A Preliminary Study of Correlation between Depth and Path Length of GO Nodes with Gene Sequence Similarity

Elham Khabiri
School of Science and Computer Engineering
University of Houston-Clear Lake
Houston, TX, USA
khabirie8390@uhcl.edu

Abstract— **We proposed a new measure (Sim_{PLD}) for calculating the semantic similarity of terms in Gene Ontology (GO) based on the depth of least common ancestor (LCA) of two terms and the path length between them in GO hierarchy. The similarity between genes is computed based on this measure when it is applied to the GO-terms related to those genes. The method is based on the average of Sim_{PLD} 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 and proteins from two genomes: Saccharomyces Database (SGD) and Drosophila Melanogaster (FlyBase); and one dataset of Human-Yeast protein pairs. The experimental results proved that the method has fairly impressive agreement with Blast sequence similarity. Therefore Sim_{PLD} can be used as an automated tool for determining the similarity between genes and proteins.**

Keywords- Gene Ontology; Least Common Ancestor;

I. INTRODUCTION

Gene Ontology is considered as the most comprehensive resource for functional gene annotation [2]. GO is a controlled and structured vocabulary that is designed mainly to describe the molecular functions, biological process and cellular components of gene products independent of the organisms. 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. The structure of GO could be used as a measure to determine the functional similarity between genes. Sequence similarity is another way to predict the functional similarity among genes and have been used as a tool for functional prediction but some flaws have been detected in it. Up to 30% of the function annotations made through sequence similarity searches were found as erroneous [20]. The reason is when the genes are not evolving from a common ancestor the sequence similarity between them are not considerable. However they may have the similar functionality which is not reflected by sequence similarity tools.

In this paper, we investigate the correlation between our new ontology structure-based method with the sequence similarity. Our method measures the functional similarity between genes using the GO term annotations related to them. The similarity between the genes are measured as the average of all Sim_{PLD} for the terms annotated for each gene in which Sim_{PLD} is based on the depth of their least common ancestor and the path length between them. The method is evaluated by a series of experiments based on the correlation between Sim_{PLD} 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 or cross genomes. In the evaluation we selected genes from FlyBase and SGD and also we applied our measure to a dataset taken from [20]. Each dataset is divided into a number of sequence similarity ranges based on the E-value in the gene pairs. Then, we grouped the genes into those with high sequence similarity (HSS), low sequence similarity (LSS) and no sequence similarity (NSS) based on their BLAST e-value. Each one of these three groups was tested separately.

II. RELATED WORK

Ontology-based semantic similarity measures have been investigated for long time in the general English domain. For example, Resnik [17], Jiang and Conrath [6] and Lin [8] proposed information-content (IC) based measures for semantic similarity between terms and these measures were designed mainly for WordNet [14]. These measures are proven to be useful in natural language processing (NLP) tasks [1, 3, 4, 15]. Resnik's measure calculates the semantic similarity between two terms [t1, t2] in a given Ontology (e.g., WordNet) as the information content (IC) of the least common ancestor (LCA) of t1, t2. 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.

$$sim_{Resnik}(t_1, t_2) = \underset{t=LCA(t_1,t_2)}{-\log p(t)} \qquad (1)$$

Resnik's measure gives a value of zero or greater, in which the value of zero means minimum similarity. There is no maximum value for his measure. In an ontology, the deeper is the LCA of the two terms, the more is the information content of the LCA of them which shows the more similarity. In 1997

Jiang and Conrath [6] combined Resnik's method with an edge based approach and came up to the formula

$$dist_{JC}(t_1, t_2) = 2\log p(t)) - (\log p(t_1) + \log p(t_2)) \quad (2)$$
$$t=LCA(t_1, t_2)$$

that measures the distance between two terms. The distance is the reverse of their similarity measure. Lin [8] has also developed a similar measure that considered how close the terms are to their least common ancestor (LCA) in the ontology. In 1998 Leacock and Chodorow [10] proposed a formula for computing the semantic similarity or the relatedness between two terms in WordNet ontology:

$$sim_{LC}(t_1, t_2) = -\log \frac{Len(t_1, t_2)}{2 \times \max_{c \in WordNet} depth(c)} \quad (3)$$

in which "Len" is the minimum path between t1 and t2. Wu and Palmer [24] also applied both the PL between each term with the LCA of two terms and the depth of LCA of them. Budanisky and Hirts [3] investigated the relatedness of Resnik [17], JC [6] and Lin's [8] measures in WordNet ontology and founded JC [6] as a superior measure to all other ones. These measures were all applied to the non-biomedical ontologies. In the Biomedical domain, Rada et al. [16] proposed the first semantic similarity measure by using path length between biomedical terms in the MeSH ontology [13] as a measure of semantic similarity. Several other biomedical ontologies, within the framework of UMLS (unified medical language system) [23], have also been used for measuring semantic similarity in bioinformatics [1], e.g. Snomed-ct [7] and ICD9CM [22].

Lord et al. (2003) [9] were the first to apply a measure of semantic similarity to Gene Ontology. They proposed an IC-based technique for calculating the semantic similarity of protein pairs based on Resnik's measure [17]. The semantic similarity between two proteins is defined as the average similarity of all GO terms with which these proteins are annotated. Sevilla et al. (2005) [18] analyzed the correlation between gene expression and Resnik's, Jiang and Conraths' and Lin's measures of semantic similarity [17, 6, 8]. They concluded that Resnik's measure correlates well with gene expression. More recently, Schlicker et al. (2006) [20] 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.

III. 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 the two terms [16, 4]. GO is considered the most comprehensive resource for gene functional information. The path length and the depth of LCA between two terms have never been investigated in Gene Ontology as a potential measure of similarity between GO terms leading to functional similarity measure between genes. In our method, we computed the depth of the least common ancestor (LCA) and the path length between the two terms. Then we measured the similarity between two genes based on the semantic similarity values between their GO term annotations. Let us define the path length function PL() between two GO terms go_x and go_y as follows:

$$PL(go_x, go_y) = \textit{the minimum path length in the GO graph between the two GO terms } go_x \textit{ and } go_y \quad (4)$$

The similarity Sim_{PLD} between two go terms go_x and go_y is defined as:

$$Sim_{PLD}(go_x, go_y) = \ln\left(\frac{depth(lca(gox, goy))}{Max\_dpth}\right) - \ln\left(\frac{PL(gox, goy)}{2 \times Maxdpth}\right) \quad (5)$$

which considers both depths of LCA between two terms and the path length between them. In equation 5, the first term is divided (scaled) by the maximum depth in the GO while the second term is scaled by 2 times the maximum depth in GO which implies the maximum PL in the gene ontology. In our experiment we got the values of Sim_{PLD} ranged between -2 and 2.

A. Path Length between Genes

Given two genes Gp and Gq such that gene Gp 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 Gq is annotated with m different GO terms. The similarity between genes are measured by calculating the average of Sim_{PLD} between the GO terms annotated for both genes in a given gene pair.

$$sim_{PLD}(g_p, g_q) = avg\{sim_{PLD}(go_x, go_y) \mid x : 1..n, y : 1..m\} \quad (6)$$

IV. Experimental Results and Evaluation

To evaluate our method, we used three datasets of genes extracted from SGD (Saccharomyces cerevisiae), FlyBase (Drosophila melanogaster) [22, 34] and a human-yeast protein pairs dataset used in [20] in which proteins are extracted from UniProt and each pair consists of one protein from Homo Sapiens and one from the Saccharomyces cerevisiae genomes. The sizes of chosen datasets are 1000, 2000, and 3000 pairs respectively.

There are few methodologies for evaluating the similarity values computed by a measure. In NLP, for example, the two common approaches for comparing the computed semantic similarity values of a given measure is (a) by the correlation 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][4].

In the scope of this paper, i.e., within the context of gene functional similarity using GO annotations, the evaluation methodologies include: - comparing the computed similarity values with gene sequence similarity [1, 4, 6, 20] or with gene expression profiles [18]. In this paper we followed the first approach, and we compared our measure with the sequential similarity measures.

We divided the datasets into different groups based on the Blast E-value of the gene pairs. E-value is a metric to show the sequence similarity among the genes and it ranges between zero and one. Those pairs with zero values are considered sequentially similar and the E-value of 1 shows that there is not a significant similarity among the genes. In our experiments we grouped the gene pairs with the Blast E-value $\leqslant$10-5 as high sequence similarity (HSS). 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.

We found three separated group for SGD dataset (HSS, LSS and NSS) and two groups for FlyBase dataset (HSS and NSS). For the third dataset we used the data from [20] paper which have already grouped the dataset into HSS, LSS and NSS.

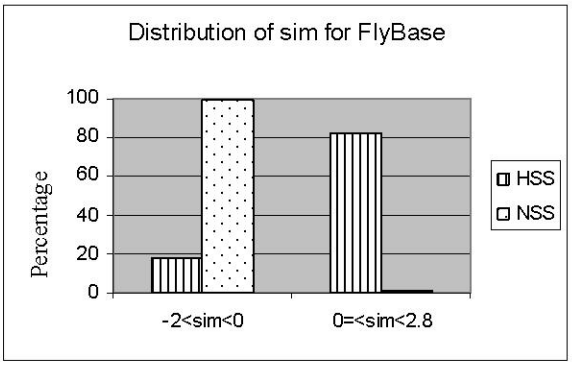

Figure 1. Distribution of Sim_{PLD} value between gene pairs in FlyBase dataset

As it is shown in figure 1, in FlyBase dataset, nearly all of the genes that have no sequence similarity have the Sim_{PLD} value of less than zero. Among those with high sequence similarity more than 80% have the Sim_{PLD} of greater than zero which shows a very high correlation of our result with the sequential similarity.

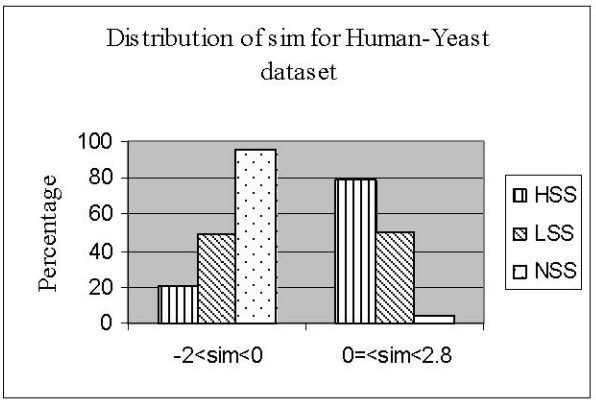

Figure 2. Distribution of Sim_{PLD} value between gene pairs in SGD dataset

In figure 2 which is related to the SGD dataset, more than 90% of NSS genes, have the Sim_{PLD} value of less than zero. More than 70% of LSS genes have the Sim_{PLD} value of less than zero and more than 60% of HSS genes have the Sim_{PLD} value of greater than zero which still shows agreement with sequential similarity.

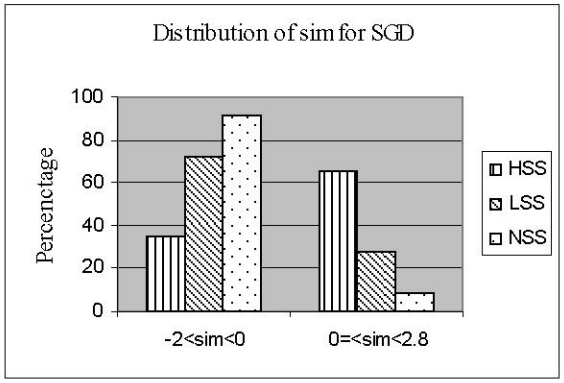

Figure 3. Distribution of Sim_{PLD} value between gene pairs in Human-Yeast dataset

In figure 3 more than 90% of NSS genes from the third dataset, have the Sim_{PLD} value of less than zero. Half of the LSS proteins have the functional similarity of less than zero and the other half have the Sim_{PLD} value of greater than zero which we expect from the proteins with low sequence similarity. Also more than 60% of HSS genes have the Sim_{PLD} value of greater than zero which is correlated with the sequential similarity measure. Therefore for the most of the genes with high sequence similarity we have found Sim_{PLD} values greater than zero and those with no sequence similarity have the Sim_{PLD} value of less than zero.

We also computed the average Sim_{PLD} value for all gene pairs in the SGD with high sequence similarity (HSS) which was 0.11 whereas the average Sim_{PLD} value for all SGD with low sequence similarity (LSS) and no sequence similarity (NSS) gene pairs were -0.54 and -0.85 respectively. For FlyBase we had the similarity values of 0.71 and -0.92 for HSS and NSS respectively. This is also another indicator that the HSS gene

pairs have significantly higher *sim* values compared with the LSS and NSS.

In summary, our evaluation experiments involved more than 3000 genes and 3000 protein pairs having high, low, or no sequence similarity from three different datasets. All the experimental results support the fact that there is significant correlation between the sequence similarity of genes and semantic similarity using Sim_{PLD}. This proves that the depth of LCA of two terms along with the path length between gene annotation terms using GO can be a reliable measure for gene functional similarity.

We have represented the results in more diagrams with analysis that shows the distribution of the Sim_{PLD} value in the three datasets with different ranges of Sim_{PLD}. For the space constraints we have not mention them in this paper. It would be freely available in our website.

V. Discussion and Conclusion

The path length and the depth of LCA between two terms have been never investigated in Gene Ontology as a potential measure of semantic similarity between GO terms leading to functional similarity measure between genes. The existing techniques for finding gene functional similarity based on GO rely mainly on information content(IC) of the terms. We presented a novel technique for finding gene functional similarity based on GO annotation terms. The method is based on the average of our measure (Sim_{PLD}) 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 and proteins from two genomes of SGD and FlyBase and a dataset of Human-Yeast protein pairs. We have shown that this method correlates very well with gene sequence similarity by comparing large numbers of gene and protein pairs with sequence similarities computed by one the most reliable algorithms for that purpose (BLAST).

References

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

[2] Ashburner M. et al. (2000). "Gene ontology: tool for the unification of biology." The Gene Ontology Consortium. Nat Genet. 2000;25:25–9. doi: 10.1038/75556.

[3] Budanitsky A. and Hirst G.," Evaluating WordNet-based measures of semantic distance," Computational Linguistics, vol.32,1, March 2006.

[4] Caviedes JE, Cimino JJ. Towards the development of a conceptual distance metric for the UMLS. Journal of Biomedical Informatics, vol. 37, no. 2, pp. 77-85, 2004

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

[6] 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, 19–33, 1997.

[7] Kuntz H. and Berkum M. V., "SNOMED CT® A standard Terminology for Healthcare". Available:http://www.sst.dk/upload/informatik_og_sundhedsdata/sundhedsinformatik/terminologi/kuntz_vanberkum_snomedct_30mar05.pdf

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

[9] Lord PW., Stevens RD., Brass A., Goble CA. (2002) "Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation."

[10] Leacock C., Chodorow M. (1998). "Combining local context and WordNet similarity for word sense identification." In Christiane Fellbaum, editor, WordNet: An Electronic Lexical Database. The MIT Press, Cambridge, MA, chapter

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

[12] Medliner Available: http://www.cas.org/ONLINE/DBSS/medliness.html

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

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

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

[16] 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.

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

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

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

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

[21] Speer, N.; Spieth, C.; Zell, A. A Memetic (2004) "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).

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

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

[24] Wu Z., Palmer M. (1994). "Verb semantics and lexical selection." In Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics, pages 133–138, Las Cruces, New Mexico, June 1994.