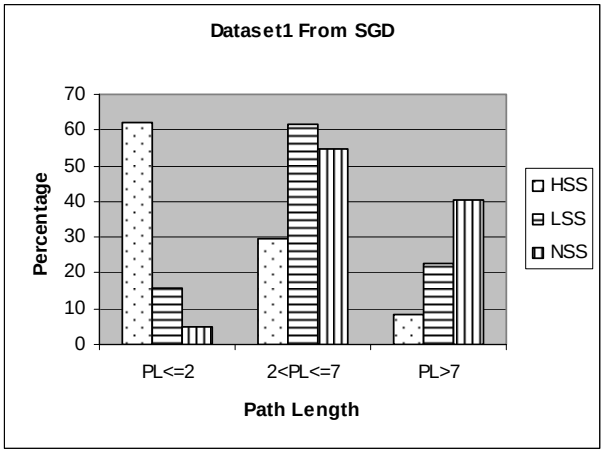

Figure 5. Distribution of path length between gene pairs in Dataset 1

Dataset 1 includes 200 gene pairs of HSS, 200 gene pairs of LSS, and 200 gene pairs of NSS extracted from SGD annotation file. Figure 5 illustrates the distribution of path length (*x-axis*) in HSS, LSS, and NSS sets. More than 60% of the gene pairs in HSS have path length of 2 or less while only 15% of LSS and 4% of NSS gene pairs have the path length 2 or less. The number of HSS gene pairs decreases as the path length increases through the x axis. We also found that more than 40% or NSS gene pairs and only less than 10% of HSS pairs have path length of 8 or more.

– We conducted another experiment on SGD genes using another dataset (Dataset2) of gene pairs having certain relations in their sequence similarity. Dataset 2 includes 139 gene pairs of HSS, 469 gene pairs of LSS, and 386 gene pairs of NSS extracted from SGD annotation. The results are illustrated in Figure 6. As we can see in these experimental results, again there is a pattern or relation between path length and sequence similarity. That is, gene pairs with high sequence similarity (HSS) tend to have low path length between their GO annotation

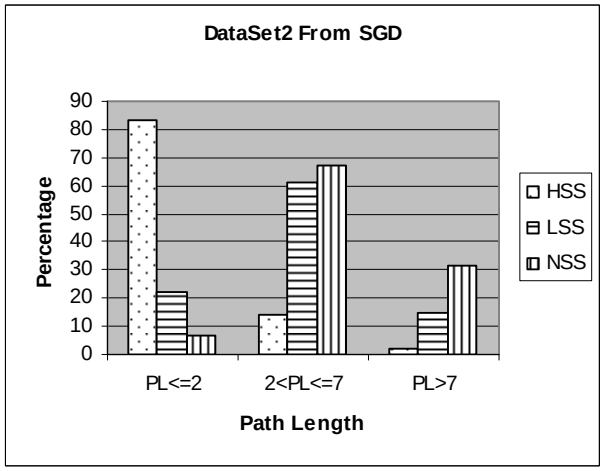

Figure 6. Distribution of path length between gene pairs in Dataset 2 from SGD

terms (more than 80% of HSS pairs have path length of 2 or less) whereas genes with no sequence similarity (NSS) lean to have relatively higher path length between their GO terms.

Next, we combined Dataset 1 and Dataset 2; we call it Dataset 3 which includes 339 HSS gene pairs, 669 LSS gene pairs, and 586 NSS gene pairs. The results of Dataset 3 are shown in Figure 7. Again, we have the same trend, majority of NSS genes (93%) have path length of 3 or more which implies that there is no significant semantic similarity in their GO terms. On the other hand, majority of HSS genes (70%) have path length of 2 or less indicating semantic similarity in their GO annotation terms.

– In another evaluation, we used genes from a different genome (Flybase) in a new dataset (Dataset_4) of gene pairs. Dataset 4 includes 60 gene pairs of HSS, 60 gene pairs of NSS extracted from FlyBase annotation database. The results of path length distribution among the Flybase gene pairs are illustrated in Figure 8. Almost 80% of HSS pairs have path length ≤ 2 while only 13% of NSS pairs have path length ≤ 2 which implies that there is a correlation between sequence similarity and path length in this dataset.

In summary, our evaluation experiments involved more than 1700 gene pairs (more than 3400 genes) having high, low, or no sequence similarity from two different organisms. Furthermore, we tested our method on 1500 gene pairs (3000 genes) randomly selected (with no particular sequence similarity) from the two organisms. All the experimental results on various gene groups, from two different genomes, support the fact that there is significant

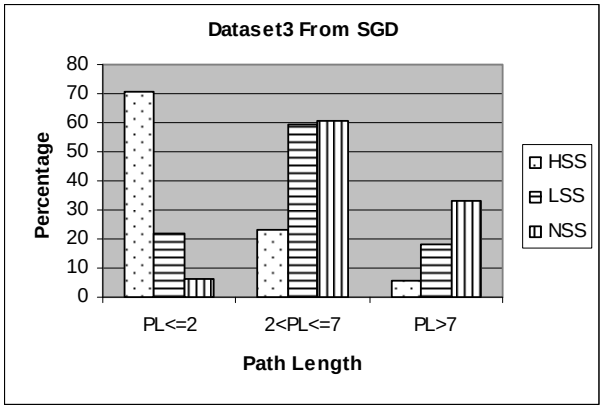

Figure 7. Distribution of path length between gene pairs in Dataset 3.

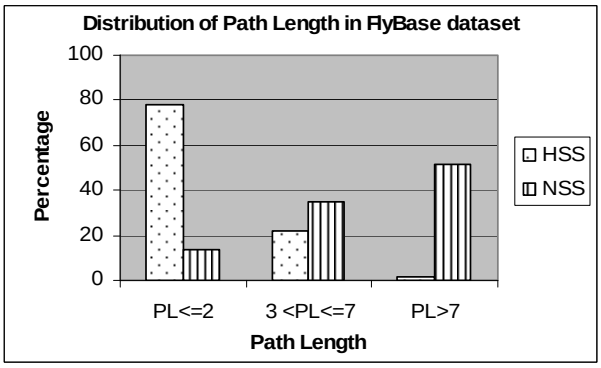

Figure 8. Distribution of path length between gene pairs in Dataset 4 from FlyBase

correlation between the sequence similarity of genes and semantic similarity using path length. This suggests and proves that path length between gene annotation terms using GO can be a good and reliable measure and metric for gene functional similarity.

5. Conclusion

Gene Ontology is considered the most comprehensive and reliable resource for functional annotations of gene products. The existing techniques for finding gene functional similarity based on GO rely mainly on IC or node depth. Path length feature has never been explored as a metric or indicator for gene functional similarities. The work presented in this paper is an attempt to fill this gap. We presented a novel technique for finding gene functional similarity based on GO annotation terms. The method is based on the average shortest path length between the GO terms annotated for both genes in a given gene pair. We evaluated the proposed method with a series of experiments on large groups of genes from two genomes SGD and Flybase. We have shown that this method correlates very well with gene sequence similarity by comparing large numbers of gene pairs with sequence similarities computed by one the most reliable algorithms for that purpose (Blast). We have shown further that randomly selected gene pairs have no significant (by-chance) pattern with path length.

Reference

[1] Gene Ontology: http://www.geneontology.org/GO.doc.shtml

[2] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenfuhrer and Thomas Lengauer (2006). "A new measure for functional similarity of gene products based on Gene Ontology." BMC Bioinformatics.

[3] Nguyen H., Al-Mubaid H. New Semantic Similarity Techniques of Concepts applied in the biomedical domain and WordNet. MS Thesis, University of Houston Clear Lake, Houston, TTX USA, 2006.

[4] H. Al-Mubaid and H.A. Nguyen. “Similarity Computation Using Multiple UMLS Ontologies in a Unified Framework.” Proceedings for the 22nd ACM Symposium on Applied Computing SAC’07, 2007.

[5] H.A. Neugyn and H. Al-Mubaid. “New Ontology-based Semantic Similarity Measure for the Biomedical Domain." IEEE conference on Granular Computing GrC-2006.pp. 623-628, 2006.

[6] H. A. Nguyen and H. Al-Mubaid. A Combination-Based Semantic Similarity Measure Using Multiple Information Sources. The 2006 IEEE Int’l Conference on Information Reuse and Integration IRI’06. Hawaii, USA, 2006.

[7] Devos D & Valencia A. (2000) "Practical limits of function prediction." PROTEINS, Structure, Function, and Genetics 41, 98-107..

[8] Devos D, Valencia A (2001), “Intrinsic errors in genome annotation.” Trends Genet.

[9] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, AHarris M, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000), “Gene ontology: tool for the unification of biology". The Gene Ontology Consortium. Nat Genet.

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

[11] Resnik, P. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. Proc 14th Int'l Joint Conf Artificial Intelligence. 1995. pp. 448–453.

[12] Jiang, J.J, and Conrath, D.W. Semantic similarity based on corpus statistics and lexical ontology. In Proc. on International Conference on Research in Computational Linguistics, 19–33, 1997.

[13] Lin, D. An information-theoretic definition of similarity. In Proc. of the Int’l Conference on Machine Learning, 1998.

[14] Lin, D. An information-theoretic definition of similarity. Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98). 1998.

[15] Lord PW, Stevens RD, Brass A, Goble CA: Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.

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

[17] Monica Chagoyen, Pedro Carmona-Saez, Concha Gil, Jose M Carazo and Alberto Pascual-Montano (2006). "A literature-based similarity metric for biological processes." BMC Bioinformatics.

[18] Jose´ L. Sevilla, Vı´ctor Segura, Adam Podhorski, Elizabeth Guruceaga, Jose´ M. Mato, Luis A. Martı´nez-Cruz, Fernando J. Corrales, and Angel Rubio (2005). “Correlation between Gene Expression and GO Semantic Similarity” IEEE/ACM Transaction on computational biology and bioinformatics, vol.2, No. 4

[19] Speer, N.; Spieth, C.; Zell, A. A Memetic “Clustering Algorithm for the Functional Partition of Genes Based on the Gene Ontology. Proceedings of the 2004 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB 2004). 2004.

[20] Maido Remm, Christian E. V. Storm and Erik L. L. Sonnhammer. (2001). "Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.

[21] Download Gene Ontology: http://geneontology.org/GO.downloads.ontology.shtml

[22] Amigo Browser: http://amigo.geneontology.org/cgi-bin/amigo/go.cgi

[23] Blast Tool : http://www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi

[24] S.Cerevisiae WU-BLAST2 Search: http://seq.yeastgenome.org/cgi-bin/blast-sgd.pl?name=YJR155W&suffix=prot

[25] Blast Help: http://www.ncbi.nlm.nih.gov/blast/

[26] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). “Gapped BLAST and PSI-BLAST: a new generation of protein data base search programs”. Nucl. Acids Res. 25, 3389-3402.

[27] MEDLINE. Available: http://www.cas.org/ONLINE/DBSS/medliness.html

[28] G.A. Miller (1995). “WordNet: A Lexical Database for English,” Comm. ACM, vol. 38, no. 11, pp. 39-41.

[29] MeSH. Available: http://www.nlm.nih.gov/mesh/meshhome.html

[30] UMLS: Unified Medical Language System. Available: http://www.nlm.nih.gov/research/umls/

[31] H. Kuntz and M. V. Berkum, “SNOMED CT® A standard Terminology for Healthcare”. Available: http://www.sst.dk/upload/informatik_og_sundhedsdata/sundhedsinfo rmatik/term inologi/kuntz_ vanberkum_snomedct_30mar05.pdf

[32] The International Classification of Diseases, 9th Revision, Clinical Modification" (ICD-9-CM)http://icd9cm.chrisendres.com/

[33] Saccharomyces Genome Database: Available: http://www.yeastgenome.org/

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

[35] MacCallum,R.M., Kelley,L.A. and Sternberg,M.J. (2000) “SAWTED: structure assignment with text description–enhanced detection of remote homologues with automated SWISS-PROT annotation comparisons.” Bioinformatics, 16, 125–129.